JYU DISSERTATIONS 680

Susanne Jauhiainen

Potential of Predictive Modeling Methods for Individual Response

Applications and Guidelines for Sports Sciences

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY

JYU DISSERTATIONS 680

Susanne Jauhiainen

Potential of Predictive Modeling Methods for Individual Response

Applications and Guidelines for Sports Sciences

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi Agoran auditoriossa 3
elokuun 18. päivänä 2023 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 3, on August 18, 2023, at 12 o'clock.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2023

Editors
Marja-Leena Rantalainen
Faculty of Information Technology, University of Jyväskylä
Päivi Vuorio
Open Science Centre, University of Jyväskylä

ISBN 978-951-39-9697-0 (PDF)
URN:ISBN:978-951-39-9697-0
ISSN 2489-9003

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-9697-0

ABSTRACT

Jauhiainen, Susanne
Potential of Predictive Modeling Methods for Individual Response: Applications and Guidelines for Sports Sciences
Jyväskylä: University of Jyväskylä, 2023, 66 p. (+included articles)
(JYU Dissertations
ISSN 2489-9003; 680)
ISBN 978-951-39-9697-0 (PDF)

The amount of data and consequently machine learning (ML) approaches are increasing at a fast pace in sports sciences, opening many new possibilities but on the other hand, also challenges. Generally limited data together with attractiveness and accessibility of ML methods without proper knowledge lead to faulty models and results with improper interpretations. Therefore, it is critical that researchers are aware of the risks related to the use of ML and that there are clear standards and robust procedures for how to perform and report ML studies. Answering the urgent need, the first aim of this thesis is to provide guidelines on how to properly perform and report (predictive) ML studies in the field of sports science. The second aim is to assess whether predictive modeling methods can be used for producing more individual information, compared to traditional statistics, namely in sports injury prediction and talent identification.

This article-style dissertation consists of four published articles. Articles I, II, and III utilize predictive modeling methods for sports injury prediction or talent identification and especially highlight the proper use of methods and data. Article IV utilizes unsupervised machine learning to discover kinematic running patterns among healthy and injured runners.

As main results of this thesis, the predictive power of multiple contemporary sports science datasets and ML approaches is assessed, and their potential for individual response discussed. Moreover, guidelines for utilizing predictive modeling are described and a framework for robust and generalizable results is introduced. Results from Article IV further confirm the need for individual approaches and provide useful information for future prediction studies. Through the included articles, advances are achieved for ACL injury prediction, recognizing predictive knee and ankle injury risk factors, utilizing ML for talent identification in soccer as well as discovering novel and useful information and patterns from running injury data. Important information about potentially best data types and variables for sports injury prediction and talent identification is produced. The approaches developed and used in this research can be utilized similarly in many other tasks and domains as well.

Keywords: Predictive modeling, Individual response, Machine learning, Sports injuries, Talent identification

TIIVISTELMÄ (ABSTRACT IN FINNISH)

Jauhiainen, Susanne
Ennustavan mallintamisen potentiaali yksilölliselle vasteelle: Sovelluksia ja ohjeita liikuntatieteiden alalle
Jyväskylä: University of Jyväskylä, 2023, 66 s. (+artikkelit)
(JYU Dissertations
ISSN 2489-9003; 680)
ISBN 978-951-39-9697-0 (PDF)

Datan määrä ja koneoppimissovellusten hyödyntäminen lisääntyvät liikuntatieteissä kovaa vauhtia, avaten monia uusia mahdollisuuksia, mutta toisaalta myös haasteita. Haastava ja rajallinen data yhdistettynä menetelmien houkuttelevuuteen ja saatavuuteen johtavat usein virheellisiin malleihin, tuloksiin ja johtopäätöksiin jos näitä ei osata hyödyntää oikein. On erittäin tärkeää että tutkijat tuntevat koneoppimismenetelmien käyttöön liittyvät riskit ja että niiden hyödyntämiselle ja tulosten raportoinnille on selkeät ja robustit standardit. Tämän väitöskirjan ensimmäinen tavoite on vastata tähän tärkeään tarpeeseen ja esitellä ohjeet (ennustaville) koneoppimismenetelmätutkimuksille, erityisesti liikuntatieteisiin keskittyen. Väitöskirjan toinen tavoite on tutkia voidaanko ennustavan mallintamisen avulla tuottaa yksilöllisempää tietoa kuin perinteisillä tilastomenetelmillä urheiluvammojen ennustamisen ja lahjakkuuksien tunnistamisen sovellusalueilla.

Tämä artikkelityylinen väitöskirja koostuu neljästä julkaistusta artikkelista. Artikkelit I, II ja III hyödyntävät ennustusmenetelmiä ja korostavat erityisesti menetelmien ja datan oikeaoppista hyödyntämistä. Artikkelissa IV tutkitaan terveiden ja loukkaantuneiden juoksijoiden juoksutyylejä ohjaamattoman koneoppimisen avulla.

Väitöskirjan tutkimuksessa arvioidaan useiden nykyaikaisten ja suurten liikuntatieteen datojen ja koneoppimismenetelmien ennustusvoimaa ja pohditaan niiden potentiaalia yksilöllisemmän tiedon tuottamiseksi. Hyödyllistä tietoa tuotetaan polven eturistisidevammojen ennustamiseen, polvi-ja nilkkavammoja ennustavien tekijöiden tunnistamiseen sekä lahjakkuuksien tunnistamiseen jalkapallossa. Artikkelin IV tulokset puolestaan vahvistavat yksilöllisten lähestymistapojen tarvetta ja tarjoavat tärkeää tietoa ennustustutkimuksia varten. Lisäksi esitellään ohjeet ennustavan koneoppimisen hyödyntämiseen liikuntatieteissä ja tuotetaan lähestymistapa jonka avulla saadaan robusteja ja yleistyviä tuloksia. Tärkeää tietoa potentiaalisesti parhaista datalähteistä ja muuttujista urheiluvammojen ennustamiseen ja lahjakkuuksien tunnistamiseen tuotetaan. Väitöskirjassa kehitettyjä lähestymistapoja ja ohjeita voidaan hyödyntää samoin myös monissa muissa aiheissa ja aloilla.

Avainsanat: Ennustusmenetelmät, Yksilöllinen vaste, Koneoppiminen, Urheiluvammat, Lahjakkuuksien tunnistaminen

Author

Susanne Jauhiainen
Faculty of Information Technology
University of Jyväskylä
Finland

Supervisors

Adjunct professor Sami Äyrämö
Faculty of Information Technology
University of Jyväskylä
Finland

Adjunct professor Jukka-Pekka Kauppi
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Pekka Neittaanmäki
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers

Professor Larisa Beilina
Department of Mathematical Sciences
Chalmers University of Technology and
University of Gothenburg
Sweden

Professor Elena M. Gutierrez-Farewik
KTH MoveAbility Lab
Department of Engineering Mechanics
KTH Royal Institute of Technology
Sweden

Opponents

Professor Olavi Airaksinen
School of Medicine
University of Eastern Finland
Finland

Adjunct professor Tuomo Kauranne
University of Eastern Finland
CEO, Arbonaut Ltd
Finland

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors Sami Äyrämö, Jukka-Pekka Kauppi, and Pekka Neittaanmäki for all of their guidance and help during this project. Sami, simply thank you for being the best supervisor a student could wish for. Your comprehensive expertise in both research domains is inspiring and your help and connections have been truly invaluable to my research and studies. Jukka-Pekka, thank you for all the work you put into providing me excellent feedback at each phase of each article, it truly improved this work so much. Pekka, thank you for all the helpful advice and for always being so encouraging.

Moreover, I would like to express my gratitude to Jenny and Antti Wihuri Foundation for generously funding most of my research and for additional funding I would like to thank the Faculty of Information Technology and the PhD program. Some research was also conducted in projects funded by Business Finland. I also want to thank Professor Tommi Kärkkäinen for offering me my first job at the university, which introduced me to the world of sports data science.

I was fortunate to spend a lot of time abroad learning from some of the top sports scientists and get experience working in truly multidisciplinary environments, thank you, KAUTE foundation, Emil Aaltonen Foundation, Ella and Georg Ehrnrooth foundation, and the University of Jyväskylä mobility grant for funding the research visits. Thank you, Professor Reed Ferber, for so warmly welcoming me (twice) into your research team at the Faculty of Kinesiology, University of Calgary. I also made so many friends in the team and the visiting students office, thank you everyone for all the fun and new experiences we had! Thank you, Professor Tron Krosshaug, for including me into your research at the Oslo Sports Trauma Research Center, Norwegian School of Sport Sciences and for making learning new things so fun and interesting.

I would also like to thank all coauthors and collaborators, this would not have been possible without your expertise and data. Thank you, Professor Larisa Beilina and Professor Elena M. Gutierrez-Farewik, for reviewing my thesis and thank you, Professor Olavi Airaksinen and Docent Tuomo Kauranne for being opponents. My colleagues and (ex-)officemates Mirka Saarela and Ilkka Rautiainen, thank you for all the conversations and peer support. And thanks to the whole Spectral Imaging Laboratory group for adopting me this past spring, you made finalizing this thesis a lot more fun!

Kiitos perheelleni ja sukulaisilleni – rakkaudesta, tuesta ja rukouksista. Vanhemmilleni erityiskiitos kasvatuksesta, luottamuksesta ja vapaudesta. Tärkeimpänä, suurimmat kiitokseni haluan lausua tyttärellemme Jennille ja miehelleni Arille. Jenni, olet jo nyt opettanut minulle enemmän kuin mikään tai kukaan muu maailmassa pystyisi. Ari, kiitos kaikesta tuesta ja rakkaudesta, maailman toiselle puolelle matkustamisesta ja videopuheluista yli aikavyöhykkeiden. On upeaa seikkailla tätä elämää eteenpäin teidän kanssa.

Jyväskylä 5.6.2023
Susanne Jauhiainen

LIST OF FIGURES

LIST OF TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

I Susanne Jauhiainen, Sami Äyrämö, Hannele Forsman, Jukka-Pekka Kauppi. Talent identification in soccer using a one-class support vector machine. *International Journal of Computer Science in Sport*, 18(3), 125-136, 2019.

II Susanne Jauhiainen, Jukka-Pekka Kauppi, Mari Leppänen, Kati Pasanen, Jari Parkkari, Tommi Vasankari, Pekka Kannus, Äyrämö. New machine learning approach for detection of injury risk factors in young team sport athletes. *International Journal of Sports Medicine*, 42(02), 175-182, 2021.

III Susanne Jauhiainen, Jukka-Pekka Kauppi, Tron Krosshaug, Roald Bahr, Julia Bartsch, Äyrämö. Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes. *The American Journal of Sports Medicine*, 50(11), 2917-2924, 2022.

IV Susanne Jauhiainen, Andrew J. Pohl, Sami Äyrämö, Jukka-Pekka Kauppi, Reed Ferber. A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns. *Scandinavian Journal of Medicine and Science in Sports*, 30(4), 732-740, 2020.

1 INTRODUCTION

This thesis focuses on the development and utilization of machine learning (ML) methods and approaches to extract useful information from datasets in the field of sports. The aim of this research is twofold. The first aim is to assess whether ML methods can be used to produce more individual information compared to traditional statistical methods for, for example, sports injury prediction and talent identification. The second aim is to describe ML analysis pitfalls common in sports sciences and provide solutions to overcome those, answering an urgent need in the field (Richter et al., 2021; Riley, 2019; Ley et al., 2022). To this end, a framework for handling uncertainty in predictive ML and producing robust predictive results is introduced. The main focus is especially on predictive ML methods, but others, such as clustering are discussed and utilized as well.

1.1 Key concepts

First, important key concepts of this research are introduced, namely predictive modeling and individual response.

1.1.1 Predictive modeling

Predictive modeling is used to make predictions (or classifications) based on some historical data. In predictive modeling, the model should be tested on separate and previously completely unseen data to test its generalization performance and predictive power (Breiman, 2001b; Ramspek et al., 2021; Hastie et al., 2009). There is a lot of confusion and contradictory definitions of prediction in the sports science literature as well as online, leading to false interpretations and conclusions and unjust comparison of different methods, approaches, or data. Terms like statistical modeling, machine learning, predictive or explanatory modeling are widely known and used but sometimes in conflicting or even incorrect ways. The problem has been widely recognized and discussed in other fields, such as

medicine (Ramspek et al., 2021; van Diepen et al., 2017; Waljee et al., 2014) and this thesis aims to provide clear definitions and guidelines for future sports science research.

Statistical modeling

Statistical modeling (also known as statistical learning) has been defined as *"learning from data"* (Hastie et al., 2009). A more precise definition by Dangeti (Dangeti, 2017) goes "Statistical modeling is applying statistics on data to find underlying hidden relationships by analyzing the significance of the variables". Shmueli (2010) divides statistical modeling to three categories, namely explanatory, predictive, and descriptive modeling. The former two are often thought as the most common approaches in statistical modeling and are described in the following subsection. Descriptive modeling on the other hand is "aimed at summarizing or representing the data structure in a compact manner" (e.g., average, standard deviation) (Shmueli, 2010). It can be thought as *learning the data*, rather than trying to learn from it through inference or prediction.

Predictive versus explanatory modeling

The two main approaches of statistical modeling are explanatory and predictive modeling (Shmueli, 2010; Sainani, 2014). Explanatory modeling is the use of statistical methods to test causal hypotheses and detect variables that are associated with the outcome (Shmueli, 2010; Sainani, 2014; Waljee et al., 2014). The focus of explanatory modeling is on understanding and explaining the phenomena of interest in the data sample. Predictive modeling, on the other hand is focused on making predictions (for future or past data) and can create new information from the data. It is defined as "the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations" (Shmueli, 2010). In predictive modeling, the model should always be tested and predictions made on separate, completely unseen by the model, data (Breiman, 2001b; Ramspek et al., 2021; Hastie et al., 2009).

While explanatory methods can only discover associations in the data, predictive modeling can be used to detect factors that can actually predict the event of interest. An important point to realize and remember is that models with high explanatory power do not necessarily have high predictive power (Shmueli, 2010). Therefore, even with a strong association between, for example, sports injury risk and a certain variable, we can not draw conclusions about that variable being predictive of injury. Predictive models with proper assessment (see Section 3.5) are necessary to be able to make any conclusions about the predictive ability. On the other hand, variables might be included in predictive models even though not causally related to the outcome (Sainani, 2014; Moons et al., 2009). For example, a variable might be correlated with the injury risk and thus included as important in the prediction but correlation does not imply causation (Altman and Krzywinski, 2015). Moreover, predictive modeling sometimes sacrifices model

interpretability against predictive accuracy (Kuhn et al., 2013) and in general, explanatory methods are necessary along predictive research and should be used to, for example, confirm the hypothesis discovered with predictive methods.

Machine learning versus statistical modeling

ML is a field of developing, understanding, and utilizing methods that learn from data to solve or improve performance on different tasks (Mitchell and Mitchell, 1997; Grus, 2019). A common outlook of differences of ML and statistical modeling is their purpose; statistical models focus on explaining the data and relationships between variables and drawing inferences from the sample while ML is focused on making accurate predictions based on the data (Ij, 2018); In this definition statistical modeling loosely corresponds to explanatory modeling and ML to predictive modeling. The distinction between ML and predictive modeling is ambiguous. One way to think about it is that predictive modeling is a subfield of ML, focused on making predictions based on historical data, while ML in general comprises of many other applications as well. Other than prediction applications in ML include, for example, unsupervised tasks such as clustering. A limitation of most explanatory modeling approaches is that they focus on a small number of variables and their linear associations while predictive ML can be used for larger data including nonlinear relationships (Ley et al., 2022). Moreover, explanatory modeling is limited by prior assumptions and hypotheses while predictive ML can be used for generating novel hypotheses in a data-driven manner. In general, many methods can be used for both explanatory (statistical) modeling and predictive ML (Shmueli, 2010; Ij, 2018). For example, linear regression in explanatory modeling models the relationship between the response and one or more explanatory variables whereas in predictive ML a linear model trained on training data can be used to make predictions on test data.

1.1.2 Individual response

Individual response refers to the unique way that each subject responds to certain training or treatment and the importance has been recognized in sports (Bouchard and Rankinen, 2001) as well as healthcare (Godman et al., 2013). Personalized medicine (also known as precision medicine) is an approach where treatment is tailored based on characteristics of an individual or a group of individuals for optimal response. Majority of existing research in training (Hecksteden et al., 2015) as well as current exercise prescriptions and recommendations are derived based on population averages, not considering individual differences and variation between people. These general prescriptions and recommendations will work well in average but the response will vary between individuals and for some people even adverse effects might occur following these general recommendations. For example Bouchard et al. (2012) detected adverse metabolic effects to regular exercise, where exercise-induced change worsened the assessed risk-factor measurements. So, while on average the exercise might have improved participant's

measures a certain amount, some participants had way larger improvements than the average while some ended up encountering negative effects.

Previously, differences in individual responses in sports have been recognized in many studies. There is strong evidence to support that response to regular physical activity is very heterogeneous and people response individually to excercise (de Lannoy et al., 2017; Bouchard and Rankinen, 2001; Ahtiainen et al., 2020). Heterogeneous responses to physical activity have been observed with, for example, the VO2max, heart rate, HDL-cholesterol levels, and systolic blood pressure (Bouchard and Rankinen, 2001; Bouchard et al., 2012; Leon et al., 2002). Therefore, considering individual response and providing more personalized training and treatment prescriptions would be highly important in order to, for example, prevent unnecessary sports injuries, optimize performance, and avoid negative health outcomes. ML can help move further from current average based practices and predictive ML in particular can be used for detection of individual probabilities for, for example, sports injury risk or the most important variables for each individual in a data-driven manner.

1.2 Background and research motivations

There is an increasing amount of data collected in sports (Brefeld and Zimmermann, 2017) and health (Murdoch and Detsky, 2013) due to advances in data collection technologies. For example, wearable devices (such as watches or clothing) that include multiple sensors and global positioning system (GPS) have rapidly advanced and become common among both recreational and elite athletes. Many sports are also utilizing novel video analysis techniques to analyze team and player tactics, passes, playing time, amount of movement, goals, or injuries et cetera during games and practice (Herold et al., 2019; Chen et al., 2012; Dietrich et al., 2014). Game score recording, prediction, and analysis have become very common across different sports and offer new information to coaches, players, viewers, and gamblers (Bunker and Thabtah, 2019). In addition, many advances have risen with research data collections; motion analysis systems have become more advanced and outside-the-lab solutions are increasing (Cuesta-Vargas et al., 2010; Fong and Chan, 2010; Mousavi et al., 2020), and more accessible and easy-to-use equipment and methods are constantly developed.

The increasing amount of versatile data has enabled the use of sophisticated ML methods and Figure 1 demonstrates the rapid increase of publications in *sports* and *machine learning* [1]. These methods enable the use of a large number or variables and accounts for their interactions and non-linear relationships as well. Compared to more traditional statistical methods, i.e., explanatory modeling, ML often focuses less on assessing predefined hypothesis but rather tries to find what novel, sometimes even surprising, information the collected data has

[1] Retrieved from PubMed https://pubmed.ncbi.nlm.nih.gov with keywords *sports* and *machine learning*

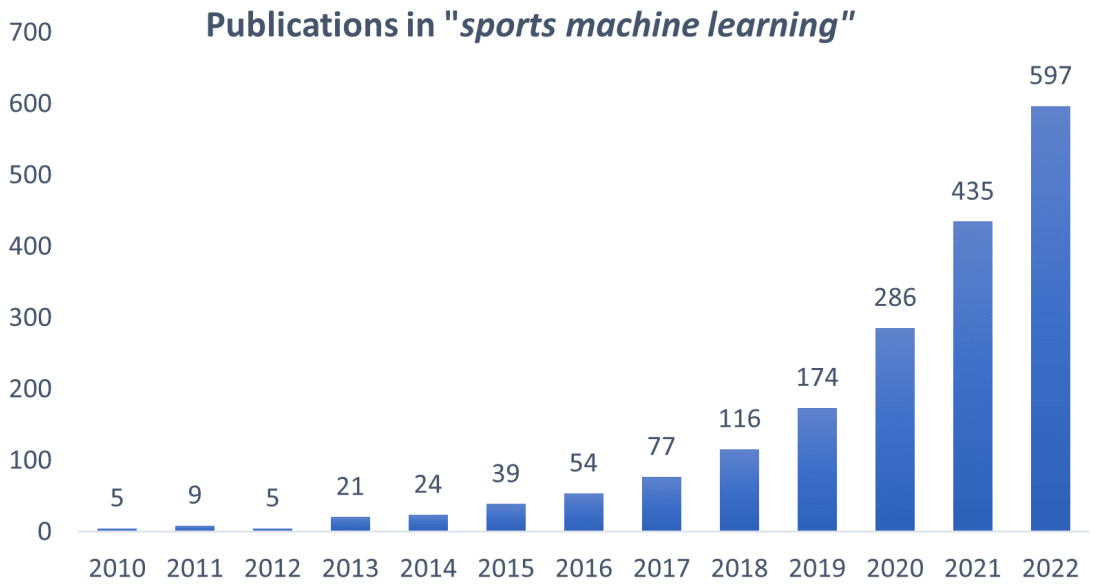


FIGURE 1 The yearly number of publications retrieved from PubMed with keywords *sports machine learning*

to offer. Additionally, while explanatory approaches often provide information about associations, correlations, or causalities on an population average level, ML can be used for prediction and analysis on a more individual level. Moreover, the data needs to be suitable for the given task as no method or approach can discover information that is not captured in the data to begin with (Riley, 2019). Therefore, this thesis assesses the potential of ML and large and contemporary sports science datasets for more individual information.

Furthermore, with the increase of ML research within sports sciences, there are many pitfalls that have not been considered and as a result a risk of faulty approaches and result interpretation. There is a need to educate scientist in ML on the topic and disciplines need to develop clear standards for how to perform and report on ML in their fields (Riley, 2019). Therefore, this thesis and included research discusses common pitfalls providing examples from sports science literature and offers solutions and guidelines for future research.

1.2.1 Sports injuries

Despite the undeniable benefits of sports and physical activity, the more exercise one does the higher the risk for suffering a sports injury gets. Many sports carry a high injury incidence rate and the worldwide prevalence of sports injuries is alarming (Hootman et al., 2007). For example, running, one of the most popular ways to increase and maintain fitness in many populations all over the world (Van Middelkoop et al., 2008; Van Gent et al., 2007), has been observed to have annual prevalence of lower extremity injuries between 19.4% to 79.3% (Van Gent et al., 2007), with a widely accepted estimate of 50% of runners experiencing an

running related injury annually (Fields et al., 2010). Moreover, in team and cutting sports, anterior cruciate ligament (ACL) injuries are a major and growing concern (Bahr and Holme, 2003).

Sports injuries can have significant effects on the health and performance of a person and may even cause prolonged problems in persons life (Myklebust et al., 2003). Sports injuries can lead to, for example, pain, loss of playing or working time, and decreased motility and stability (Myklebust et al., 2003). Prolonged problems can eventually even lead to artificial joints surgeries, which cause financial burden to the society. Some other possible acute or prolonged outcomes are back and joint pain, which are both major public health problems that cause huge economic burdens to individuals, companies and social systems (Maniadakis and Gray, 2000; Gaskin and Richard, 2012). In addition, there is a large variety of mental issues that can often follow sports injuries, like depression, increased anger and fatigue and cognitive impairment (Hutchison et al., 2009; Mainwaring et al., 2010; Hutchison et al., 2011). All the above conditions can be largely prevented if appropriate actions are taken in time. Sports injury prevention and treatment should also consider individual response. For example, some athletes might require more rest or have genetic factors increasing injury risk and injury rehabilitation should also be tailored individually. Predictive modeling can be used to, for example, recognizing predictive injury risk factors or providing individual injury risks based on athlete data. Furthermore, sports injuries are complex and multifactorial (Meeuwisse et al., 2007) and ML can utilize a large number of variables including their linear and non-linear relationships in prediction.

1.2.2 Talent identification

In many sports as well as other areas in life, the detection of talented individuals at early age can open better opportunities and possibilities and greater development. In sports, detected talents can be offered higher quality training and environments. Furthermore, considering individual response to training, more emphasis can be put on planning and offering personalized training programs to talents. With personalized, quality training, their performance development can be optimized and further accelerated. In many fields, talent and future potential is a sum of variety of different skills and qualities, making detection a complicated and multifaceted task. Therefore ML, being able to account for a large set of variables and their interactions, has potential for future talent identification (i.e., prediction) or the detection of predictive variables.

1.3 Research questions

The main research questions are as follows:

RQ1 What is the potential of machine learning methods and current datasets for individual injury prediction or talent identification in sports?

RQ2 What are the most common pitfalls in predictive machine learning and data in the field of sports science and how to tackle these?

1.4 Structure of the thesis

Chapter 2 introduces common ML pitfalls in sports science and an overview of the current state of predictive modeling research in sports injury prediction and talent identification. Chapter 3 focuses on the methodological background of the thesis and introduces the approaches and guidelines to overcome analysis pitfalls in predictive modeling. Chapter 4 includes an overview of the included publications and their most important results and Chapter 5 discusses the research questions, conclusions and new implications to the fields.

2 PREDICTIVE MODELING AND MACHINE LEARNING IN SPORTS SCIENCE – CURRENT STATE

Traditionally, research in sports science has mainly been based on explanatory modeling (Richter et al., 2021; Rossi et al., 2018) with a focus on explaining or understanding phenomena of interest in the data. Recently, the increasing amounts of data and availability and recognizability of ML methods has lead to more studies utilizing predictive modeling. However, a big problem in the field is that the difference between explanatory and predictive modeling is unclear, leading to contradictory and misleading interpretations. The difference is described in detail in Section 1.1.1, but shortly, in order to make any conclusions about the predictive ability of data or the model it has to be tested on completely unseen data. A recent example of the issue is a sports injury prediction review Bullock et al. (2022) that straightforwardly compares the performance metrics from studies using both explanatory or predictive modeling. The results in predictive studies are calculated based on separate test data and can thus seem lower but are generalizable and opposite to explanatory studies, directly indicate something about the predictive ability. This chapter presents the current state of predictive ML in sports sciences and discusses most common pitfalls and challenges both in analysis and the data, focusing on the applications in included research, namely sports injuries and sports talent identification.

2.1 Common predictive machine learning pitfalls

The following terms are important to understand for predictive modeling research and are all closely related and overlapping.

Model generalization performance refers to its ability to make accurate predictions on independent (i.e., unseen) test data and should be assessed appropriately (Hastie et al., 2001).

Overfitting of a model refers to model fitting to correspond too closely the training data and as a consequence not being generalizable, i.e., not performing well on unseen test data.

Chance result refers to a result where the predictive performance is obtained by chance.

Uncertainty refers to the lack of generalizability, reliability (consistency), or validity (data and methods actually measuring what they are intended to measure) of the model and its results.

With sufficient computational resources, many ML methods have unlimited ability to fit to complex data, causing a great risk for overfitting. Therefore, the model generalization performance needs to be assessed appropriately to avoid overfitting and also to estimate the risk of chance results and uncertainty around the results. Chance results are widely recognized in, for example, neuroscience (Hosseini et al., 2020; Combrisson and Jerbi, 2015) and can happen, for example, due to model learning some noise or unintentional variations in data (Riley, 2019) or by randomly having a favorable test dataset. The importance of considering uncertainty in prediction has been previously recognized, for example, in healthcare (Chua et al., 2022). It can enter the analysis from different sources, including the data (quality, noise, size), model (which methods and hyperparameters to use), or sampling (variability from data splitting). In general, the more uncertainty there is, the higher the risk of chance results and the less generalizable results are. On the other hand, comprehensive assessment of generalization performance and uncertainty will exclude the risk of overfitting and chance results.

Having separate test data is not enough to prevent overfitting, but it also has to be treated properly to avoid data leakage, a common problem in sports sciences. Data leakage refers to information in the test data somehow leaking into training of the model which sacrifices the generalization performance of a model through risk of overfitting. Data leakage can happen in many parts of the ML process. For example, feature selection based on the whole data before model training is a typical approach but often causes clear overfitting (Hastie et al., 2001). Data leakage can also happen due to improper preprocessing (e.g., normalization or imputation) of the whole data at once (i.e., both test and training data) (Tampu et al., 2022; O'Neil and Schutt, 2013) as is common, for example, in sports injury prediction studies (Van Eetvelde et al., 2021). Also, improper hyperparameter optimization (Kaufman et al., 2012) where the analysis is repeated and hyperparameters (or any other part of the analysis) manually tuned based on the results leads to test data not being completely unseen anymore. A good example of a possible data leakage and a chance result is a recent ACL injury prediction study (Tamimi et al., 2021). While the study does use separate test data to assess model generalization performance, the high test accuracy (92%) compared to clearly lower training accuracy (70%) suggests overfitting to test data either by (unconsciously) repeatedly resampling the test dataset or purely by chance.

Moreover, dividing data to training and testing at random brings uncertainty to the analysis and results can vary largely based on the data split (Forman and Scholz, 2010), leading to a risk of chance results due to favorable test data. For example, this risk was highlighted in the results of a recent hamstring injury prediction study (Ruddy et al., 2018) where results from different k-fold cross-validation splits vary largely across repetitions with multiple methods (e.g., AUC from 0.37 to 0.65 with neural networks). Additionally, chance results due to learning noise in data can happen especially when the true phenomena is weak to begin with or when there is data leakage through which the noise is amplified. Results with low predictive performance should be considered at a risk of being due to noise in data and not true phenomena and therefore confirmatory analysis should be used. The effect of chance results is even more important to consider in the case of small and/or high-dimensional (i.e., large number of variables) datasets as well as imbalanced data, which are common cases in sports sciences. Thus, to reduce uncertainty, the data for assessing generalization performance should be chosen appropriately and based on the task and data at hand (Riley, 2019).

2.2 Sports injury prediction

Sports injury prediction is an emerging area but the validity of many studies is questionable and models are poorly developed (Van Eetvelde et al., 2021; Bullock et al., 2022). A recent review by Bullock et al. (Bullock et al., 2022) conclude that "Ninety-eight percent of sport musculoskeletal injury prediction models (and 79% of studies) were rated as high risk of bias." However, only fourteen out of the thirty included studies (47%) used separate data to test the model predictive ability, meaning more than half of the studies are more explanatory by nature, despite arguing about prediction. Another review by Eetvelde et al. (Van Eetvelde et al., 2021) that focused only on predictive ML studies concludes that "although the majority of the analyzed studies did apply ML methods properly to predict injuries, the methodological study quality was moderate to very low". Table 1 summarizes current sports injury prediction studies. The studies are selected based on the previous reviews (Van Eetvelde et al., 2021; Bullock et al., 2022), references included in the thesis articles and prompt literature scoping. All data in the table are collected from the original studies by the author of this thesis.

TABLE 1 Summary of the sports injury studies. Results are mean (or median) of metrics over the folds and/or repetitions, except for (Thornton et al., 2017) a mean over data representations. "Best" refers to best result across different methods, models, and/or data representations and preprocessing approaches, picked from the text by the author of this thesis (otherwise only single result was produced).

Study	Results	Repetitions	Model Assessment	Model selection	Data leakage	N	Injuries
(Rossi et al., 2018)	Best mean AUC 0.78	10 000	Stratified 2-fold CV with 70% of data	30% for feature selection and model selection		26 (952 sessions)	2 %
(Karuc et al., 2021)	Best mean AUC 0.62	20	20-fold CV	Inner CV on training data		556	16 %
(Oliver et al., 2020)	Best (mean?) AUC 0.66	NR	Stratified 5-fold CV	NR		355	28 %
(Rommers et al., 2020)	F1-score 85%, sensitivity 85%, precision 85%	NR	80% for training, 20% for testing	Inner CV on training data		734	50%
(Colby et al., 2018)	Best (mean?) AUC 0.64	NR	10-fold CV	NR		60 (7147 sessions)	58 injuries in total
(Carey et al., 2018)	Best mean AUC 0.72	50	2 seasons for training, 1 season for testing	Inner 10-fold CV on training data	Unclear if PCA on whole data	75 (13867 sessions)	3%
(López-Valenciano et al., 2018)	Best mean AUC 0.75	NR	5-fold CV	NR	Imputation done to whole data	122	22 %
(Ruddy et al., 2018)	Best median AUC 0.58	10 000	70% for training, 30% for testing	Inner 10-fold CV on training data	Normalization done to whole data	362	15 %
(Ayala et al., 2019)	Best mean AUC 0.84	NR	Stratified 3-fold CV	NR	Imputation done to whole data, feature discretization as well as imbalance handling (ratio) selected based on predictive performance	86	21 %

(Thornton et al., 2017)	Best mean AUC 0.74	NR	70% for training and 15% for testing	15% for validation (unlear whether used at all)	Feature selection done based on whole data	25 (Unclear how many days)	2,5% of days unavailable (i.e., injured)
(McCullagh and Whitfort, 2013)	Mean accuracy 83%, sensitivity 95%, and specificity 81%	NR	10-fold CV	Hyperparameters "*were derived by conducting a series of trials involving varying parameters and assessing the effect on the neural network's output*"	Hyperparameters tuned using test data	39 (1210 sessions)	13%
(Whiteside et al., 2016)	Mean accuracy 75%, sensitivity 74%, and specificity 75%	NR	Stratified 5-fold CV	"*The optimized classifier was that which produced the greatest classification accuracy*"	Feature and model selection done based on whole data	208	50%
(Feijen et al., 2021)	(Mean?) AUC 0.71	250	Bootstrapping	NR	Feature selection done based on whole data, imputation done to whole data,	129	32%
(Luu et al., 2020)	Best mean AUC 0.95	NR	10-fold CV	Unclear	Feature selection done based on whole data	2322 (unclear how many sessions/ seasons)	6982 in total
(Tamimi et al., 2021)	Best accuracy 92%	NR	76% for training, 24% for testing	NR	Training accuracy 70% suggest overfitting to test data	100	50%

The achieved predictive performance are varying in accordance with the ML analysis protocols. There is a risk of overfitting in majority of the sports injury prediction studies through data leakage. Especially those where feature selection or hyperparameter tuning (i.e., model selection) were done utilizing the whole data (i.e., including test data) can be considered at high risk of overfitting. Only five studies show no signs of data leakage and two of them also utilized repetitions to account for variation in the data splitting (Karuc et al., 2021; Rossi et al., 2018). Additionally, in (Carey et al., 2018) it was unclear whether preprocessing was done on whole data or not. Importantly, many of these studies would benefit from more detailed description on how preprocessing was implemented with data splitting or mentioning if no preprocessing was done. In total, five studies used repetitions to assess the variability across different train/test data splits, most demonstrating a large variation in results with almost all trained models. This highlights how studies without repetitions include uncertainty and are at a risk of change results. Five studies use inner cross-validation to avoid data leakage from hyperparameter tuning (Karuc et al., 2021; Ruddy et al., 2018; Rommers et al., 2020; Rossi et al., 2018; Carey et al., 2018). Additionally, one study reports dividing a separate training, validation and test data (Thornton et al., 2017) but from the text it is unclear whether the validation data was used at all. Furthermore, only two studies report metrics for the training data (Tamimi et al., 2021; Rommers et al., 2020) and for the first (Tamimi et al., 2021), the high test accuracy (92%) compared to clearly lower training accuracy (70%) strongly suggests overfitting to test data. If model selection criteria are not clearly described, there is a risk for (unconscious) overfitting through manual repetitions to tune hyperparameters to achieve higher performance and therefore it is important for future studies to clearly describe how the final model was selected.

2.3 Talent identification

In sports, the concept of *talent* is very multifaceted and sport-specific (Baker et al., 2019). In soccer, for example, talent identification requires a great variety of physical features and technical skills (Reilly et al., 2000; Baker et al., 2019), including linear and non-linear relationships between those (Sarmento et al., 2018). Moreover, psychological skills and characteristics also play an important role at elite level (Macnamara and Collins, 2011). Most sports talent identification research have been done utilizing traditional statistical approaches (Reyaz et al., 2022). ML, however, enables the consideration of large set of variables as well as their linear and non-linear relationships. Furthermore, predictive ML can be used for predicting potential future elite athletes or recognizing variables most predictive of future success. While many previous talent identification studies report about prediction or detecting predictive variables, only a handful of studies utilizing predictive modeling exist to date.

Taha et al. (2018) first cluster 50 youth archers into high and low performing clusters and then achieve a 97.5% accuracy for prediction. They split 80% of data for training and 20% for testing and use inner 5-fold CV to optimize hyperparameters on the training data. Normalization was done to whole data, leading to data leakage. No repetitions or confirmatory analysis were utilized, but could be beneficial considering the small sample size. Barron et al. (2018) predict career development of 966 soccer players based on key performance indicators. Their model achieves an accuracy between 61.5% and 78.8% for predicting players moving down to lower level, remaining in Football League Championship, or moving up to the English Premier League. Feature selection is first done based on the whole data which causes data leakage. For prediction, they divide 60% of data for training, 20% for validation and 20% for testing. No repetitions or confirmatory test were done. PECOTA (Player Empirical Comparison and Optimization Test Algorithm) is a famous commercial system for predicting player performance in baseball based on player comparisons and their historical data (Prospectus, 2003). The methodology is based on utilizing similarity scores and projection but exact formulas are proprietary. While other similar systems have been developed, PECOTA has been considered as one of the most accurate (Lyle, 2007).

2.4 Data characteristics and common challenges

Traditionally, many datasets in sports science can be considered relatively small (Phinyomark et al., 2018), the main reason for this being the laboriousness of data collection methods. This is a challenge for the utilization and testing of ML methods (Richter et al., 2021), especially as for predictive modeling we should have separate test data. In addition to proper sample sizes, the data itself and features extracted from it need to be meaningful for the task (Richter et al., 2021). For example, in sports injury prediction, while current research suggest some variables that have (or do not have) predictive value, more research is still needed to find best types of predictors in different populations and settings.

Additionally, in many sports and sports medicine datasets, classes are slightly or even extremely imbalances. For example, the amount of healthy subjects in data is often inevitable larger than the amount of those that have certain sports injury, disease, symptom etc. and similarly in talent identification, the group of "normal" players is without exception much larger than the group of those that have the prerequisites to become elite athletes. Moreover, drop-outs are a common challenge in many longitudinal sports science studies, resulting into missing data. Common challenges in data related to human movements are that they are most often collected in a lab and the generalization to real world is unclear and their reproducibility is weak as different marker placements affect results largely (Gorton III et al., 2009). As solutions, recently there have been applications to utilize more data from wearable devices to advance or replace more traditional mo-

tion analysis data (Fong and Chan, 2010; Cuesta-Vargas et al., 2010) and suggestion to combine datasets within sports science to achieve larger samples (Richter et al., 2021; Ferber et al., 2016). Moreover, approaches to handle class imbalance and missing data are common in ML and most suitable ones for sports science tasks can be selected based on the task at hand as well as the (still limited) previous research.

3 FOUNDATIONS OF CONCEPTS AND METHODS

This chapter includes the methodological background for the included publications and introduces approaches and guidelines to overcome common analysis pitfalls in sports sciences discussed in Chapter 2. Section 3.2 discusses preprocessing of data and sections 3.3 and 3.4 shortly introduce the utilized ML methods. Sections 3.5 and 3.6 focus on introducing the approaches for overcoming the pitfalls, including proper assessment of predictive models and confirmation of results.

3.1 Mathematical definitions

Let us define a data with N observations and each having p attributes as a matrix

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \boldsymbol{x}_{1,2} & \cdots & \boldsymbol{x}_{1,p} \\ \boldsymbol{x}_{2,1} & \ddots & & \\ \vdots & & \ddots & \\ \boldsymbol{x}_{N,1} & \cdots & \cdots & \boldsymbol{x}_{N,p} \end{bmatrix} \quad \boldsymbol{X} \in \mathbb{R}^{N \times p} \tag{1}$$

Then the ith observation is a vector $\boldsymbol{x}_{i,:} = (x_{i,j})$ with $j = 1, \ldots, p$. The observations, i.e., rows of the matrix, can also be referred to as points or objects and represent the amount of data. The attributes of the observation, i.e., columns of the matrix, are also commonly known as variables, features or dimensions. The jth variable in data is a vector $\boldsymbol{x}_{:,j} = (x_{i,j})$ with $i = 1, \ldots, N$. Additionally, in classification and regression we denote the observed response of each observation $\boldsymbol{x}_{i,:}$ with a y_i. Moreover, ^{T} means the transpose of a vector or matrix. An indicator function is denoted with I so that $I(S)$ maps elements in (sub)set S to one and all other elements to zero.

3.2 Preprocessing

Preprocessing is a crucial initial step with any type of data analysis and can largely affect obtained results (Kuhn et al., 2013). Preprocessing and reporting practices in sports science are varying and partly incomplete and faulty (Van Eetvelde et al., 2021; Bullock et al., 2022; Phatak et al., 2022). Future sports science ML studies should familiarize themselves with proper data preprocessing and report each step in detail for reproducibility. This section shortly describes most common preprocessing steps within the context of sports sciences. The steps and their order will be chosen based on the data and study design at hand, considering that some ML methods can inherently handle these data "problems" (e.g., missingness, outliers) differently. Proper preprocessing of separate training and test data to avoid data leakage are described in Section 3.5.1.

3.2.1 Data selection and dimension reduction

After data collection, a target dataset for analysis is selected based on several criteria. For example, exclusion of variables and subjects with mostly missing data or excluding subjects with "uninteresting" or missing category (e.g., only interested in predicting non-contact injuries so contact injuries are excluded) are common steps. Further variable selection and dimension reduction can be done based on prior literature or expert knowledge, by summarizing metrics from signal (e.g., maximum or mean of steps from motion analysis), or by data-driven ML approaches such as principal component analysis (PCA). For achieving an optimal and robust set of variables for ML in sports science, Richter et al. (2021) suggest combining domain-specific knowledge-based and objective data-driven approaches together.

3.2.2 Normalization

Commonly the purpose of normalization is to scale all variables to comparable magnitudes or centering to a certain point, e.g., zero mean. Some ML methods have strict assumptions and absolutely require scaling and centering of data but even without such requirements, many methods benefit from unified data and numerical stability of calculations can improve (Kuhn et al., 2013). Other normalization techniques might include, for example, transforming skewed data to more symmetric (Kuhn et al., 2013). Some sports domain specific normalization methods have also been developed, such as a recent normalization key performance indicators to be used across a wide range of sports (Phatak et al., 2022).

A typical normalization is the z-score transformation where each variable $\boldsymbol{x}_{:,j}$ is transformed to follow the standard normal distribution with the following

$$\boldsymbol{x}'_{:,j} = \frac{\boldsymbol{x}_{:,j} - \mu_j}{\sigma_j}, \tag{2}$$

where μ_j is the mean and σ_j the standard deviation of the variable. Min-max scal-

ing, on the other hand, transforms the variable to a given interval, for example, between [0,1] with

$$\boldsymbol{x}'_{:,j} = \frac{\boldsymbol{x}_{:,j} - \min_{x_{:,j}}}{\max_{x_{:,j}} - \min_{x_{:,j}}}, \tag{3}$$

where $\min_{x_{:,j}}$ and $\max_{x_{:,j}}$ are the minimum and maximum values of the variable. Moreover, normalization between any interval [a,b] is achieved by

$$\boldsymbol{x}'_{:,j} = (b-a)\frac{\boldsymbol{x}_{:,j} - \min_{x_{:,j}}}{\max_{x_{:,j}} - \min_{x_{:,j}}} + a. \tag{4}$$

3.2.3 Handling missing data

Missing data is an almost unavoidable issue with any real world data (Kotsiantis et al., 2006) and common in sports science data. Missing values can occur as a result of human or equipment errors, but often the values are simply unavailable due to e.g., sensor has not been worn, athlete was injured during data collection, or poor athlete and team adherence (Benson et al., 2021). (Little and Rubin, 2002) divide missing data into three categories, namely:

MCAR: Missing completely at random - missingness is not depended on the data values, either missing or observed. In practice, data is seldom MCAR but, for example, athletes test data could be missing if someone forgot to save it or deleted it by accident. Dealing with MCAR data will not introduce bias to the data (Batista and Monard, 2003).

MAR: Missing at random - missingness depends only on the data that are observed and not on that which are missing. For example, athletes in certain teams (team information available for everyone) might have more missing data due to poor team and coach attitude toward that data collection.

MNAR: Missing not at random - missingness is not MCAR or MAR and the value of that variable is related to the reason why it is missing. For example, a player has (unreported) pain and is therefore missing some data due to not being able to participate the data collection. Dealing with MAR and especially MNAR data can introduce bias in data.

The way to handle missing data will depend on, for example, the reasons why they are missing (Little and Rubin, 2002), the amount of missing values and importance of reserving certain data. Observations with missing data can be excluded completely or whole data used with methods that can handle missing data (Little and Rubin, 2002). Alternatively, missing points can be imputed with statistics (e.g. mean, median) or predictive modeling (Batista and Monard, 2003).

3.2.4 Data cleaning and outliers

Real world data can often include incorrect, corrupted, or even impossible datapoints, that can be caused by measurement errors (Hammer, 1976). Data points with values that violate common sense (e.g., height 320cm) or seem to be out of

scale compared to other data (e.g., immense force measurements) are called outliers. They can be detected automatically based on their difference or distance from other datapoints and their center. For example, median absolute deviance (MAD) is suggested as a robust measure for detecting outliers (Leys et al., 2013) and for variable $\boldsymbol{x}_{:,j}$ is defined as

$$\text{MAD} = median(|x_{i,j} - median(\boldsymbol{x}_{:,j})|) \tag{5}$$

and values, e.g., 2.5 times away from MAD are outliers (Leys et al., 2013).

Outliers can be deleted (i.e., set to missing values) or replaced similarly as missing values. The whole data can also be transformed to resolve outliers (Kuhn et al., 2013). It is important to carefully consider the impacts of outlier handling on further analysis (Bramer, 2007). In many sports science applications it can be important to include outliers because, for example, an athlete producing distinct forces can have a higher risk of injury or an athlete with anomalous data can be a future talent to be identified.

Additionally, many of the data types in sports sciences can benefit from filtering and it is a common method used in biomechanical data analysis (Campbell et al., 2020). The purpose of filtering is to remove noise in data and, for example, to separate different components such as the gravitational and body acceleration components in accelerometer data (Yang and Hsu, 2010; Karantonis et al., 2006). A recent study with a scope in gymnastics skills, suggests that excess noise should be filtered out for research combining data from many participants but for individual purposes, such as daily monitoring and injury risk screening, no filter should be used to preserve extreme peak values (Campbell et al., 2020).

3.2.5 Imbalance handling

Class imbalance refers to data having clearly more observations from the other class, called minority, than the other, called majority, see Figure 2. Imbalance handling has become common in sports sciences and currently, for example, about half of sports injury prediction studies utilize some type of imbalance handling (Bullock et al., 2022; Van Eetvelde et al., 2021). Some common ways to deal with class imbalance are random under- and oversampling, the Synthetic Minority Oversampling Technique (SMOTE), and cost-sensitive learning. Random oversampling works by randomly duplicating observations in the minority class (e.g., injured players) while random undersampling randomly deletes observations from the majority class (e.g., non-injured players).

Similarly to random oversampling, SMOTE (Chawla et al., 2002) augments the minority class but instead of duplicating observations, it combines variables based on the k-nearest neighbors algorithm to create new synthetic observations. In cost-sensitive learning, misclassifying a minority observation has a higher cost than misclassifying a majority example. For example, in sports injury prediction or talent identification, not identifying potential injury or a talent is more harmful than incorrectly identifying some healthy athletes as injured or "normal" players as talents.

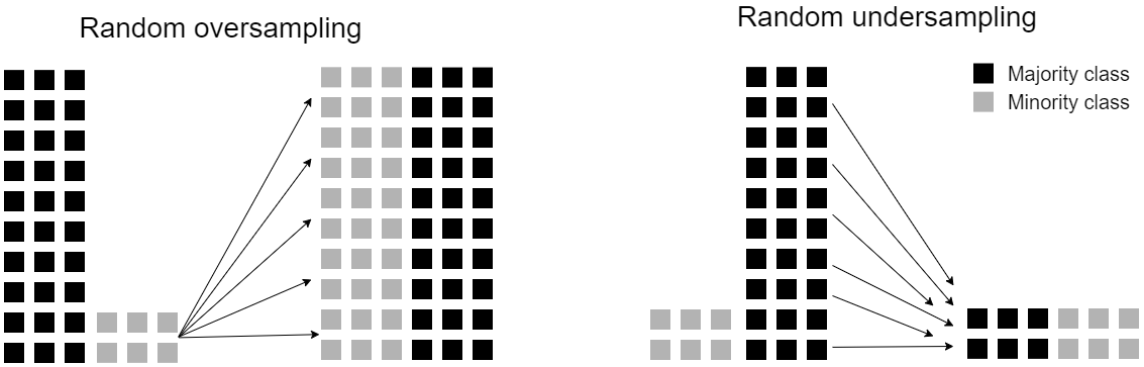


FIGURE 2 An example of class imbalance and the random over- and undersampling approaches

3.3 Supervised (predictive) machine learning

Supervised ML is based on labeled data with the goal of predicting or classifying the value of an outcome measure based on a number of input measures (Hastie et al., 2001).

3.3.1 Regression based methods

Linear regression is based on modeling the relationship between data observations $\boldsymbol{x}_{i,:}$ and a scalar response variable y_i and, loosely following the notation in (Hastie et al., 2001), the model can be written as

$$f(\boldsymbol{x}_{i,:}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{\beta}^T \boldsymbol{x}_{i,:}, \tag{6}$$

where $\boldsymbol{\beta}$ includes the regression coefficients and β_0 is called the intercept (or bias) term, and in the latter form we assume the observation vector $\boldsymbol{x}_{i,:}$ includes a constant term 1 to accommodate the intercept. This is most commonly fit with the least-squares methods (Hastie et al., 2001), where the coefficients are chosen by minimizing the residual sum of squares (RSS)

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_{i,:}))^2. \tag{7}$$

Logistic regression, on the other hand, includes a discrete response y_i, where the aim is to model the posterior probabilities of the classes with the logistic (or sigmoid) function $\sigma(z)$ that maps values to interval $(0, 1)$

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{8}$$

In the classification task $z = \boldsymbol{\beta}^T \boldsymbol{x}_{i,:}$. Research in this thesis focused on binary logistic regression and for simplicity, the definitions for binary classification are

introduced. Now the logistic regression can be written as

$$f(\boldsymbol{x}_{i,:}) = \frac{1}{1 + e^{-(\boldsymbol{\beta}^T \boldsymbol{x}_{i,:})}}. \tag{9}$$

In this case of two classes, namely c_1 and c_2, the probability of observation $\boldsymbol{x}_{i,:}$ belonging to c_1 is $p(\boldsymbol{x}_{i,:}; \boldsymbol{\beta})$ while the probability of belonging to class c_2 is $1 - p(\boldsymbol{x}_{i,:}; \boldsymbol{\beta})$. Futhermore, we can define that $y_i = 1$ for class c_1 and $y_i = 0$ for class c_2. The usual way to fit logistic regression is with the maximum likelihood. For the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \log p(\boldsymbol{x}_{i,:}; \boldsymbol{\beta}) + (1 - y_i) \log(1 - p(\boldsymbol{x}_{i,:}; \boldsymbol{\beta}))\} \tag{10}$$

the goal is to find coefficients $\boldsymbol{\beta}$ that maximize the function value. The problem can be transformed to minimization by multiplying with -1 and solved with, for example, the coordinate-descent algorithm that is based on successively minimizing each individual coordinate direction, i.e., dimension, while holding other values fixed (Hastie et al., 2001).

Regularized regression is an extension based on shrinking the coefficients $\boldsymbol{\beta}$ by imposing a penalty λ on their size. It can be utilized with any type of regression and essentially just has an additional penalty term at the end of the optimization problem. If the regularization is based on the L^1 norm, we end up with the Least Absolute Shrinkage and Selection Operator (LASSO) regression with the penalty term as

$$\lambda \sum_{j=1}^{p} |\beta_j|, \tag{11}$$

while for L^2 norm we have the Ridge regression with the penalty term as

$$\lambda \sum_{j=1}^{p} (\beta_j)^2. \tag{12}$$

The bigger the parameter λ is, the more coefficients are shrunk towards zero. With LASSO regression and large enough λ, some coefficients are shrunk to exactly zero, resulting into a sparse model with less variables. Therefore, regularized regression provides more interpretable models and information about variable importances.

3.3.2 Random forests

Random forest is a nonlinear classification and regression method that is very common in different fields such as medicine and bioinformatics (Boulesteix et al., 2012) and has been used widely in sports science studies as well, including sports injury prediction (Bullock et al., 2022; Van Eetvelde et al., 2021). Random forests are a combination of multiple decision tree predictors (Breiman, 2001a) that can be used for both regression and classification. Each of the decision trees provides

a separate prediction value and the final result of the forest is based on the average or major vote. Decision trees can capture complex relationships in data and if sufficiently deep, have relatively low bias but also include lot of noise and for these reasons, benefit from combining (Hastie et al., 2001). In a tree (see Figure 3), the branches represent rules based on which separation to further nodes is made. These rules are decided based on impurity measures. The nodes at the end of the tree, called leaf nodes, represent the prediction (or classification) of that tree.

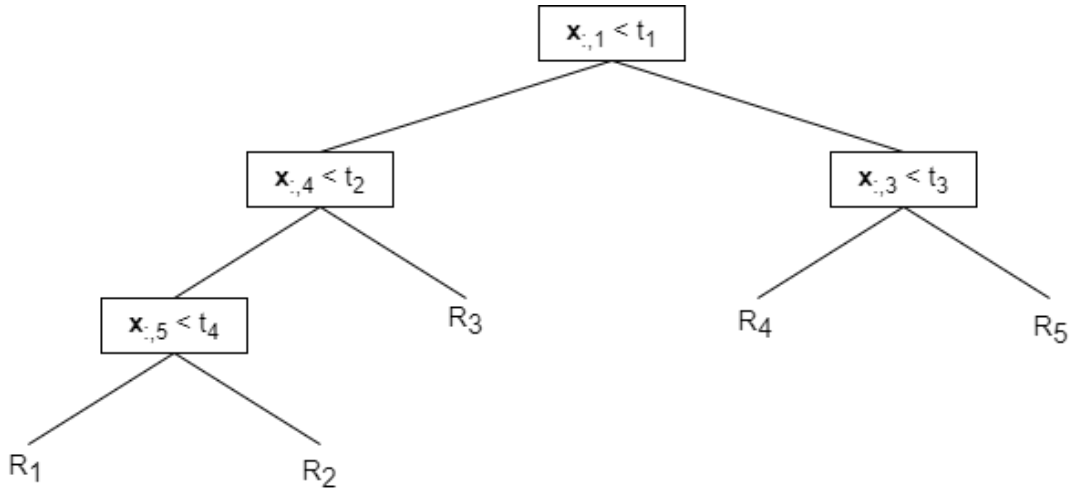


FIGURE 3 An example of a decision tree. At each step, the node is split based on a variable (e.g., $\boldsymbol{x}_{:,1}$) and a rule (e.g., $< t_1$), which are decided based on the impurity measure. The final nodes represent the prediction.

Loosely following the notation in (Hastie et al., 2001), for a regression tree, if we have a partition into M regions $R_1, \ldots, R_M$, the response y_i in region R_m is modeled as

$$f(\boldsymbol{x}_{i,:}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x}_{i,:} \in R_m), \tag{13}$$

and c_m is an average of y_i in region R_m

$$c_m = \frac{1}{N_m} \sum_{\boldsymbol{x}_{i,:} \in R_m} y_i, \tag{14}$$

where N_m is the number of observations in region R_m. For a regression tree, the splitting of nodes is based on squared-error node impurity measure (Hastie et al., 2009). For any subtree $T \subset T_0$ that can be obtained by splitting T_0, the impurity measure is then defined as

$$Q_m(T) = \frac{1}{N_m} \sum_{\boldsymbol{x}_{i,:} \in R_m} (y_i - c_m)^2. \tag{15}$$

For a classification tree with classes $k = 1, \ldots, K$, the impurity measure $Q_m(T)$ needs to be changed. First, for node m, the proportion of class k observations p_{mk} is defined as

$$p_{mk} = \frac{1}{N_m} \sum_{\boldsymbol{x}_{i,:} \in R_m} I(y_i = k). \tag{16}$$

And the most common impurity measures for classification are the Gini index

$$\text{Gini} = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}), \tag{17}$$

and cross-entropy

$$\text{Cross-entropy} = -\sum_{k=1}^{K} p_{mk} \log(p_{mk}). \tag{18}$$

A tree in a random forest is grown in the following way (Hastie et al., 2001):

Algorithm 1 Growing a random forest tree

for each terminal node of the tree
  while node_size $> n_{min}$
    Step 1: From the p variables, randomly select m variables.
    Step 2: Based on the impurity measure, pick the best variable, i.e., split-point among the m.
    Step 3: Split the node in two
  end while
end for

The minimum node size (n_{min}) and number of variables to sample at each split (m) are hyperparameters to be optimized.

Random forests provide inbuilt importance values for each variable, commonly with the out-of-bag estimates. In each tree, the values of jth variable are randomly permuted in the out-of-bag samples. Then the decrease in prediction ability of the whole forest is thought as the importance value of that variable. If there is no (considerable) decrease in model performance when a variable is permuted, we can assume it is not important for the task and vice versa, for a large decrease, the variable is considered important.

3.3.3 Support vector machines

Support vector Machines (SVMs) are commonly used for sports injury prediction (Bullock et al., 2022; Van Eetvelde et al., 2021). They are powerful and flexible classifiers (Kuhn et al., 2013) that work for linearly separable as well as overlapping or non-linearly separable classes. SVMs can be adapted for regression as well, but to support methodology in thesis articles and for simplicity, we focus on the case of classifying two classes. The goal is to find a hyperplane that best separates the classes from each other, or more technically maximizing the separation, or margin, of classes (Cortes and Vapnik, 1995).

For a linear SVM, let us first define a hyperplane as

$$\boldsymbol{\beta}^T \boldsymbol{x}_{i,:} + b = 0, \tag{19}$$

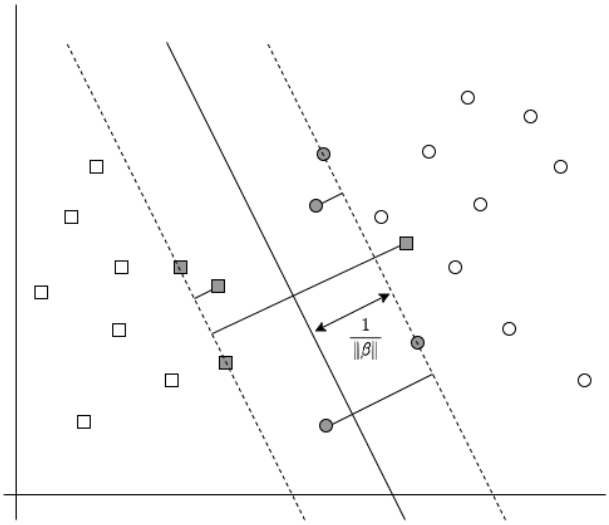


FIGURE 4 A soft margin SVM hyperplane in a linear, non-separable (class overlap) case. Grey points correspond to the support vectors. The margin is $1/\|\beta\|$, and points with positive slack values are shown with black lines. Adapted from (Zaki and Meira, 2014).

where $\boldsymbol{\beta}$ is a vector normal to hyperplane (scalars with at least one non-zero) and b is a constant that determines the offset, the minimization problem of SVM can be written as

$$\min_{\boldsymbol{\beta},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{N}\xi_i \qquad \text{s.t.} \begin{cases} y_i(\boldsymbol{\beta}^T\boldsymbol{x}_{i,:} + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \end{cases} \tag{20}$$

where C is a cost parameter (i.e., penalty for each misclassified observation) and ξ_is are the slack variables that are introduced to allow some observations to be on the wrong side of the margin. See Figure 4 for an illustration of a linear and non-separable classification.

For non-linearity, SVMs use kernel functions to transform data into higher dimension and make the classification more separable than in the lower dimension. We can map the features (i.e., variables) into higher dimensional space with a feature map $\phi(\boldsymbol{x}_i)$ and then utilize these mapped features instead of the original variables. Then a kernel function

$$K(\boldsymbol{x}_{i,:}, \boldsymbol{x}'_{i,:}) = \phi(\boldsymbol{x}_{i,:})\phi(\boldsymbol{x}'_{i,:}), \tag{21}$$

where $\boldsymbol{x}_{i,:}$ and $\boldsymbol{x}'_{i,:}$ are two observations, can be used to calculate the dot product in the original lower dimensional space. For example, Radial Basis Function (RBF) is a common kernel suited for many situations

$$K(\boldsymbol{x}_{i,:}, \boldsymbol{x}'_{i,:}) = e^{-\frac{\|\boldsymbol{x}_{i,:} - \boldsymbol{x}'_{i,:}\|^2}{2\sigma^2}}, \tag{22}$$

where σ is a hyperparameter impacting the decision boundary.

One-class SVM (Schölkopf et al., 2001) on the other hand is an anomaly detection method that first trains a SVM model on the observations from the normal

i.e., majority class and then predicts whether new observations belong to this normal region or not. For example, in this research, one-class SVM was used for talent identification where training was first done with the majority of players (i.e., not the talent class) and talented players were then identified from a test data including both talented and "normal" players. Following notation from (Schölkopf et al., 2001), the one-class SVM minimization problem is defined as

$$\min_{\boldsymbol{\beta},\boldsymbol{\xi},p} \frac{1}{2}\|\boldsymbol{\beta}^2\| - p + \frac{1}{vN}\sum_{i=1}^{N}\xi_i \qquad \text{s.t.} \begin{cases} \boldsymbol{\beta}^T\phi(\boldsymbol{x}_{i,:}) \geq p - \xi_i \\ \xi_i \geq 0, \end{cases} \tag{23}$$

where v is the upper bound of the fraction of outliers, p is an offset parameterizing the region, and ϕ is a feature map that transforms data points into a higher dimension.

3.4 Unsupervised machine learning

Unsupervised learning methods do not use labeled data but the goal is to describe the associations and patterns in a dataset (Hastie et al., 2001). There is no clear outcome measure for assessing the results as in supervised ML.

3.4.1 Clustering

Clustering is the (unsupervised) division of observations into groups (called "clusters") so that points in the same group are as similar to each other as possible and points in different groups are as dissimilar to each other as possible (Jain et al., 1999). Similarity of observations can be measured in many different ways and therefore the concept of 'cluster' is hard to define precisely (Estivill-Castro, 2002) and. A common measure for similarity is distance, which can also be defined with different measures, such as the Euclidean distance, correlation based distance or Ward's method.

A common way to divide clustering methods is into *partitional* and *hierarchical* where partitional clusters are non-overlapping and each observation belongs to one cluster and hierachical clusters can nested, i.e., clusters can have subclusters (Tan et al., 2007). Partitional clustering requires the user to define a number of clusters and then form the clusters by iteratively minimizing the chosen criteria. Clusters are represented by their prototypes (e.g., mean, median) and thus each cluster can be interpret based on its most representative point (Saarela, 2017). This research utilized hierarchical clustering, which in comparison to partitional clustering does not need a predefined number of clusters but only the measure of similarity. It is based on a hierarchy of representations where at the bottom each observation is their own cluster and each step upwards merges the most similar clusters together. Hierarchical clustering can be done in agglomerative (from bottom-up) or divisive matter (from top-down, starts with all observations in one cluster).

3.5 Model selection and assessment

Selecting and assessing predictive models is an important step that has to be done carefully to avoid data leakage.

Model selection refers to estimating the performance of different models and choosing the best one.

Model assessment refers to measuring the predictive ability (i.e., generalization performance) of the final selected model on unseen test data.

An optimal approach would be to split data into training, validation, and test data, where *training data* is used for fitting the model, *validation data* is used for estimating the fitted models and model selection, and *test data* is used for assessing the generalization performance of the final chosen model (Hastie et al., 2001). However, in many real-world problems and especially in sports sciences, data size is limited and other approaches more useful. Another approach is to only split training and test data, commonly 70% for training and 30% for testing (Vrigazova, 2021). Additionally, cross-validation (CV) is a standard approach (Vrigazova, 2021) common in sports sciences as well (Richter et al., 2021) and useful when data size is limited as the whole data can be utilized for training the model. K-fold CV (see Algorithm 2), for example, splits the data into k sets and trains k models using each sets as test data at a time. Stratified k-fold CV is a modification where the distribution of samples from different classes is preserved for both training and test data when data is split at step 1. It can be useful in the case of imbalanced classes or unevenly distributed variables. For example, data collected at different years or at different laboratories (with possible differences or noise) can be divided evenly through the training and test datasets to avoid under- or over-representation.

Algorithm 2 K-fold cross-validation

Step 1: Divide dataset into k parts
Step 2: Fit the model with $k - 1$ parts (training data) and leave one part out for model assessment (test data)
Step 3: Repeat k times until each part has been used as test data and combine estimates

It is beneficial to report metrics on both training and test data for assessment of, e.g., over- or underfitting. Moreover, due to the considerable variability in results with different test datasets (e.g., different folds), it is also important to report this through, for example, standard deviation or standard error (Hastie et al., 2001).

3.5.1 Preventing data leakage to test data

To prevent data leakage from model selection (i.e., hyperparameter tuning), it must be done without any guidance from the test data. Manual repetitions where any part of the analysis or hyperparameters is modified after seeing the performance on test data easily lead to data leakage and overfitting, often unconsciously and without realizing. Additional validation data for model selection solves the problem as test data is only used after model selection, but is unpractical with small data. In the absence of separate validation data, model selection is done on training data, e.g., with an inner (nested) CV where the training data is further split for training and validation. All suitable or interesting hyperparameter options can be included from the beginning, and the method optimizes itself inside the training phase.

Preprocessing of test and training data should also be done appropriately, and it is important to distinct two types of preprocessing in this context:

Type 1: Approaches where values from other observations do not directly affect the results, e.g., deleting observations with more than 50% of missing values, exclusion of uninteresting categories, and feature selection based on previous literature or expert-knowledge.

Type 2: Approaches where the data itself and values of other observations affect the results, e.g., data transformation like normalization and imputation or feature selection based on correlations or other metrics.

Type 1 preprocessing can be done before data splitting and does not cause data leakage. Type 2 preprocessing opens a risk for overfitting as preprocessing the whole data at once leads to data leakage. On the other hand, completely separate type 2 preprocessing of training and test data can lead to lower predictive performance due to different transformations (Van Eetvelde et al., 2021).

The proper way to handle preprocessing in prediction, is to think about the test data as future data, unavailable at training (Van Eetvelde et al., 2021) and use transformations calculated on training data to the test data. Similarly type 2 feature selection should be done on the training data alone. So for example with K-fold CV, normalization and imputation would be done separately inside each fold, first for the training data and then test data are normalized and imputed using coefficients and values estimated from the training data. Separate validation data should be treated similarly to test data to maximize generalization performance of the model.

3.5.2 Model assessment metrics

In addition, the model assessment metrics needs to be chosen carefully, especially in case of imbalanced datasets. In classification there are variety of measures that can be calculated based on the confusion matrix with some commonly used measure in sports injury prediction and talent identification being accuracy, sensitivity, specificity, precision and the F1-score (Van Eetvelde et al., 2021; Reyaz

et al., 2022). Again, focusing on the binary classification task, a confusion matrix (see Figure 5) visualizes a prediction result against the actual classes, where true positives (TP) are those correctly classified as positive, true negatives (TN) those correctly classified as negative, false positives (FP) those incorrectly classified as positive and false negatives (FN) those incorrectly classified as negative. P′ refers to the number of actual positive cases, N′ to actual negative cases while P_{pred} and N_{pred} refer to the number of predicted positive and negative cases.

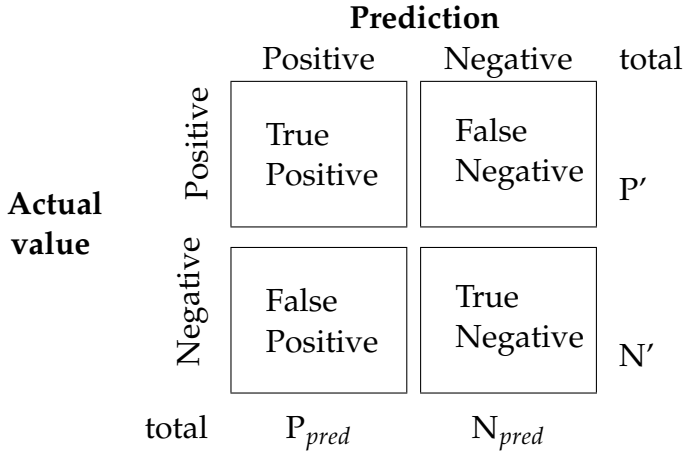


FIGURE 5 A confusion matrix of binary classification.

In the case of sports injury prediction, where injured players would be the positive cases in data, accuracy is simply the proportion of correctly classified players

$$\text{Accuracy} = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + FN + TN} \tag{24}$$

and would not work well with imbalanced classes as classifying everyone as non-injured would yield high accuracy. Sensitivity (also known as recall or true positive rate) measures the proportion of correctly classified positive observations (e.g., how many injured players are correctly classified as injured)

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{25}$$

while specificity (also known as true negative rate) measures the proportion of correctly classified negative (e.g., how many non-injured players are correctly classified as healthy)

$$\text{Specificity} = \frac{TN}{N} = \frac{TN}{TN + FP}. \tag{26}$$

Precision on the other hand measures the proportion of relevant observations among those classified as positive (e.g., how many of the players classified as injured are actually injured)

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{27}$$

False positive rate (FPR) measures the ratio of false negatives and number of actually negative (e.g., how many of the healthy players were classified as injured)

$$\text{FPR} = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - \text{Specificity} \tag{28}$$

and F1-score is the harmonic mean of sensitivity and precision

$$\text{F1} = \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}. \tag{29}$$

Additionally, Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is commonly used in sports injury prediction (Van Eetvelde et al., 2021), including the studies in this thesis. The ROC-curve is based on plotting true positive rate (sensitivity) against the false positive rate and the area under provides a summary metric of model performance. Area Under the Precision Recall curve (AUC-PR), on the other hand, works similarly but the curve is plotted as precision against recall (sensitivity). For regression, measures like Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i), \tag{30}$$

where $\hat{y}_i$ is the predicted response, or Root Mean Squared Error (RMSE, square root of MSE) can be used. With CV, it is important to choose how to combine metrics over different folds (Forman and Scholz, 2010). For example, for AUC an average over folds is recommended (over merging folds to a singe curve) while for F1-score recording the total number of true positives and false positives over the folds and then computing the final value is recommended (Forman and Scholz, 2010).

3.6 Confirmatory analysis

Confirmatory analysis is useful for handling uncertainty and avoiding chance results in predictive modeling. For this purpose, this thesis utilizes permutation tests (Combrisson and Jerbi, 2015) with repetitions as described in Algorithm 3. In permutation tests, a random reference model is trained by randomly shuffling the labels in the training phase and then the prediction performance is compared with the true model trained with actual labels. Then if the performance of the true model is (statistically significantly) better compared to the random model, the results are confirmed not being chance results or due to some noise in the data. Repetitions of the analyses assess the uncertainty and variation in results from random data splits. By repeating analyses, for example, a hundred times, an average of these represents a robust estimate of the true performance. For each repeated run, the fold division is the same for random and true models to allow fair pairwise comparison (e.g., by setting a seed number). Variation of results

across repetitions can be assessed through, for example, the standard deviation or range of results.

Algorithm 3 Confirmatory analysis with permutation tests and repetitions

Step 1: Set a seed number to initialize the (random) data split to training and test data

TRUE MODEL	RANDOM MODEL
Step 2 –	Step 2: Randomly shuffle the class labels
Step 3: Train the true model with actual labels	Step 3: Train a random model with the shuffled labels

Step 4: Save model and related metrics
Step 5: Repeat the whole process from Step 1 multiple times
Step 6: Compare the results from true and random models over the repetitions

Figure 6 summarizes the suggested predictive framework for producing robust and generalizable results that consider uncertainty and risk of chance results. Separate validation data is not included in the figure for simplicity but could be used for model selection if split at the data splitting phase.

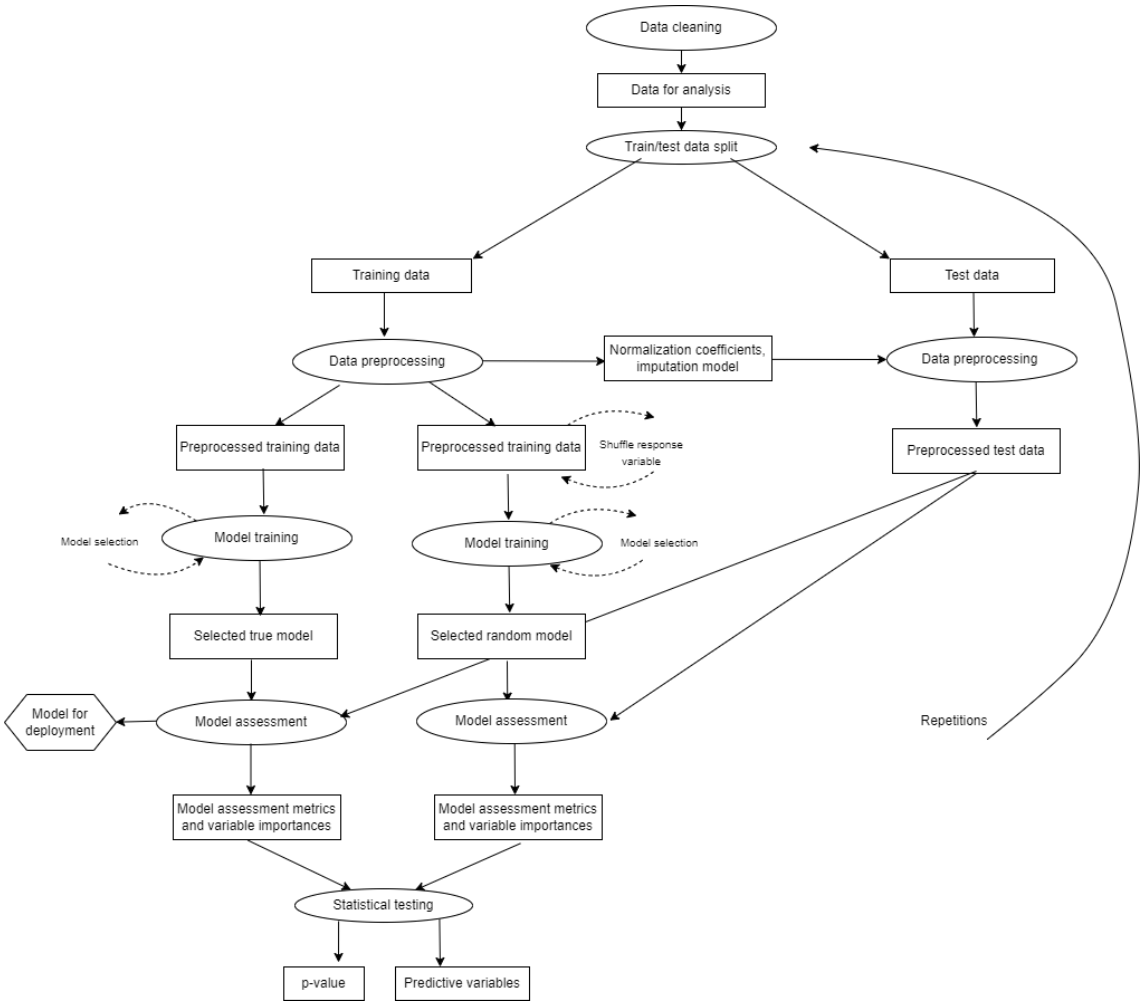


FIGURE 6 The predictive modeling framework. Model selection (i.e., hyperparameter tuning) is done at the training phase with, for example, inner CV with training data. Model deployment is discussed in Section 5.4 and a single trained model needs to be chosen for deployment.

4 OVERVIEW OF THE INCLUDED ARTICLES

The aim of this thesis was to assess the potential of predictive modeling and other ML methods for individual response within the area of sports science data. Furhermore, most common challenges and pitfalls in predictive ML and data in the field are considered as well as the use of these approaches for individualized injury prediction and talent identification. This chapter provides an overview of the included publications and their individual contribution to the research.

4.1 Article I: Talent identification in soccer using a one-class support vector machine

This article was published in 2019 in *International Journal of Computer Science in Sport*, 18(3).

Objectives

The objective of this study was to assess the potential of ML for talent identification in soccer. Moreover, the goal was to identify potential future elite players from the majority based on their physical and psychological test information as 14-year-old juniors.

Methods

The data in this study is from 4991 junior soccer players (age 12.41±1.53 years, range 8-18 year), collected for monitoring the development of young soccer players by *The training and research centre for Finnish football*. They performed physical tests (e.g., technical, speed, and agility) and self-assessment tests (e.g., perceived competence, tactical skills, and motivation) twice a year between 2011 and 2017. Included teams were the best of their age group in Finland. Players who had already signed a contract with an international academy ("academy players") were

labeled as talent category in the study. All of the academy players were boys and had performed the tests at the age of 14 and therefore, 14-year-old Finnish boys were selected for our analyses. Furthermore, as the age limit for signing a contract to an academy is 16 years, players born in 2003 or later were dropped out of the analysis as they might be future academy players but were too young to yet have a contract. Due to a significant proportion of values not missing at random (NMAR) in many variables, the data was further pruned into four separate representations; 1)*"phys large"* with 16 physical variables and N=951) *"phys+quest"* with 18 additional questionnaire variables and N=468 3) *"phys"* with the 16 physical variables and N=468 participants from *"phys+quest"*, and 4) *"quest"* with the 18 questionnaire variables and N=468 participants from *"phys+quest"*.

All datasets were extremely imbalanced with only 14 academy players in representation *"phys large"* and 10 in the rest of data representations. Therefore the problem was approached by anomaly detection with one-class support vector machine (SVM) that first models the normal region based on observations from the majority class, i.e., the non-academy players and then predicts whether new observations belong to this normal region or not. Sixteen different hyperparameter combinations were evaluated. To assess the generalization ability of our model, 10-fold CV was used and mean (and standard deviation) of AUC-ROC, AUC-PR, sensitivity and specificity values calculated for all data representations separately. A mean over all hyperparameter combinations was chosen in order to avoid chance findings.

Results and contribution to the whole

The highest mean AUC-ROC value was 0.763 for data representation *"phys large"*. For representations *"phys"*, *"phys+quest"*, and *"quest"*, the mean AUC-ROC values were 0.665, 0.643, and 0.585, respectively. While the estimated sensitivity of *"phys large"* representation was nearly 0.80, the estimated specificity of 0.614 shows that a large proportion of the players without an academy contract is "misclassified" into the class of potential academy players. So in a sense, the method recognizes player potential but there can be many reasons why these players have not (yet) signed an academy contract, such as young age, injury, choosing not to continue playing et cetera.

Methodologically, the study presents a useful approach for extremely imbalanced data (often the case in, e.g., talent identification) as well as for data with a lot of NMAR values (as common with longitudinal data in sports). Domain-wise it shows that players can already at the age of 14 show qualities in their measurable test data that can predict future potential. Furthermore the model provides an individual likelihood for elite potential that can be used to support professionals in coaching and player management.

Author's contributions

The author of this thesis is the main and corresponding author of this journal article. She preprocessed the data, run all analyses, produced the tables and wrote the majority of the article. The analysis approach was invented by the co-authors of the article and its design as well as interpretation of results was done together by all authors.

4.2 Article II: New machine learning approach for detection of injury risk factors in young team sport athletes

This article was published in 2021 in *International Journal of Sports Medicine*, 42(02), 175-182.

Objectives

The objective of this article was to present a framework for detecting variables with predictive power on sports injuries in a data-driven manner. It also coherently describes differences of predictive and explanatory methods within the field of sports injuries and discusses common pitfalls in sports injury prediction and how to tackle them. Interpretability and generalization ability of models and results is highlighted in order to produce trustworthy results that have practical benefit. A secondary purpose was to assess differences between linear and non-linear methods for the task.

Methods

The data was three-dimensional motion analysis and physical data from 314 young basketball and floorball players (48.4% males, 15.72±1.79 years, 173.34±9.14cm, 64.65±10.4kg) collected at the *UKK institute* between 2011 and 2015. The focus was on moderate to severe (unable to fully participate in training or match play for at least 8 days ((Fuller et al., 2007)) acute non-contact knee and ankle injuries (N=57). Altogether 54 variables were included in classification after careful selection by a group of experts in sports medicine and exclusion of variables with more than 50% of missing data.

Two commonly used ML methods that provide information about the variable importance in prediction were selected, namely random forests and L1- regularized logistic regression, were used to predict the injuries. Generalization ability of the models was assessed with 10-fold CV and test performance with the Area Under the Receiver Operating Characteristics Curve (AUC-ROC). Due to large variation in results from different K-fold splits, the analysis was repeated a hundred times and results averaged over all data splits. Permutation tests were used to confirm the significance of the results as well as to detect significantly

consistent injury predictors. The class labels (injured/healthy) in training data were randomly shuffled for training of a "random" reference model. Then by comparing results of the true and random model, we can confirm the achieved performance or important variable is significant and not observed by chance. Significantly consistent injury predictors were identified by comparing the variable importance values of the true and random model from the hundres repetitions with Wilcoxon signed-rank test.

Results and contribution to the whole

Twelve and twenty consistent injury predictors were suggested by random forests and the L1-regularized logistic regression, respectively. The following ten variables were suggested as consistent injury predictors ($p<0.01$) by both models: sex, body mass index, hamstring flexibility (non-dominant leg), KT1000 (dominant leg), hamstring flexibility (dominant leg), medial knee displacement (dominant leg), height, ankle (plantar) flexion at IC (dominant leg), leg press one repetition maximum (1RM), and knee valgus at IC (dominant leg). Predictive accuracy of the models remained low and was similar between the linear (regression AUC=0.65) and non-linear (random forests AUC=0.63) models.

Methodologically, this study offers an approach for finding new hypotheses for injury risk factors and for confirming the predictive power of risk factors detected in previous explanatory studies, while taking into account the most common pitfalls in predictive modeling. Great benefit of the presented approach is that consistent injury predictors can be detected even from data with weak phenomena. Therefore, it can be useful with small datasets (that do not necessarily possess high predictive power or strong phenomena) as often is the case in sports science and sports medicine. Domain-wise, the identified consistent injury predictors can guide the planning of future (predictive and explanatory) sports injury studies. Both, the information about these injury predictors as well as the individual injury probabilities from the trained model can be used for more individualized injury prevention and prediction in the future.

Author's contributions

The author of this thesis is the main and corresponding author of this journal article. She preprocessed the data, run all analyses, produced all tables and figures and wrote the majority of the article. The analysis approach was invented by the co-authors of the article and its design as well as interpretation of results was done together by all authors.

4.3 Article III: Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes

This article was published in 2022 in *The American Journal of Sports Medicine*, 50(11), 2917-2924.

Objective

The purpose of this article was to carefully investigate the predictive potential of multiple predictive ML methods on a large set of risk factor data for anterior cruciate ligament (ACL) injury. The approach is based on Article II and was extended by increasing the hypothesis space with more ML methods as well as preprocessing techniques for handling class imbalance in the data. Moreover, it follows up the previous article on clarifying the distinction between predictive and explanatory approaches and how these have been confused in previous literature in the field. Most common pitfalls in sports injury prediction with examples from the literature are also further discussed and considered in the methodology.

Methods

The data consist of a comprehensive ACL injury screening test battery including demographic, neuromuscular, biomechanical, anatomic, and genetic ACL injury risk factors collected at the *Norwegian School of Sport Sciences* between the years 2007 and 2015. Participants were elite female soccer (N=451) and handball (N=429) players (21±4years, 170±6cm, 66±8kg), making it the largest ACL injury data collected to date.

After preprocessing (e.g., players with more than 50% of missing data or noncontact and indirect contact injuries excluded) the dataset used for injury prediction had 791 players with 60 ACL injuries and 283 variables. Four common methods, random forest, L2-regularized logistic regression, and support vector machines (SVMs) with both linear and nonlinear kernel, were chosen. Data imbalance handling with random undersampling, The Synthetic Minority Oversampling Technique (SMOTE) as well as class weight vector in the training phase were experimented. Generalization ability of the models was assessed with 5-fold CV and test performance with the Area Under the Receiver Operating Characteristics Curve (AUC-ROC). Permutation tests, a hundred repetitions and Wilcoxon signed-rank test were used to confirm the significance of the results, similarly as in the above study.

Results and contribution to the whole

Linear SVM without any imbalance handling achieved the highest mean AUC-ROC value of 0.63. The predictive ability was relatively consistent between the

methods and significantly higher (P< 0.001) with the real responses than with the random models. With all of the classifiers, there was a large variation in performance across the repetitions, highlighting the effect of data split and need for repetitions. Additionally, class imbalance handling did not improve the prediction results but they remained similar or even slightly worse.

Methodologically, this study presents how to avoid most common pitfalls in predictive ML in sports sciences. Domain-wise it assesses the predictive potential of a large ACL injury data and suggests that one single screening test dataset is potentially not enough for ACL injury prediction. Moreover, individual injury probabilities of the trained model can be used for more individualized injury prevention and prediction in the future.

Author's contributions

The author of this thesis is the main and corresponding author of this journal article. She preprocessed the data, run all analyses, produced the tables and wrote the majority of the article. The analysis approach idea was based on Article II and its design as well as interpretation of results was done together by all authors.

4.4 Article IV: A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns

This article was published in 2020 in *Scandinavian Journal of Medicine and Science in Sports*, 30(4), 732-740.

Objectives

The primary purpose of this article was to examine if unsupervised ML can be used to discover similar gait patterns from 3D kinematic data of injured and healthy runners. A secondary purpose was to analyze the 3D running kinematics between the subgroups to better understand differences in running gait patterns.

Methods

The data in this study consisted of 3D running kinematics of 291 runners (255 injured, 146 females, 39.51±11.21 years). Injuries were grouped by location with: 72 knee, 58 ankle/foot, 51 hip/pelvis, 42 thigh, and 39 lower leg (shin) injuries and twenty-five individuals were confirmed as injury free for at least six months prior to data collection. Each subject's running gait pattern was described by 62 extracted kinematic (e.g., peak knee flexion and adduction angles, heel strike angle) and functional (e.g., step width, vertical oscillation, stride rate and length) variables. A PCA was performed to reduce multi-collinearity between biomechanical variables and a subset of principal components (PCs) that explained 80% of the total variance in the dataset were selected.

Runners were clustered into subgroups of homogeneous gait patterns with hierarchical cluster analysis (HCA). Number of clusters was determined by on a stopping rule (a large percentage decrease in the coefficient followed by a plateau) and confirmed by visual inspection of the dendrogram. Subgroups were compared with univariate analysis of variance (ANOVA) and the injury distribution of the formed subgroups was assessed with the adjusted Rand index, that measures the similarity between two different clusterings of the same data.

Results and contribution to the whole

The first 16 PCs, explaining 80.98% of the total variance, were used for clustering. Five subgroups with specific gait patterns were discovered from the data but despite being distinct, the population of injured and healthy runners was randomly scattered among those. This was confirmed with the very low Rand index score of r = 0.012 when comparing the cluster partition and the original injury groups.

Based on these results, the location of injury seems not to be related to specific running kinematic patterns. This suggest that the traditional method of creating and analyzing subgroups of subjects based on a pre-defined injury might not consider variance of gait biomechanical patterns that exists independent of the injury location. Additionally, prediction of injuries, (solely) based on specific kinematic gait patterns, is not supported. Therefore, an initial additional step to segment subjects according to gait patterns could be beneficial to future biomechanical investigations as well as ML approaches.

Author's contributions

The author of this thesis is the main and corresponding author of this journal article. She preprocessed the data, run the analyses, produced tables and figures, and wrote the majority of the article. Analysis design and interpretation of results was done together by all authors.

5 DISCUSSION AND CONCLUSION

The use of ML methods has recently generalized in sports sciences but many common pitfalls lurk researchers (Richter et al., 2021; Riley, 2019). Generally small sample sizes together with attractiveness and accessibility of ML methods without proper knowledge lead to faulty models and results with improper interpretations. For example, Table 1 summarizes how ML approaches in most existing studies are flawed in sports injury prediction. It is critical that researchers are aware of the risks related to the use of ML and that there are clear standards and robust procedures for how to perform and report ML studies in sports sciences. Answering the urgent need, this thesis provides guidelines on how to properly perform and report predictive ML studies in the field of sports science and introduces a framework for producing robust and trustworthy results.

Moreover, this thesis assesses the potential of ML for more individual injury prediction and talent identification, utilizing large, contemporary datasets. Through the included articles, advances are achieved for predicting ACL injuries, recognizing predictive knee and ankle injury risk factors, assessing how to use ML with talent identification in sports settings as well as utilizing unsupervised methods to discover novel and useful information and patterns from running injury data. The approaches developed and used in this research can be utilized similarly in many other tasks and domains as well.

5.1 Relation to previous work

The results from the predictive studies in this thesis are among only a handful of others in sports injury prediction and talent identification that utilize robust approaches, i.e., no signs of data leakage and of these only a few also consider uncertainty in results and the risk of chance results at least to some degree (see Table 1). There seem to be no previous sports science studies that utilize confirmatory analyses, such as permutation tests, to exclude the risk of chance results. The other sports injury studies have achieved similar or slightly higher predictive

ability as Articles II and III (Colby et al., 2018; Karuc et al., 2021; Carey et al., 2018; Rossi et al., 2018; Whiteside et al., 2016). Rommers et al. (2020) achieved promising results (85% sensitivity and precision) predicting acute and overuse injuries in 734 elite youth soccer players. Their study differs from the others in its age range (11.7 ± 1.7 years) and they reported that the five most important variables in prediction were anthropometric measures, indicating injuries might be easier to predict among teenagers during the growth spurt. The results from Articles II and III add important information to the body of knowledge about sports injury prediction, such as assessing the sufficiency of two large and contemporary datasets and importance of the variables in prediction. Moreover, Article II introduces an approach that can be used for finding variables most predictive of injury and thus for generating new hypotheses in a data-driven manner. In talent identification, Article I is among the first studies utilizing predictive modeling. Moreover, it is the first to predict future athlete performance based on data collected at young age, an approach that can be very helpful for offering suitable resources and training for performance development. It also provides information about the most important variables and data properties for talent identification in soccer.

Article IV is the first study to assess homogeneous gait patterns across a population of runners with a wide variety of injuries and healthy runners. Previous studies have found similarities in kinematics across different running injuries (Bramah et al., 2018), but Article IV suggests there are subgroups with similar running patterns across both injured and healthy runners. Contrary to some previous results and practices, the results imply that certain running styles do not increase risk of injury and that grouping based on injury location might not be best practice, which can be important to consider in future predictive and unsupervised ML studies.

5.2 Potential of machine learning and current datasets for individual response (RQ1)

All included articles provide some information in the individual response context. The predictive models trained in Articles II and III can be utilized for calculating an individual sports injury risk based on the provided class probabilities of the model. However, the low performance of these models (AUCs 0.65 and 0.63) suggest the current data is not suitable or sufficient enough for injury prediction, at least by itself, and further research can utilize the approach with more adequate data. In a similar fashion, the model trained in Article I could be used to asses individual chances on future elite soccer career success. Based on the results, it is possible to predict and provide an individual probability of becoming an academy player based on test data collected at already 14-years-old juniors with an 80% sensitivity. A similar approach can be used for talent identification in other sports and levels of play as well as other areas in life. Moreover, the re-

sults show that larger data achieved higher predictive performance but utilizing a more versatile set of variables (i.e., including both physical tests and questionnaire data) provided higher sensitivity.

Article IV, on the other hand, confirms that injury prevention and rehabilitation strategies should be individualized and not based on certain running patterns. Among the large group of injured and healthy runners, no associations between specific running patterns and injury locations were detected. Therefore, the results contradict and dispute previous studies and assumptions that certain running style would increase the risk of certain running injuries or that specific changes in running style would prevent or help rehabilitate injuries.

5.3 Predictive machine learning pitfalls in sports sciences (RQ2)

All three articles utilizing predictive modeling methods (Articles I, II, and III) clarify the distinction between predictive and explanatory analyses, including examples from sports injury prediction and talent identification. Moreover, most common pitfalls in predictive ML research in sports sciences were further discussed and solutions introduces in Articles II and III. In addition to selecting separate and appropriate test data, the pitfalls to be avoided include: avoidance of data leakage (e.g., feature selection, preprocessing, hyperparameter tuning and model selection) and assessment of uncertainty and chance results (e.g., random test data, noise in data, weak phenomena, and small, imbalanced, or high-dimensional data). Articles II and III utilize the presented framework with permutation tests and repetitions to overcome these pitfalls and for robust, reproducible results. Article III reports performance metrics for the training data (and the trained random model as well), including variance across the repetitions. Article II reports results for the training data and as the focus was on detecting predictive variables, variance of results is assessed through the consistency of variables selected as important by the models.

Article I did not include confirmatory analysis but the risk of chance results is considerably lower with the one-class SVM model as it is trained only on the majority class while all minority examples (i.e, academy players/talents) were always included in the test data. Furthermore, the negligible variation reported in results across the folds further confirms reliability of results and the relatively high predictive performance suggest models learned true phenomena and not only noise.

5.4 Model deployment

The studies in this thesis, as well as majority of (if not all) existing research, focuses on model assessment and development of accurate models. In practice,

however, the final goal would be to be able to use a predictive model for a practical task. In the sports injury prediction context, for example, a clinician needs a single ready-to-use trained model which would output individual probabilities and information about important variables when given new athlete data. However, deployment can be very time-consuming and challenging (Baier et al., 2019). While research studies should train multiple models (due to, e.g., repetitions or CV data splits) a single model needs to be selected for deployment. So, from deployment point of view, training a single model with randomly split train and test data at once would be most simple but then the generalization performance would be questionable due to lot of uncertainty. Larger data and established ML practices are the first steps towards successful deployment, but more research is still needed to decide the best deployment practices. Whichever way the model for deployment is selected, it should not be blindly trusted in practical use, especially if trained on small data collected from a short period. Moreover, after deployment, the model should be continually monitored and improved based on user input and input data.

5.5 Model interpretability

Most of the existing ML research uses the methods as black-boxes (Rudin, 2019; Lipton, 2018). This means that even though the method is able to predict or classify something well, the reasons behind how it does this are hidden. Interpretability has been considered highly important especially in medicine (Bellazzi and Zupan, 2008), with ethical reasons clearly essential as well. ML approaches should only be used as support for human experts and therefore the more reasoning a method can provide behind its suggestions, the more beneficial it will be for the end-users. In the sports injury prediction context, for example, in addition to calculating a injury risk for an athlete, we might be interested in the variables which are causing the injury risk. With this information a coach or a team physician can try to minimize the risk by focusing on these most important variables. In this thesis, random forests and regularized logistic regression used in Articles II and III, for example, are interpretable (at least to some degree) as they provide information about the most important variables in prediction.

5.6 Limitations and future research

In ML, the amount of methodological approaches available is immense and naturally only a limited portion of these can be utilized in this research. Further research should assess potential of different methods as well as further tuning of hyperparameters and the data preprocessing techniques. Moreover, the research in this thesis was limited on previously collected data and variables. The 3D mo-

tion analysis data, for example, despite being a gold standard in human movement research, can have low reproducibility due to marker placements (Gorton III et al., 2009) and low generalization to real world as it is collected in laboratory settings. For example in sports injury prediction, future research is needed to detect the most relevant data sources and variables for accurate prediction. Continuous time-series data of, for example, training loads from GPS or accelerometers could capture important information in the context of injury prevention or performance development, as the results from few previous studies suggest (Carey et al., 2018; Colby et al., 2018; Rossi et al., 2018; Thornton et al., 2017; McCullagh and Whitfort, 2013). Larger datasets will benefit future ML research in sports (Richter et al., 2021) and open new possibilities. It would allow training of multiple, more individual models based on subsamples as well as the usage of deep learning. Moreover, continuous data that includes measurements of individual response, such as training loads and HR, would also open new ML possibilities, particularly for research on individual response. Additionally, a more balanced class distribution, where possible, might improve the performance of predictive models as some previous sports injury studies suggest (Rommers et al., 2020; Tamimi et al., 2021; Whiteside et al., 2016).

Importantly, there should be clear standards for how to perform and report ML studies in sports sciences (Riley, 2019; Richter et al., 2021) and the research in this thesis provides fundamental guidelines to build on. Following the guidelines and utilizing the introduced framework, more reliable, robust, and comparable results can be achieved leading to more impactful research. Moreover, these multidisciplinary studies should always include experts from both fields (i.e., ML and sports science) and the ML tools must be understood by those using them (Riley, 2019).

YHTEENVETO (SUMMARY IN FINNISH)

Liikuntatieteissä data on usein haastavaa ja määrältään rajallista, mikä yhdistettynä koneoppimismenetelmien houkuttelevuuteen ja saatavuuteen johtaa helposti virheellisiin ennustusmalleihin, tuloksiin ja johtopäätöksiin. On kriittisen tärkeää, että tutkijat tuntevat koneoppimismenetelmien ja datan oikeaoppisen käytön ja ovat tietoisia näihin liittyvistä riskeistä. Tämän tutkimuksen ensimmäinen tavoite on ohjeistaa ennustavan koneoppimisen oikeaoppista hyödyntämistä ja raportoimista erityisesti liikuntatieteissä. Toisena tavoitteena on tutkia voidaanko ennustavan koneoppimisen avulla tuottaa yksilöllisempää tietoa kuin perinteisillä tilastomenetelmillä urheiluvammojen ennustamisen ja lahjakkuuksien tunnistamisen sovellusalueilla.

Ennustava mallintaminen viittaa ennusteiden (tai luokitteluiden) tekemiseen datan pohjalta ja sen riskit liittyvät muun muassa ennustusmallien yleistettävyyteen, ylioppimiseen sekä sattumatuloksiin. Mallin yleistettävyys viittaa sen ennustuskykyyn erillisellä (mallin aiemmin näkemättömällä) datalla ja sen asianmukainen mittaaminen on äärimmäisen tärkeää luotettavien tulosten saavuttamiseksi. Ylioppiminen puolestaan tarkoittaa joko mallin liiallista sovittumista sen opettamisessa käytettyyn dataan, jolloin sen yleistettävyys uudelle datalle kärsii tai liiallista sovittumista käytettyyn testidataan. Sattumatulos viittaa tuloksiin jotka saavutetaan sattumalta, esimerkiksi suotuisan datajaon seurauksena tai datassa olevan ylimääräisen kohinan oppimisen seurauksena. Yksi yleisimmistä virheistä liikuntatieteissä on datan vuotaminen eli informaatiota testidatasta vuotaa mallin opettamiseen jostain kohtaa analyysiä. Esimerkiksi urheiluvammatutkimuksessa suurimmassa osassa ennustustutkimuksista on selkeitä datavuodon lähteitä ja seurauksena mallien yleistyvyys on kyseenalaista.

Artikkelit I, II ja III hyödyntävät ennustusmenetelmiä ja korostavat niihin liittyviä riskejä sekä kuvaavat menetelmien ja datan oikeaoppista hyödyntämistä. Artikkelissa I tunnistetaan lahjakkuuksia 14-vuotiaiden jalkapalloilijoiden testidatasta, saavuttaen noin 76% kokonaistarkkuus tulevaisuudessa akatemiaan pelisopimuksen sopineiden pelaajien tunnistamisessa. Tulosten perusteella tulevaisuuden eliittipelaajien tunnistaminen jo 14-vuotiaana kerätyn datan perusteella on mahdollista jalkapallossa ja lähestymistapaa voidaan hyödyntää muun muassa lahjakkuuksien tunnistamiseen muillakin sovellusaloilla.

Artikkelit II ja III keskittyvät urheiluvammojen ennustamiseen sekä ennustusvoimaisten muuttujien tunnistamiseen joukkueurheilijoiden datasta. Ennustusvoimaltaan tulokset (kokonaistarkkuus 65% ja 63%) ovat samaa luokkaa aiempien oikeaoppisesti toteutettujen urheiluvammoja ennustavien tutkimusten kanssa ja korostavat urheiluvammojen ennustamisen haastavuutta. Artikkelin II päätavoitteena on tunnistaa ennustusvoimaisia muuttujia ja kehitettyä lähestymistapaa voidaankin käyttää tarkoitukseen muissakin tutkimuksissa.

Artikkelissa IV tutkitaan terveiden ja loukkaantuneiden juoksijoiden juoksutyylejä ja niiden yhteyksiä vammojen sijaintiin ohjaamattoman koneoppimisen avulla. Tulokset vahvistavat yksilöllisempien lähestymistapojen tarvetta ja ovat

tärkeää tietoa myös tuleville ennustustutkimuksille. Tietyt juoksutyylit eivät ole tuloksissa yhteydessä tiettyihin vammoihin, joten vammojen ehkäisy sekä kuntouttaminen tulisi suunnitella yksilöllisesti, eikä esimerkiksi juoksutyyliin tai vamman sijaintiin perustuen.

Tutkimuksen päätuloksena kehitetään lähestymistapa joka arvioi kattavasti ja luotettavasti mallien yleistyvyyttä ja huomioi koneoppimiseen liittyvät riskit. Väitöskirjassa esitellään kattavasti eri analyysin vaiheet ja datan oikeaoppinen käsittely erityisesti liikuntatieteisiin keskittyen. Huomioimalla riskit ja huolellinen raportointi voidaan varmistaa tulevien tutkimusten luotettavuus, robustisuus ja vertailukelpoisuus ja saavuttaa vaikuttavampia tuloksia.

Lisäksi tutkimuksessa tuotetaan tärkeää tietoa yksilöllisemmän tiedon saavuttamiseksi koneoppimisen avulla. Kehitettyjä malleja voidaan hyödyntää esimerkiksi yksilöllisen urheiluvammariskin tai jalkapallomenestyksen laskemiseksi. Urheiluvammojen tunnistamisessa nykyisten datojen ennustusvoima jäi kuitenkin käytännön kannalta matalaksi, mutta hyödyllistä tietoa tulevia tutkimuksia varten tuotetaan muun muuassa uusien hypoteesien muodostamiseksi sekä datan ja mittausten uudelleen suuntaamiseksi tehtävän kannalta relevanttien tekijöiden löytämiseksi.

REFERENCES

Juha P Ahtiainen, Janne Sallinen, Keijo Häkkinen, and Elina Sillanpää. Inter-individual variation in response to resistance training in cardiometabolic health indicators. *Scandinavian Journal of Medicine & Science in Sports*, 30(6):1040–1053, 2020.

Naomi Altman and Martin Krzywinski. Points of significance: Association, correlation and causation. *Nature methods*, 12(10), 2015.

Francisco Ayala, Alejandro López-Valenciano, Jose Antonio Gámez Martín, Mark De Ste Croix, Francisco J Vera-Garcia, Maria del Pilar Garcia-Vaquero, Iñaki Ruiz-Pérez, and Gregory D Myer. A preventive model for hamstring injuries in professional soccer: Learning algorithms. *International journal of sports medicine*, 40(05):344–353, 2019.

R Bahr and I Holme. Risk factors for sports injuries – a methodological approach. *British journal of sports medicine*, 37(5):384–392, 2003.

Lucas Baier, Fabian Jöhren, and Stefan Seebacher. Challenges in the deployment and operation of machine learning in practice. In *ECIS*, volume 1, 2019.

Joseph Baker, Nick Wattie, and Jörg Schorer. A proposed conceptualization of talent in sport: The first step in a long and winding road. *Psychology of Sport and Exercise*, 43:27–33, 2019.

Donald Barron, Graham Ball, Matthew Robins, and Caroline Sunderland. Artificial neural networks and player recruitment in professional soccer. *PloS one*, 13 (10):e0205818, 2018.

Gustavo EAPA Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17 (5-6):519–533, 2003.

Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2): 81–97, 2008.

Lauren C Benson, Carlyn Stilling, Oluwatoyosi BA Owoeye, and Carolyn A Emery. Evaluating methods for imputing missing data from longitudinal monitoring of athlete workload. *Journal of Sports Science & Medicine*, 20(2):188, 2021.

Claude Bouchard and Tuomo Rankinen. Individual differences in response to regular physical activity. *Medicine and science in sports and exercise*, 33(6 Suppl): S446–51, 2001.

Claude Bouchard, Steven N Blair, Timothy S Church, Conrad P Earnest, James M Hagberg, Keijo Häkkinen, Nathan T Jenkins, Laura Karavirta, William E Kraus,

Arthur S Leon, et al. Adverse metabolic response to regular exercise: is it a rare or common occurrence? *PloS one*, 7(5):e37887, 2012.

Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

Christopher Bramah, Stephen J Preece, Niamh Gill, and Lee Herrington. Is there a pathological gait associated with common soft tissue running injuries? *The American journal of sports medicine*, 46(12):3023–3031, 2018.

Max Bramer. *Principles of data mining*, volume 180. Springer, 2007.

Ulf Brefeld and Albrecht Zimmermann. Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, 31:1577–1579, 2017.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001a.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231, 2001b.

Garrett S Bullock, Joseph Mylott, Tom Hughes, Kristen F Nicholson, Richard D Riley, and Gary S Collins. Just how confident can we be in predicting sports injuries? a systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports medicine*, 52(10):2469–2482, 2022.

Rory P Bunker and Fadi Thabtah. A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1):27–33, 2019.

Rhiannon A Campbell, Elizabeth J Bradshaw, Nick Ball, Adam Hunter, and Wayne Spratford. Effects of digital filtering on peak acceleration and force measurements for artistic gymnastics skills. *Journal of sports sciences*, 38(16): 1859–1868, 2020.

David L Carey, K Ong, Rod Whiteley, Kay M Crossley, Justin Crow, and Meg E Morris. Predictive modelling of training loads and injury in australian football. *International Journal of Computer Science in Sport*, 17(1):49–66, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Hua-Tsung Chen, Chien-Li Chou, Tsung-Sheng Fu, Suh-Yin Lee, and Bao-Shuh P Lin. Recognizing tactic patterns in broadcast basketball video using player trajectory. *Journal of Visual Communication and Image Representation*, 23(6):932–947, 2012.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, pages 1–8, 2022.

Marcus J Colby, Brian Dawson, Peter Peeling, Jarryd Heasman, Brent Rogalski, Michael K Drew, and Jordan Stares. Improvement of prediction of noncontact injury in elite australian footballers with repeated exposure to established high-risk workload scenarios. *International journal of sports physiology and performance*, 13(9):1130–1135, 2018.

Etienne Combrisson and Karim Jerbi. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods*, 250:126–136, 2015.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Antonio I Cuesta-Vargas, Alejandro Galán-Mercant, and Jonathan M Williams. The use of inertial sensors system for human motion analysis. *Physical Therapy Reviews*, 15(6):462–473, 2010.

Pratap Dangeti. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

Louise de Lannoy, John Clarke, Paula J Stotz, and Robert Ross. Effects of intensity and amount of exercise on measures of insulin and glucose: analysis of inter-individual variability. *PloS one*, 12(5):e0177095, 2017.

Carlos Dietrich, David Koop, Huy T Vo, and Cláudio T Silva. Baseball4d: A tool for baseball game reconstruction & visualization. In *2014 IEEE conference on visual analytics science and technology (VAST)*, pages 23–32. IEEE, 2014.

Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

Stef Feijen, Thomas Struyf, Kevin Kuppens, Angela Tate, and Filip Struyf. Prediction of shoulder pain in youth competitive swimmers: the development and internal validation of a prognostic prediction model. *The American Journal of Sports Medicine*, 49(1):154–161, 2021.

Reed Ferber, Sean T Osis, Jennifer L Hicks, and Scott L Delp. Gait biomechanics in the era of data science. *Journal of biomechanics*, 49(16):3759–3761, 2016.

Karl B Fields, Jeannie C Sykes, Katherine M Walker, and Jonathan C Jackson. Prevention of running injuries. *Current sports medicine reports*, 9(3):176–182, 2010.

Daniel Tik-Pui Fong and Yue-Yan Chan. The use of wearable inertial motion sensors in human lower limb biomechanics studies: A systematic review. *Sensors*, 10(12):11556–11565, 2010.

George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57, 2010.

Colin W Fuller, Michael G Molloy, Christian Bagate, Roald Bahr, John HM Brooks, Hilton Donson, Simon PT Kemp, Paul McCrory, Andrew S McIntosh, Willem H Meeuwisse, et al. Consensus statement on injury definitions and data collection procedures for studies of injuries in rugby union. *British journal of sports medicine*, 41(5):328–331, 2007.

Darrell J Gaskin and Patrick Richard. The economic costs of pain in the united states. *The journal of pain*, 13(8):715–724, 2012.

Brian Godman, Alexander E Finlayson, Parneet K Cheema, Eva Zebedin-Brandl, Inaki Gutiérrez-Ibarluzea, Jan Jones, Rickard E Malmström, Elina Asola, Christoph Baumgärtel, Marion Bennie, et al. Personalizing health care: feasibility and future implications. *BMC medicine*, 11:1–23, 2013.

George E Gorton III, David A Hebert, and Mary E Gannotti. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait & posture*, 29 (3):398–402, 2009.

Joel Grus. *Data science from scratch: first principles with python*. O'Reilly Media, 2019.

Michael Hammer. Error detection in data base systems. In *Proceedings of the June 7-10, 1976, national computer conference and exposition*, pages 795–801, 1976.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. springer series in statistics. *New York, NY, USA*, 2001.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Anne Hecksteden, Jochen Kraushaar, Friederike Scharhag-Rosenberger, Daniel Theisen, Stephen Senn, and Tim Meyer. Individual response to exercise training-a statistical perspective. *Journal of applied physiology*, 118(12):1450–1459, 2015.

Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6):798–817, 2019.

Jennifer M Hootman, Randall Dick, and Julie Agel. Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. *Journal of athletic training*, 42(2):311, 2007.

Mahan Hosseini, Michael Powell, John Collins, Chloe Callahan-Flintoft, William Jones, Howard Bowman, and Brad Wyble. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, 119:456–467, 2020.

Michael Hutchison, Lynda M Mainwaring, Paul Comper, Doug W Richards, and Sean M Bisschop. Differential emotional responses of varsity athletes to concussion and musculoskeletal injuries. *Clinical Journal of Sport Medicine*, 19(1): 13–19, 2009.

Michael Hutchison, Paul Comper, Lynda Mainwaring, and Doug Richards. The influence of musculoskeletal injury on cognition: implications for concussion research. *The American journal of sports medicine*, 39(11):2331–2337, 2011.

H Ij. Statistics versus machine learning. *Nat Methods*, 15(4):233, 2018.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Dean M Karantonis, Michael R Narayanan, Merryn Mathie, Nigel H Lovell, and Branko G Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE transactions on information technology in biomedicine*, 10(1):156–167, 2006.

Josip Karuc, Marjeta Mišigoj-Durakovic, Marko Šarlija, Goran Markovic, Vedran Hadžic, Tatjana Trošt-Bobic, and Maroje Soric. Can injuries be predicted by functional movement screen in adolescents? the application of machine learning. *The Journal of Strength & Conditioning Research*, 35(4):910–919, 2021.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. Data preprocessing for supervised leaning. *International journal of computer science*, 1(2): 111–117, 2006.

Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

AS Leon, SE Gaskill, T Rice, J Bergeron, J Gagnon, DC Rao, JS Skinner, JH Wilmore, and C Bouchard. Variability in the response of hdl cholesterol to exercise training in the heritage family study. *International journal of sports medicine*, 23(01):1–9, 2002.

Christophe Ley, R Kyle Martin, Ayoosh Pareek, Andreas Groll, Romain Seil, and Thomas Tischer. Machine learning and conventional statistics: making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(3):753–757, 2022.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

Alejandro López-Valenciano, Francisco Ayala, José Miguel Puerta, Mark De Ste Croix, Francisco Vera-García, Sergio Hernández-Sánchez, Iñaki Ruiz-Pérez, and Gregory Myer. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Medicine and science in sports and exercise*, 50(5): 915, 2018.

Bryan C Luu, Audrey L Wright, Heather S Haeberle, Jaret M Karnuta, Mark S Schickendantz, Eric C Makhni, Benedict U Nwachukwu, Riley J Williams III, and Prem N Ramkumar. Machine learning outperforms logistic regression analysis to predict next-season nhl player injury: an analysis of 2322 players from 2007 to 2017. *Orthopaedic journal of sports medicine*, 8(9):2325967120953404, 2020.

Arlo Lyle. *Baseball prediction using ensemble learning*. PhD thesis, University of Georgia, 2007.

Áine Macnamara and Dave Collins. Development and initial validation of the psychological characteristics of developing excellence questionnaire. *Journal of Sports Sciences*, 29(12):1273–1286, 2011.

Lynda M Mainwaring, Michael Hutchison, Sean M Bisschop, Paul Comper, and Doug W Richards. Emotional response to sport concussion compared to acl injury. *Brain injury*, 24(4):589–597, 2010.

Nikolaos Maniadakis and Alastair Gray. The economic burden of back pain in the uk. *Pain*, 84(1):95–103, 2000.

J McCullagh and T Whitfort. An investigation into the application of artificial neural networks to the prediction of injuries in sport. *International Journal of Sport and Health Sciences*, 7(7):356–360, 2013.

Willem H Meeuwisse, Hugh Tyreman, Brent Hagel, and Carolyn Emery. A dynamic model of etiology in sport injury: the recursive nature of risk and causation. *Clinical journal of sport medicine*, 17(3):215–219, 2007.

Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Prognosis and prognostic research: what, why, and how? *Bmj*, 338, 2009.

Seyed Hamed Mousavi, Juha M Hijmans, Forough Moeini, Reza Rajabi, Reed Ferber, Henk van der Worp, and Johannes Zwerver. Validity and reliability of a smartphone motion analysis app for lower limb kinematics during treadmill running. *Physical Therapy in Sport*, 43:27–35, 2020.

Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

Grethe Myklebust, Inger Holm, Sverre Mæhlum, Lars Engebretsen, and Roald Bahr. Clinical, functional, and radiologic outcome in team handball players 6 to 11 years after anterior cruciate ligament injury. *The American journal of sports medicine*, 31(6):981–989, 2003.

Jon L Oliver, Francisco Ayala, Mark BA De Ste Croix, Rhodri S Lloyd, Greg D Myer, and Paul J Read. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of science and medicine in sport*, 23(11):1044–1048, 2020.

Cathy O'Neil and Rachel Schutt. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.

Ashwin A Phatak, Saumya Mehta, Franz-Georg Wieland, Mikael Jamil, Mark Connor, Manuel Bassek, and Daniel Memmert. Context is key: normalization as a novel approach to sport specific preprocessing of kpi's for match analysis in soccer. *Scientific Reports*, 12(1):1–6, 2022.

Angkoon Phinyomark, Giovanni Petri, Esther Ibáñez-Marcelo, Sean T Osis, and Reed Ferber. Analysis of big data in gait biomechanics: Current trends and future directions. *Journal of medical and biological engineering*, 38:244–260, 2018.

Baseball Prospectus. *Baseball Prospectus 2022*. Stylus Publishing, LLC, 2003.

Chava L Ramspek, Ewout W Steyerberg, Richard D Riley, Frits R Rosendaal, Olaf M Dekkers, Friedo W Dekker, and Merel van Diepen. Prediction or causality? a scoping review of their conflation within current observational research. *European journal of epidemiology*, 36:889–898, 2021.

Thomas Reilly, A Mark Williams, Alan Nevill, and Andy Franks. A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences*, 18 (9):695–702, 2000.

Nahida Reyaz, Gulfam Ahamad, Naveed Jeelani Khan, and Mohd Naseem. Machine learning in sports talent identification: A systematic review. In *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, pages 1–6. IEEE, 2022.

Chris Richter, Martin O'Reilly, and Eamonn Delahunt. Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, pages 1–7, 2021.

Patrick Riley. Three pitfalls to avoid in machine learning. *Nature*, 572(7767):27–29, 2019.

Nikki Rommers, Roland RÃ¶ssler, Evert Verhagen, Florian Vandecasteele, Steven Verstockt, Roel Vaeyens, Matthieu Lenoir, Eva D'Hondt, and Erik Witvrouw. A machine learning approach to assess injury risk in elite youth football players. *Medicine and science in sports and exercise*, 52(8):1745–1751, 2020.

Alessio Rossi, Luca Pappalardo, Paolo Cintia, F Marcello Iaia, Javier Fernández, and Daniel Medina. Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264, 2018.

Joshua D Ruddy, Anthony J Shield, Nirav Maniar, Morgan D Williams, Steven Duhig, Ryan G Timmins, Jack Hickey, Matthew N Bourne, and David A Opar. Predictive modeling of hamstring strain injuries in elite australian footballers. *Med Sci Sports Exerc*, 50(5):906–14, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Mirka Saarela. Automatic knowledge discovery from sparse and large-scale educational data: case finland. *Jyväskylä studies in computing*, (262), 2017.

Kristin L Sainani. Explanatory versus predictive modeling. *PM&R*, 6(9):841–844, 2014.

Hugo Sarmento, M Teresa Anguera, Antonino Pereira, and Duarte Araújo. Talent identification and development in male football: A systematic review. *Sports medicine*, 48:907–931, 2018.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Galit Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

Zahari Taha, Rabiu Muazu Musa, Anwar PP Abdul Majeed, Muhammad Muaz Alim, and Mohamad Razali Abdullah. The identification of high potential archers based on fitness and motor ability variables: A support vector machine approach. *Human movement science*, 57:184–193, 2018.

Iskandar Tamimi, Joaquin Ballesteros, Almudena Perez Lara, Jimmy Tat, Motaz Alaqueel, Justin Schupbach, Yousef Marwan, Cristina Urdiales, Jesus Manuel Gomez-de Gabriel, Mark Burman, et al. A prediction model for primary anterior cruciate ligament injury using artificial intelligence. *Orthopaedic Journal of Sports Medicine*, 9(9):23259671211027543, 2021.

Iulian Emil Tampu, Anders Eklund, and Neda Haj-Hosseini. Inflation of test accuracy due to data leakage in deep learning-based classification of oct images. *Scientific Data*, 9(1):580, 2022.

Pang-Ning Tan, Michael Stienbach, and Vipin Kumar. Introduction to data mining, 139-20, 2007.

Heidi R Thornton, Jace A Delaney, Grant M Duthie, and Ben J Dascombe. Importance of various training-load measures in injury incidence of professional rugby league athletes. *International journal of sports physiology and performance*, 12(6):819–824, 2017.

Merel van Diepen, Chava L Ramspek, Kitty J Jager, Carmine Zoccali, and Friedo W Dekker. Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrology Dialysis Transplantation*, 32(suppl_2):ii1–ii5, 2017.

Hans Van Eetvelde, Luciana D Mendonça, Christophe Ley, Romain Seil, and Thomas Tischer. Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of experimental orthopaedics*, 8(1):1–15, 2021.

RN Van Gent, Danny Siem, Marienke van Middelkoop, AG Van Os, SMA Bierma-Zeinstra, and BW Koes. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *British journal of sports medicine*, 41(8):469–480, 2007.

Marienke Van Middelkoop, Jelle Kolkman, John Van Ochten, SMA Bierma-Zeinstra, and Bart W Koes. Risk factors for lower extremity injuries among male marathon runners. *Scandinavian journal of medicine & science in sports*, 18 (6):691–697, 2008.

Borislava Vrigazova. The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the Society for Advancing Innovation and Research in Economy*, 12 (1):228–242, 2021.

Akbar K Waljee, Peter DR Higgins, and Amit G Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014.

David Whiteside, Douglas N Martini, Adam S Lepley, Ronald F Zernicke, and Grant C Goulet. Predictors of ulnar collateral ligament reconstruction in major league baseball pitchers. *The American journal of sports medicine*, 44(9):2202–2209, 2016.

Che-Chang Yang and Yeh-Liang Hsu. A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772–7788, 2010.

Mohammed J Zaki and Wagner Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

ORIGINAL PAPERS

I

TALENT IDENTIFICATION IN SOCCER USING A ONE-CLASS SUPPORT VECTOR MACHINE

by

Susanne Jauhiainen, Sami Äyrämö, Hannele Forsman, Jukka-Pekka Kauppi
2019

International Journal of Computer Science in Sport, 18(3), 125-136

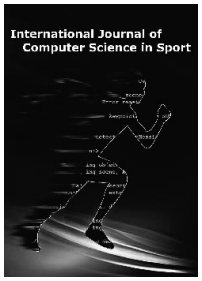
DOI:10.2478/ijcss-2019-0021

International Journal of Computer Science in Sport

Volume 18, Issue 3, 2019

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.2478/ijcss-2019-0021



Talent identification in soccer using a one-class support vector machine

Jauhiainen S.[1], Äyrämö S.[1], Forsman H.[2], Kauppi J-P.[1]

[1]Faculty of Information Technology, University of Jyväskylä, Finland

[2]Eerikkila Sports & Outdoor Resort, Training and Research Centre forFinnish Football Urheiluopistontie Eerikkila, Tammela

Abstract

Identifying potential future elite athletes is important in many sporting events. The successful identification of potential future elite athletes at an early age would help to provide high-quality coaching and training environments in which to optimize their development. However, a large variety of different skills and qualities are needed to succeed in elite sports, making talent identification generally a complex and multifaceted problem. Due to the rarity of elite athletes, datasets are inherently imbalanced, making classical statistical inference difficult. Therefore, we approach talent identification as an anomaly detection problem. We trained a nonlinear one-class support vector machine (one-class SVM) on a dataset (N=951) collected from 14-year-old junior soccer players to detect potential future elite players. The mean area under the receiver operating characteristic curve (AUC-ROC) over the tested hyperparameter combinations was 0.763 (std 0.007). The most accurate model was obtained when physical tests, measuring, for example, technical skills, speed, and agility, were used. According to our results, the proposed approach could be useful to support decision-makers in the process of talent identification.

KEYWORDS: TALENT IDENTIFICATION, ANOMALY DETECTION, ONE-CLASS SVM

Introduction

The amount of data in sports is rapidly increasing due to advances in data collection technologies (Brefeld & Zimmermann, 2017). This has opened many possibilities for data analysis and application development across all sports. Even though sports analytics is a relatively new field, a variety of different research questions, approaches and data sources are already documented in the literature (Brefeld & Zimmermann, 2017). For example, data analysis has been used for predicting outcome of a game (Aoki, Assuncao, & de Melo, 2017), decision support for passing in soccer (Power, Ruiz, Wei, & Lucey, 2017), and optimization of a training schedule (Knobbe, Orie, Hofman, van der Burgh, & Cachucho, 2017). Commonly used data analysis methods with examples from elite sports are introduced in (Ofoghi, Zeleznikow, MacMahon, & Raab, 2013). One example of data analysis utilized in talent identification is PECOTA (Player Empirical Comparison and Optimization Test Algorithm). It calculates different career paths for baseball players and forecasts player's performance utilizing similarity scores and projection (Silver, 2003). Another study, focusing on decision making in sport management, used the ordered weighted averaging (OWA) operator for selection of players (Merigó & Gil-Lafuente, 2011). In addition to talent identification, applications for suggesting the best sports in terms of athlete's individual capabilities have been developed (Papić, Rogulj, & Pleština, 2009).

The identification and selection of talented players at an early age is important in many sporting events. It will enable offering high-quality coaching and training environments for talented players and thereby accelerate their development (Williams & Reilly, 2000). However, the identification task, especially in team games, is a very complex process (Reilly, Williams, Nevill, & Franks, 2000). In soccer, for example, a great variety of physical features and technical skills are needed for success (Reilly et al., 2000). Moreover, psychological skills and characteristics also play an important role at elite level (Macnamara & Collins, 2011). Therefore, talent identification in sports should be based on a versatile set of variables.

Furthermore, the datasets in talent identification are inherently imbalanced due to the rarity of elite athletes in sports. In practice, this means that there are typically significant differences in the number of observations available from different classes (Chawla, Japkowicz, & Kotcz, 2004), which must be carefully taken into account when designing a machine learning method for the case at hand. Imbalanced datasets are common in many other real-life applications as well (He & Garcia, 2009). In this study only 14 observations were available for the minority class, whereas majority class consisted of almost thousand observations.

The two main approaches in data analysis are explanatory and predictive modeling (Breiman, 2001). Many previous studies on talent identification have concentrated on explanatory data analysis (Nieuwenhuis, Spamer, & Rossum, 2002; O'Connor, Larkin, & Mark Williams, 2016; Woods, Raynor, Bruce, McDonald, & Robertson, 2016). Although differences between talented and other players have been found, the predictive power of their models is unclear. Therefore, there is a need for models whose predictive power has been evaluated on independent test data. This need has also been noticed in other studies, such as the one by Smiths, Lipscomb, and Simkins (2007), where a predictive approach is applied on award prediction in baseball, a task that has been previously approached with explanatory methods.

In this research, we studied the potential of machine learning in talent identification, using data containing a diverse set of variables. Our goal was to analyze whether potential future elite players can be distinguished from majority of the players based on their test information already as juniors. Because of the limited number of observations in the minority class, the use of supervised machine learning methods would easily lead to overfitting. Therefore, we used one-class classification approach, where the training stage was completely unsupervised, and

information about the class labels was only used in model assessment (Goldstein & Uchida, 2016).

Methods

The target data of this study were collected by *The training and research centre for Finnish football* for monitoring the development of young soccer players. A total of 4991 junior soccer players (age mean±std 12.41±1.53 year, range 8-18 year) participated in the specific test events organized by the centre between the years 2011 and 2017. Out of the 47 participating teams, 41 were Finnish, but 293 players came from Sweden, Denmark, England, and the Netherlands. The participating Finnish teams are the best of their age group in Finland.

Each player participated in the test events twice a year together with their team. During the events, the players performed physical tests including, e.g., technical, speed, and agility tests. Moreover, they completed a self-assessment test including, e.g., perceived competence, tactical skills, and motivation. Description of the test protocol can be found in Forsman et al. (2016). The physical tests were measured in a continuous scale (such as length of a 5-jump or time of a speed test). The questionnaire scale was a discrete 5-point Likert scale concerning sport performance, anchored with 1 (almost never) and 5 (almost always).

Data selection

Our goal was to investigate how accurately we can detect potential future elite players among the large pool of players based on the collected test information. Some of the tested Finnish players in our dataset are currently pursuing an international soccer career and have already signed a contract with an international academy. In the absence of senior players who have reached the absolute elite performance level by playing, e.g., for a national team, these international academy players (from now on called ”academy players”) were labelled as talent category for the present study. The player categorization was defined by an educated person in charge of player development at the training and research centre for Finnish football. The players representing other than Finnish teams were excluded from the further analysis due to the insufficient information of their current career development.

All academy players were boys and all of them had performed the tests at the age of 14. For these reasons, 14-year-old Finnish boys were selected for our analyses. The age limit for signing a contract to an academy is 16 years. Therefore, we dropped out of the study those players born in 2003 or later as at the time of this study they could not have a signed contract even though they might be future academy players. In the whole dataset, the total number of academy players was 26. Twelve of these players were dropped out from the analysis due to the overly many missing test results. The final data set included 14 academy players.

Further pruning of the data set was performed due to a large number of variables with a significant proportion of non-random missing values, i.e., they followed *not missing at random* (NMAR) pattern (Little & Rubin, 2002). These missing values were caused, for example, by adjustments to the test protocols and questionnaires or inability of a player to participate all the test events.

Finally, the used data representation (called ”*phys large*”, N=951) consisted of 16 variables in which the test results were measured for at least half of the players (see Table 1). Since the questionnaire answers were missing from more than half of the players they were not included in “*phys large”*. In order to characterize univariate differences between the academy and nonacademy players two-sample t-tests were performed. Normality of the variables was tested

using Shapiro-Wilk test (if $n < 50$) or Kolmogorov-Smirnov test (otherwise). Homogeneity of the variances was tested with Levene test. When the assumption on normality failed Wilcoxon rank sum test was used. Significance level $\alpha = 0.05$ with Bonferroni correction was used and effect size Cohen's *d* reported (Cohen, 1988). All test were performed using MATLAB version R2016b.

Table 1: Mean/median and standard deviation of the variables in "*phys large*" separately for the 14 academy players and 937 non-academy players. A statistically significant difference between the groups was found with 5-jump, height, and weight (*$p < 0.05, d > 0.8$; **$p < 0.01, d > 0.8$).

Countinous variables	**Mean (std) of non-academy players**	**Mean (std) of academy players**
5 jump (m)**	10.90 (0.85)	11.69 (0.63)
Agility (sec)	6.98 (0.58)	6.87 (0.26)
Countermovement jump (cm)	29.87 (4.58)	32.54 (4.76)
Driving and shooting (sec)	15.03 (4.19)	13.36 (4.18)
Speed 10 meters (sec)	1.80 (0.09)	1.74 (0.06)
Speed 20 meters (sec)	3.19 (0.16)	3.05 (0.10)
Speed 30 meters (sec)	4.49 (0.25)	4.30 (0.14)
Speed 5 meters (sec)	1.03 (0.05)	0.99 (0.04)
Weight (kg)*	55.59 (9.48)	62.41 (5.65)
Height (cm)**	167.75 (8.93)	176.33 (6.06)
Juggling (sec)	24.63 (7.58)	22.34 (8.01)
Dribbling (sec)	25.74 (2.78)	25.20 (1.92)
Passing (sec)	37.15 (6.12)	34.94 (6.01)
Gymnastics (points)	12.22 (1.96)	12.05 (2.47)
Yo-Yo endurance (m)	2248.18 (319.00)	2480.00 (330.32)
Discrete variable	Median of non-academy Players	Median of academy Players
Mobility (1-3)	3	3

In order to investigate the predictive power of questionnaire data, another representation (called "*phys+quest*") of the data with fewer players (N = 468), but a greater number of variables including all the 16 "*phys large*" variables and additionally 18 variables from a specific questionnaire consisting of self-assessment of perceived competence, was defined (see Table 2). The questionnaire measures how the players rate their skills in offense, defense, and one versus one situations. Four out of the 14 academy players did not answer the questionnaire and they were dropped off. A detailed description of the questionnaire can be found in Forsman et al. (2016).

Table 2: Questionnaire variables used in this study.

M1. Mean of offensive skill questions
M2. Mean of 1-on-1 skill questions
M3. Mean of defensive skill questions

Q1. I can schedule my own movement correctly in offensive and defensive play

Q2. I have clear solution models about how to win 1-on-1 situations

Q3. I am usually the first player to reach the ball

Q4. I can easily lose my opponent in different game situations

Q5. I feel strong in match ups

Q6. In 1-on-1 situations, I am stronger/faster than my opponent

Q7. I can accomplish the typical play for my position in defensive play

Q8. I can, if necessary, help/support my teammates in defensive situations

Q9. I have a soft ”touch” on the ball

Q10. I dare to keep the ball to myself even in tight spaces

Q11. I have clear solution models about how I score in the different situations in the games

Q12. I can move to the empty spaces on the field, so that my teammates can pass me the ball

Q13. I can find my teammates with my sharp and accurate passes

Q14. I can accomplish the typical play for my position in offensive play

Q15. I know how my teammates are moving in attack situations and it is easy for me to pass them the ball

In order to analyze the predictive power of physical tests and perceived competence self-assessment independently of each other, the variables in ”*phys+quest*” were further split into two smaller representations consisting of only either physical (called ”*phys*”, N =468) or questionnaire (called ”*quest*”, N = 468) variables. The number of academy players was ten in both presentations. A comparison between ”*phys*” and ”*phys large*” representations was performed in order to evaluate the effect of sample size to the predictive accuracy. The summary of the four data representations is shown in Table 3.

Table 3: Different representations of the player data analyzed in this study.

Data representation	N	Variables	D
Phys large	951	Physical variables	16
Phys+quest	468	Physical variables + questionnaire	34
Phys	468	Physical variables	16
Quest	468	Questionnaire	18

Data preprocessing

Prior to model fitting all the physical variables were normalized and the Likert scale questionnaire variables were min-max scaled to range $[\frac{1}{10}, \frac{9}{10}]$. After removing observations and variables due to NMAR values, the remaining missing values were imputed using a self-implemented k-nearest neighbor (knn) imputation algorithm on MATLAB (Bishop, 2006). The estimate for each missing value was computed as the sample mean of the ten nearest neighbours based on Euclidean distance. Although the standard MATLAB knn classier uses $k = 1$ as default,

larger values of k have been found to control noise and perform better. Several studies have found the method relatively insensitive to the exact value of k between 10-20 (Beretta & Santaniello, 2016; Troyanskaya et al., 2001). In addition, principal component analysis (PCA) (Jolliffe, 1986) was used to eliminate correlations from the data. The minimal number of PCs which explained at least 90% of the total variance of the data was chosen, as suggested by Jolliffe (1986). The number of chosen PCs was ten for "*phys large*", "*phys*", and "*quest*". In the case of "*phys+quest*", PCs were calculated separately for physical and questionnaire variables yielding ten PCs for both subsets, and thereby altogether twenty PCs for "*phys+quest*".

One-class support vector machine

Because we only had 14 academy players available for our analysis, training a classifier with supervised methods can be highly sensitive to overfitting. For this reason, we trained one-class support vector machine (one-class SVM) (Chandola, Banerjee, & Kumar, 2009) to model the normal region of the data based only on the observations from the majority class, i.e., the non-academy players. The trained model can then be used to predict whether new observations belong to this normal region or not.

The primal problem of one-class SVM is (Chang & Lin, 2011):

$$\min_{\boldsymbol{w},\xi,p} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - p + \frac{1}{\upsilon N}\sum_{i=1}^{N}\xi_i \tag{1}$$

$$s.t.\begin{cases}\boldsymbol{w}^T\phi(\boldsymbol{x}_i) \geq p - \xi_i \\ \xi_i \geq 0,\end{cases} \tag{2}$$

where ϕ is a feature map that transforms data point $\boldsymbol{x}_\mathrm{i}$ into higher-dimensional space, $\boldsymbol{w}$ is a weight vector and p an offset parameterizing the region. ξ_is are slack variables, N is the number of observations, and υ is an upper bound on the fraction of training errors. More detailed description can be found in (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001). We evaluated the performance of one-class SVM (Python's scikit-learn library, version 0.20.0) using 16 different combinations of hyperparameters γ (0.1, 0.2, 0.3, and 0.4) and υ (0.05, 0.1, 0.2, and 0.4). Radial basis function (RBF) has been found to work best with one-class SVM (Bounsiar & Madden, 2014) and was chosen here as the kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}, \tag{3}$$

where γ is the kernel coefficient. To verify the need of non-linearity, a baseline classifier, where the RBF-kernel was replaced with the linear kernel, was trained.

Performance evaluation

The performance of the different one-class SVM models were assessed with 10-fold cross validation. The majority class was first divided into ten folds, one for testing and nine for training, and then the players of the minority class were added to each test fold. The learning process is unsupervised, because the information from the minority class is used only for performance evaluation and not for classifier training. Preprocessing, including normalization, knn-imputation and PCA, was performed separately in each fold, first for training data and then the obtained parameters were applied to the test cases before predicting the classes.

As performance metrics, we used the mean area under the receiver operating characteristic curve

(AUC-ROC) and the mean area under the precision recall curve (AUC-PR). The mean value of the metrics were calculated across all the ten folds. Although AUC-ROC is a widely used performance measure in machine learning (Narasimhan & Agarwal, 2013), AUC-PR might be a better option for highly imbalanced datasets (Davis & Goadrich, 2006). For both AUC measures the ideal classifier would yield the maximum score of one. The AUC-ROC score of a completely random classifier is 0.5, whereas the baseline level of AUC-PR depends on the class ratio in the data. The per fold AUC-PR baseline values are 0.869 and 0.818 for "*phys+large*" and the three other representations, respectively. Mean AUC-ROC values were compared with Kruskal Wallis test and in case of differences, Tukey's post-hoc tests were performed. Limit of statistical significance was set to p=0.05 and Bonferroni corrected. Based on Kolmogorov-Smirnov test, the values were not normally distributed.

In addition, mean sensitivity and specificity values across the test folds were also calculated using the default decision threshold. Sensitivity measures the proportion of correctly detected academy players, and specificity measures the proportion of correctly detected non-academy players.

Results

One-class SVM results

In Table 4, one-class SVM results in the talent identification task are summarized. The highest mean AUC-ROC value over the tested hyperparameters was 0.763 for data representation *"phys large"*. For representations "*phys*", "*phys+quest*", and "*quest*", the mean AUC-ROC values over the tested hyperparameters were 0.665, 0.643, and 0.585, respectively.

Table 4: Talent identification results for the proposed one-class SVM classifier using the four different data representations. The mean values over hyperparameter combinations and the cross-validation folds are reported for each performance measure (AUC-ROC, AUC-PR, sensitivity, specificity).

	"phys large"	***"phys"***	***"phys+quest"***	***"quest"***
Mean AUC-ROC	0.763(±0.007)	0.665(±0.016)	0.643(±0.013)	0.585(±0.062)
Mean AUC-PR	0.960(±0.002)	0.913(±0.009)	0.880(±0.003)	0.313(±0.194)
Mean sensitivity	0.795(±0.184)	0.732(±0.226)	0.838(±0.120)	0.313(±0.194)
Mean specificity	0.614(±0.142)	0.520(±0.176)	0.355(±0.235)	0.789(±0.125)

Differences in AUC-ROC values were significant between all data representation ($p<0.001$), except between "*phys*" and "*phys+quest*" (p=0.113). It can also be observed from the estimated accuracies of "*phys*" and "*phys+quest*" models, that the questionnaire variables did not improve the performance of the models. The results obtained with "*phys*" and "*phys large*" demonstrate, in line with the expectations, that the estimated classification performance tends to improve along with the number of available observations.

All the AUC-PR values were in line with the above-mentioned results. Note that in the case of AUC-PR, the baseline depends on the class ratio and therefore the results for "*phys*" and "*phys large*" are not directly comparable. When the non-linear kernel was replaced with the linear kernel in the one-class SVM classification model, the performance decreased notably for all the data representations. The mean of the AUC-ROC values in this case were: 0.548, 0.496, 0.612, and 0.582, for "*phys large"*, "*phys*", "*phys+quest*", and "*quest*", respectively.

Discussion

The aim of this study was to investigate whether potential future elite soccer players can be identified from a large group of players using machine learning and data collected by physical and psychological tests in their youth. Application of data-driven approaches to talent identification can be generally considered a cumbersome research problem due to the scarcity of childhood data from elite players. Previous research on talent identification has focused on explanatory methods, i.e., explaining relationships and dependencies between variables without assessing generalization abilities of the fitted models on independent observations (Nieuwenhuis et al., 2002; O'Connor et al., 2016; Woods et al., 2016). In this study we evaluated the predictive ability of the one-class SVM anomaly detection method when trained on four different representations of the soccer player test data set.

The best classification performance (mean AUC-ROC value 0.763) was obtained with the set of variables representing physical tests and the greatest number of players (see Table 4). According to classification proposed by Youngstrom (2013), this result can be considered as "fair". In addition, one might argue that this result is satisfactory considering that the classification model has been fitted in an unsupervised manner using only cases from the category of non-academy players and tested using independent data (using CV) involving players from both categories. Besides, since the number of academy players was limited, one-class SVM hyperparameters γ and υ were not optimized in this study, but the average results over multiple classification models (trained using several combinations of γ and υ) were reported. Once more data for classifier validation becomes available, model selection based on CV can be applied to improve the current results.

While the estimated sensitivity of the "*phys large*" representation in the identification task was nearly 0.80, the estimated specificity of 0.614 shows that yet a large proportion of the players without an academy contract will likely be misclassified into the class of potential academy players. The results prove that there is still a long way to go before talent identification can be made by data-driven machine learning tools independently of human expertise. Realistically, the goal should not be full automation of the selection process, but rather modeling of the talent detection expertise possessed by the best professionals in coaching and player management. These data-driven decision support tools may be able to transfer knowledge and enhance decision making in local and regional development organizations.

Several studies have reported relatively high classification performance measures for various models, but the results can be optimistic from the predictive ability point of view, as their performances have not been tested on independent test observations. A multi-dimensional approach for talent identification among young soccer players with AUC-ROC value of 0.954 was presented by Woods et al. (2016). In O'Connor et al. (2016), 93.7% of young soccer players were correctly classified based on selection or nonselection for a full-time elite player scholarship. Nieuwenhuis et al. (2002) reported accuracy of 90.5% when young female field hockey players were classified as successful or less successful. A web-oriented expert system for talent identification in soccer was developed by Louzada, Maiorano, and Ara (2016). It applies principal component and factor analysis to compute general scores for the players in real time. However, without estimation of the generalization ability on independent test observations, the results are not directly comparable to ones presented in this study.

We also studied the relevance of two different types of tests for measuring players physical and psychological abilities by constructing four different representations of the data. The largest data representation, "*phys large*", produced the greatest overall classification performance. The highest sensitivity was achieved with the most versatile set of variables "*phys+quest*". The

representation ”*quest*” achieved the highest specifity, but the sensitivity was low. While some promising players may not become detected due to the low sensitivity, the higher specificity will lead to a lower number of false positives. This enables detection of a smaller group of players with potentially higher chances to succeed. In small countries, such as Finland, elite soccer players are a rarity and higher sensitivity should probably be preferred in order to prevent loss of unrecognized talents. Downsizing of the training group can be completed by coaches when necessary, and thereby ensure that all the talents receive special attention from their training organizations.

It should be noted that without a doubt some of the players that were assigned in the minority class by the SVM model can be potential future elite players, but they have been still too young for signing a contract at the time of this study. In addition, even if a player shows exceptional potential at the age of 14, numerous factors, such as maturity, injuries, coaching/scouting, or decision about whether to stay in Finland to finish school, affects her/his future as an elite soccer player. This is a limitation of this study, and must be taken into account when interpreting the results.

Another limitation of this study is the high number of missing values. Our results indicate that the performance of the model can be improved when more players will be available in both minority and majority classes. Thus, in the future, we can expect improvements due to the continuous accumulation of the data. In addition, some of the included players may sign a contract with a soccer academy after this study, which will enable further improvement to the current models. Moreover, increase in the size of the minority class would enable more thorough validation of the one-class model, or even use of supervised machine learning methods. Also, with larger data it can be possible to utilize different classifications as well, for example look at whether the player made it into the national team.

Furthermore, many of the relevant observations were incomplete. For instance, the self-assessment questionnaire measuring the player’s motivation could improve the classification performance, but in the present study these variables had to be discarded due to missing values. These missing values were caused by refinement and changes in the questionnaires and tests over the years as well as the fact not all questions were compulsory to answer. These issues have been considered and in the future, more complete data will be attained. Also, with more research, the most relevant features can be detected to improve the model. In the long-term perspective, it might become possible to include even more complex data types, such as player tracking or video data, to the machine learning process.

In this study, the parameters of the missing value imputation and dimension reduction methods were fixed based on the existing literature. However, data-based optimization of the parameters might improve the performance of the models.

Conclusion

Identifying talented athletes at young age is an interesting but difficult problem to be successfully solved by machine learning. Accurate identification may, however, enable better career development and level of performance for talented players. In this study, an unsupervised anomaly detection method, one-class SVM, was used to detect potential elite soccer players based on their test data in youth. The best results (mean AUC-ROC 0.763) were achieved when the largest dataset including physical test measurements was used. Considering the size and quality of the available data the present results are promising, but not yet able to provide practical tools to the field. The results also suggest that non-linear methods might be more efficient in the talent identification task than linear ones. Follow-up studies should focus on repeating the study with larger number of players and a more versatile set of variables.

Acknowledgment

This work has been carried out in two projects "Value from health data with cognitive computing" and "Watson Health Cloud", funded by Business Finland. Jukka-Pekka Kauppi was funded by the Academy of Finland Postdoctoral Researcher program (Research Council for Natural Sciences and Engineering; grant number 286019).

References

Aoki, R., Assuncao, R. M., & de Melo, P. O. S. (2017). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1367–1376).

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(3), 197–208.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.

Bounsiar, A., & Madden, M. G. (2014). Kernels for one-class support vector machines. In *2014 International Conference on Information Science & Applications (ICISA)* (pp. 1–4).

Brefeld, U., & Zimmermann, A. (2017). Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, *31*(6), 1577–1579.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Chawla, N. V, Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1–6.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).

Forsman, H., Gråstén, A., Blomqvist, M., Davids, K., Liukkonen, J., & Konttinen, N. (2016). Development and validation of the perceived game-specific soccer competence scale. *Journal of Sports Sciences*, *34*(14), 1319–1327.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS One*, *11*(4), 1–31.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Jolliffe, I. T. (1986). *Principal component analysis* (1st ed.). Springer.

Knobbe, A., Orie, J., Hofman, N., van der Burgh, B., & Cachucho, R. (2017). Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, *31*(6), 1872–1902.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.

Louzada, F., Maiorano, A. C., & Ara, A. (2016). iSports: A web-oriented expert system for talent identification in soccer. *Expert Systems with Applications*, *44*, 400–412.

Macnamara, Á., & Collins, D. (2011). Development and initial validation of the psychological characteristics of developing excellence questionnaire. *Journal of Sports Sciences*, *29*(12), 1273–1286.

Merigó, J. M., & Gil-Lafuente, A. M. (2011). Decision-making in sport management based on the OWA operator. *Expert Systems with Applications*, *38*(8), 10408–10413.

Narasimhan, H., & Agarwal, S. (2013). A structural SVM based approach for optimizing partial AUC. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 516–524).

Nieuwenhuis, C. F., Spamer, E. J., & Rossum, J. H. A. van. (2002). Prediction function for identifying talent in 14-to 15-year-old female field hockey players. *High Ability Studies*, *13*(1), 21–33.

O'Connor, D., Larkin, P., & Mark Williams, A. (2016). Talent identification and selection in elite youth football: An Australian context. *European Journal of Sport Science*, *16*(7), 837–844.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. *Measurement in Physical Education and Exercise Science*, *17*(3), 171–186.

Papić, V., Rogulj, N., & Pleština, V. (2009). Identification of sport talents using a web-oriented expert system with a fuzzy module. *Expert Systems with Applications*, *36*(5), 8830–8838.

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not All Passes Are Created Equal: Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1605–1613).

Reilly, T., Williams, A. M., Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences*, *18*(9), 695–702.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471.

Silver, N. (2003). Introducing PECOTA. *Baseball Prospectus*, *2003*, 507–514.

Smith, L., Lipscomb, B., & Simkins, A. (2007). Data mining in sports: Predicting cy young

award winners. *Journal of Computing Sciences in Colleges*, *22*(4), 115–121.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sport Science*, *18*(9), 657–667.

Woods, C. T., Raynor, A. J., Bruce, L., McDonald, Z., & Robertson, S. (2016). The application of a multi-dimensional assessment approach to talent identification in Australian football. *Journal of Sports Sciences*, *34*(14), 1340–1345.

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221.

II

NEW MACHINE LEARNING APPROACH FOR DETECTION OF INJURY RISK FACTORS IN YOUNG TEAM SPORT ATHLETES

by

Susanne Jauhiainen, Jukka-Pekka Kauppi, Mari Leppanen, Kati Pasanen, Jari Parkkari, Tommi Vasankari, Pekka Kannus, Sami Äyrämö 2021

International Journal of Sports Medicine, 42(02), 175-182

DOI:10.1055/a-1231-5304

New machine learning approach for detection of injury risk factors in young team sport athletes

Abstract

The purpose of this article is to present how predictive machine learning methods can be utilized for detecting sport injury risk factors in a data-driven manner. The approach can be used for finding new hypotheses for risk factors and confirming the predictive power of previously recognized ones. We used three-dimensional motion analysis and physical data from 314 young basketball and floorball players (48.4% males, 15.72±1.79yr, 173.34±9.14cm, 64.65±10.4kg). Both linear (L1-regularized logistic regression) and non-linear methods (random forest) were used to predict moderate and severe knee and ankle injuries (N=57) during three-year follow-up. Results were confirmed with permutation tests and predictive risk factors detected with Wilcoxon signed-rank-test ($p<0.01$). Random forest suggested twelve consistent injury predictors and logistic regression twenty. Ten of these were suggested in both models; sex, body mass index, hamstring flexibility, knee joint laxity, medial knee displacement, height, ankle plantar flexion at initial contact, leg press one-repetition max, and knee valgus at initial contact. Cross-validated areas under receiver operating characteristic curve were 0.65 (logistic regression) and 0.63 (random forest). The results highlight the difficulty of predicting future injuries, but also show that even with models having relatively low predictive power, certain predictive injury risk factors can be consistently detected.

Keywords: Sport medicine, Predictive methods, Machine learning, Knee injuries, Ankle injuries, Basketball and floorball

1. Introduction

Sport injuries are very common across different sports, among both elite and recreational athletes [1–3]. They can have significant effects on the health and performance and may even cause prolonged problems in persons life [3]. Sport injuries can lead to, for example, pain, loss of playing or working time, and decreased motility and stability [3]. The incidence rate of some injuries, such as the anterior cruciate ligament (ACL) injury, is a growing case of concern [4]. Effective prevention of injuries presumes that the most relevant risk factors are found. Even though many intrinsic and extrinsic risk factor have been identified, there is no clear consensus with the findings [5].

A large majority of existing sport injury studies rely on explanatory analysis approach [6, 7]. Explanatory methods have played an important role in the development of sport injury research and will be needed in future research as well. They are used when the purpose is to explain or understand data or phenomena of interest. However, high explanatory power does not necessarily imply high predictive power [8]. Therefore, risk factors that are identified by explanatory methods only demonstrate a statistically significant association with injuries, but might not have predictive power on them [6, 7].

Another limitation of explanatory analysis is that they often focus on a small number of variables and their linear associations with injuries in isolation. However, underlying causes behind sport injuries have been considered to be multifactorial, indicating that a high number of variables and their inter-relationships should be considered [9, 10]. It has also been suggested that using cut-off values and studying only linear interactions between isolated variables can not successfully identify injury predictors, but more complex models should be applied [11]. To overcome these limitations, predictive analysis should be utilized alongside explanatory methods. This has been previously suggested specifically for sport injury research as well [12].

Predictive analysis focuses on predicting new or future observations from data [8]. By exploiting computational power, predictive methods are able to analyze a larger set of variables including their interactions and nonlinear relationships as well as to efficiently remove redundant variables from a

model. Therefore, they can be used for generating new hypotheses for sport injury risk factors in a data-driven manner.

In predictive analysis, the generalization ability of a model should always be assessed on independent test data, i.e., data that have not been used in the training phase. This measures how accurate the trained model will be on new unseen observations and only after this validation can any conclusions about the predictive power be drawn [8]. In addition, when constructing a predictive model it is necessary to confirm that the prediction results were significantly above the random chance level. This kind of confirmatory analysis is especially relevant with smaller sample sizes. If this issue is not considered, in the worst case it can lead to false interpretations and conclusions. For example, in neuroscience the problem has been widely recognize [13]. One way to confirm significance of the models and relevance of the chosen predictors are permutation tests [13].

Another important issue related to predictive analysis is the explainability of a model. Explainability means that the model somehow explains its predictions, for example, gives information on how individual variables contribute to the prediction outcome, and does not only predict as black box [14]. Explainable models and their predictions are more informative, easier to trust, and therefore can provide more practical benefits. A term widely used with sophisticated machine learning methods is Explainable Artificial Intelligence (XAI) [14]. In some domains, such as medicine, model explainability is considered highly important [15] and should be pursued in sport science and medicine as well.

During the last couple of years, the first studies using predictive analysis in sport injury research have been conducted [6, 9, 16, 17]. The previous studies have, however, focused solely on the prediction task without paying attention to the explainability of the models. In addition, two of the studies also used a very low number of variables (from three to eleven), although a larger set might have increased the accuracy [9, 16]. The need and potential of predictive machine learning methods in sport injury prediction have been recognized but more research is needed [12, 17].

Therefore, the aim of this study is to utilize predictive machine learning methods to detect variables with predictive power on sport injuries. We present a framework that can be used to detect consistent injury predictors in a data-driven manner and validate their predictive power on independent test data. Consistent means that the variable is constantly chosen as an important predictor in the used model. Our framework utilizes both linear and non-linear classification methods, namely L1-regularized logistic regression and random forests, to predict moderate and severe knee and ankle injuries. Generalization ability of these models is assessed with 10-fold cross-validation. A reference model based on randomized labels is constructed to confirm that the observed prediction performance is not achieved by chance. Consistent injury predictors are detected with Wilcoxon signed-rank test. This approach can be used for finding new hypotheses for injury risk factors as well as confirming the predictive power of previously recognized risk factors. Our secondary aim is to compare linear and non-linear methods for the task.

2. Methods

2.1. Participants

The data were collected in the Predictors of Lower Extremity Injuries in Team Sports (PROFITS) study [18]. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Pirkanmaa Hospital District, Tampere, Finland (ETL-code R10169). The authors declare that this study meets the ethical requirements of the journal [19]. Altogether 175 basketball and 139 floorball youth (12-21 years) players, including 162 females (15.44±1.95 years, 167.92±6.44 cm, 60.86±8.58 kg) and 152 males (16.03±1.59 years, 179.13±8.00 cm, 68.68±10.76 kg) from the two highest junior league levels of the Tampere city district, Finland, were recruited. To be included they had to be official team members (i.e., have valid playing contract and licenses), 21 years old or younger at baseline, and free from injury at baseline. Information about previous injuries, their treatment, and whether the player was fully recovered were assessed with a baseline questionnaire. The players entered the study during the preseason in 2011, 2012, or 2013.

They signed a written informed consent form before inclusion (including parental consent for players aged ≤18 years).

2.2. Data collection

At baseline, each player participated in physical tests including a vertical drop jump (VDJ) (3D motion analysis), height, weight, isokinetic concentric quadriceps and hamstring strength, isometric hip abductor strength, one repetition maximum (1RM) leg press, knee joint laxity (KT-1000), generalized joint laxity (Beighton scale), genu recurvatum, navicular drop, hip anteversion, and hamstring flexibility (for more details see Supplementary Table 1 and online supplementary appendices in [18]).

The VDJ was performed from a 30-cm box. Players were instructed to drop off the box and perform a maximal jump upon landing with their feet on two separate force platforms (BP6001200; AMTI). The 3D motion analysis was carried out using sixteen reflective markers placed over anatomic landmarks on the lower extremities according to the Plug-In Gait Marker set (Vicon Nexus v.i.7; Oxford Metrics) and eight highspeed cameras (Vicon T40). Kinetics and kinematics variables were extracted using the Vicon Nexus Plug-in Gait model. Medial knee displacements were extracted using a custom MATLAB script (MathWorks Inc). For more detailed description of the motion data collection and variable extraction see [18, 20].

The injury definition was based on the time-loss definition by Fuller et al. [21]. We focused on moderate to severe acute non-contact knee and ankle injuries that resulted in an athlete being unable to fully participate in training or match play for at least 8 days. Non-contact injury was defined as an injury which occurred without direct contact to the injured body part. Injuries were recorded by a team coach or another designated team member. For injury registration, the study physicians contacted the team coach or designate on a weekly basis by phone or email. Designate was someone who was always present at practice and matches, e.g., head, assistant, or strength and conditioning coach, team manager, or physiotherapist. The study physicians contacted the athlete after each injury and collected information about the injury time, place, cause, type, location, and the time-loss due to

the injury in a standardized phone interview. For exposure registration, the team coaches recorded player participation in team practice and game play and emailed the records to the study group at the end of each month.

2.3. Data preprocessing

All data analysis was performed with MATLAB R2016b (MathWorks Inc) and classification methods run with the *Statistics and machine learning toolbox 11.0*. For classification, the players with moderate and severe acute ankle and knee injuries formed the first group (group A, n=57) and players with no injuries formed the other (group B, n=257). Athletes with mild injuries (time-loss ≤7days, n=21) were excluded from the analysis. Altogether 58 variables were chosen for further analysis by a group of experts in sport medicine, including a sports medicine researcher and four clinical researchers (one physiotherapist and three physicians). Four variables had more than 50% of missing values (iliopsoas and quadriceps extensibility from both legs) as they were added to the test patterns only in the second year of testing and these were excluded from the analysis, resulting into 54 variables. The chosen variables are described in the Supplementary Table 1.

After dropping out irrelevant and sparse variables, 22 variables with missing data remained and were imputed with K-nearest neighbour imputation with k value of 10. On average, each of these 22 variables had five missing values (1.6% of the 314 observations). Data was normalized to have mean of zero and standard deviation of one for each column. The variables that had been measured separately for both right and left legs were transformed to dominant (leg used for kicking a ball) and non-dominant leg variables.

2.4. Choice of classification methods

Two commonly used methods, random forest and L1-regularized logistic regression, were chosen for the binary classification task in our framework. These methods were selected because of their inbuilt variable importance features. Random forest is a nonlinear classification and regression method that has become a standard data analysis tool in different fields such as medicine and bioinformatics [22] and has been used in sport injury research as well [23]. It is based on building an ensemble of multiple

decision trees [24]. The model was trained with a hundred trees [24] and the minimum number of observations per tree leaf and the number of predictors to sample at each split were chosen with Bayesian optimization. To estimate the predictive power of the variables, we recorded and analyzed the out-of-bag estimates of variable importance [24].

L1-regularized logistic regression, in turn, is a linear classification method that has been used to model sport injury outcomes [23]. A benefit of this method is that it is capable of automatically discarding redundant and/or irrelevant variables from the model. This is done by penalizing the model with the L1 norm and as a result, some of the variable coefficients tend to shrink to exactly zero. The optimal amount of penalization was estimated with stratified 10-fold cross-validation.

Variable importance for logistic regression was based on the variable coefficient values. We analyzed whether a variable was chosen as a predictor in the model, i.e., the variable coefficient was not shrunk to zero. Variable importance was then the number of times the variable was chosen over the ten CV folds (a value between zero and ten). The sign of each variable coefficient was also assessed in order to perceive whether the variable decreased or increased the injury risk.

2.5. Validation

Generalization ability of our models was assessed with 10-fold cross-validation (CV). K-fold CV is based on randomly splitting the data into K sets and leaving each set at a time for testing while the rest of the sets are used to train a model. Test performance was assessed with Area Under the Receiver Operating Characteristics Curve AUC-ROC [25]. It is based on both true positive and false positive rates and it can be used with imbalanced class distributions which is the case in our data. AUC-ROC provides a value 1.0 for perfect prediction and 0.5 for purely random prediction.

AUC-ROC and variable importance values were estimated by ten-fold cross-validation. Normalization and imputation of the training data were done separately inside each fold and the test data were then normalized using coefficients estimated from the training data. Because K-fold CV is based on random splitting of the data, there is variation in the K-fold validation estimates [26].

Therefore, the analysis was repeated a hundred times and results were averaged over the runs to obtain a more reliable estimate for the generalization ability.

2.6. Confirmatory data analysis

To confirm the significance of our results, permutation tests were used [13]. A reference model was constructed by randomly shuffling the class labels in the training data. By comparing the outcome of the true models to the distribution of values from the random models we confirmed that the performance was not observed by chance. In addition, we can detect significantly consistent injury predictors by comparing the variable importance of the true and the random reference models. If a variable is consistently important in the true model, but not in the reference model, that confirms its significance in the prediction.

To confirm the significance of obtained performance, a paired comparison between AUC-ROC values of the true and random model from a hundred repeated 10-fold CV runs was conducted based on a Wilcoxon signed-rank test. In each CV run, the fold divisions were kept the same for random and true models to allow fair pairwise comparison.

To detect significantly consistent injury predictors, we compared the variable importance values. Again, the values from the hundred repetitions were compared between the random and true models but with Wilcoxon signed-rank test. The limit of significance was set to $\alpha=0.01$ and corrected with Bonferroni correction. The used framework is summarized in Figure 1.

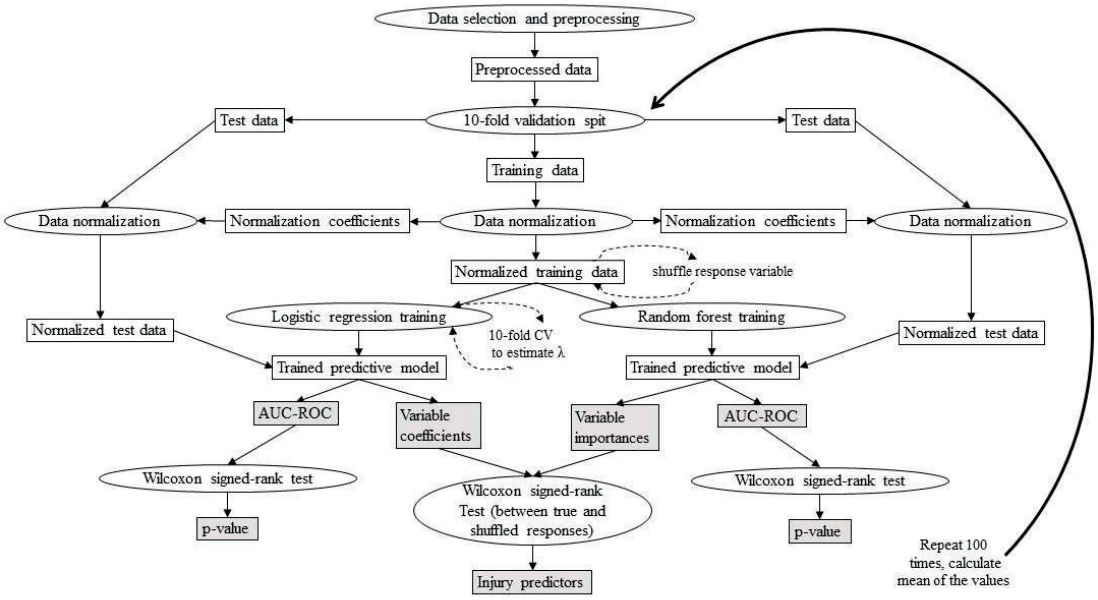


Figure 1. Framework of the proposed predictive analysis approach.

3. Results

3.1. Random forest

Random forest suggested twelve consistent injury predictors ($p < 0.01$). The variable importance values averaged over the CV folds and hundred repeated runs can be seen in Figure 2. The larger the importance value, the greater the importance of the variable is for the prediction task. By comparing the values between true and randomized results, variables with true predictive power can be detected. If the value of true model is significantly larger than the value of random model, its predictive power is not likely result of chance or noise in data. Negative values indicate the variable was not important in prediction.

As seen in the figure, sex, hamstring flexibility (both dominant and non-dominant legs), body mass index (BMI), KT1000 (dominant leg), and height show the highest random forest importance values. Other suggested predictors include leg press 1RM, knee valgus at IC (dominant leg), knee flexion peak (non-dominant leg), medial knee displacement (dominant leg), ankle flexion at IC (dominant leg), and navicular drop (non-dominant leg).

The mean AUC-ROC value for random forest was 0.63 (0.94 for the training data). The AUC-ROC values were higher ($p<0.001$) with real responses than the randomized ones (mean AUC-ROC 0.48), which confirms the significance of the random forest models.

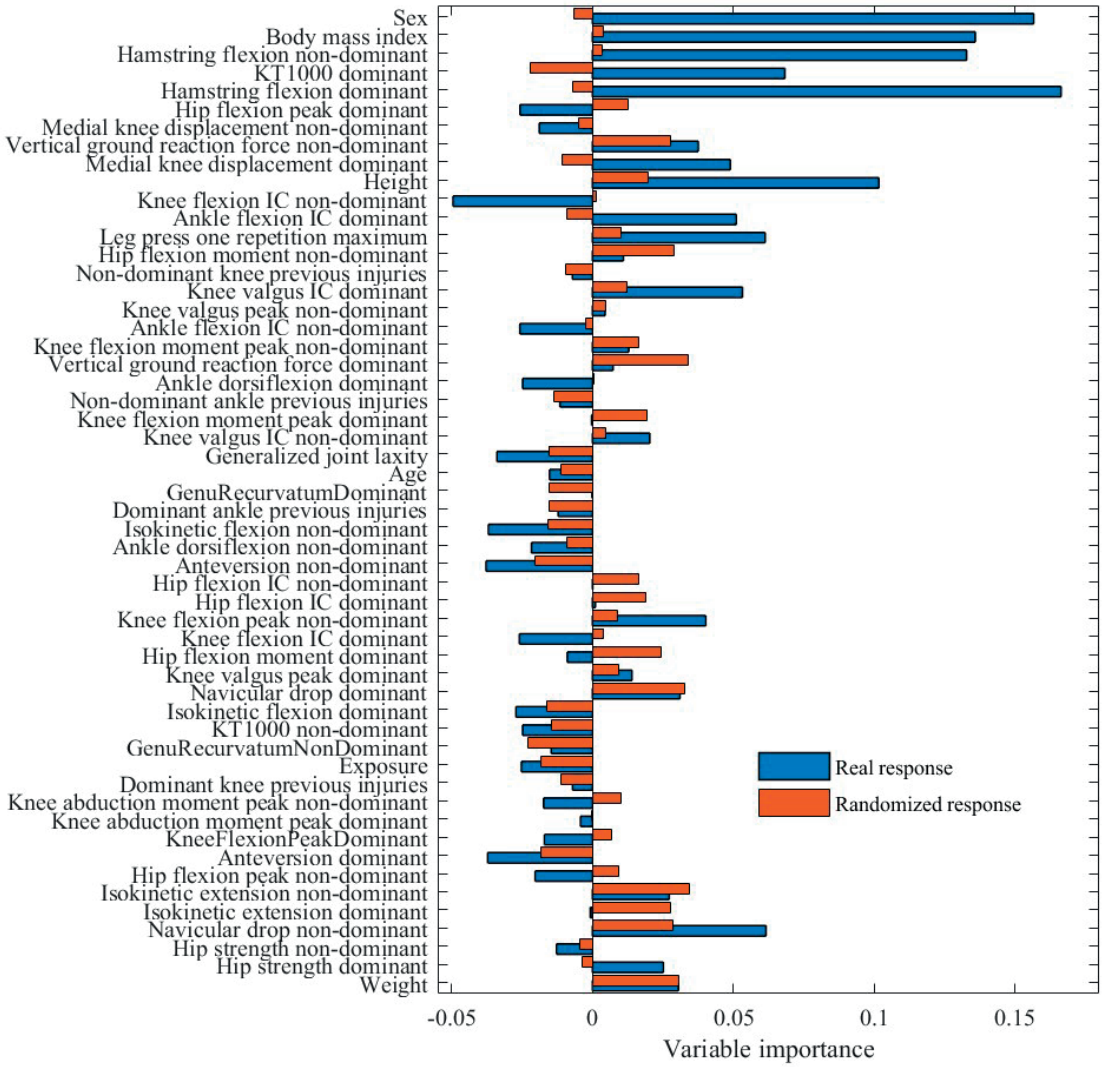


Figure 2. Variable importance values from random forest. Blue bars correspond to the results with real response, red ones with randomized response.

3.2. Logistic regression

Figure 3 shows the variables chosen most frequently as predictors in the L1-regularized logistic regression. The bars represent the number of CV folds where a variable was chosen for the predictive model (i.e., its coefficient was not shrunk to zero). As can be seen in the figure, a part of variables were chosen for prediction in almost every CV split , whereas the others were regarded as not important and their coefficients shrunk to exactly zero. Twenty variables were suggested as consistent injury predictors ($p<0.01$) with the logistic regression model.

The suggested variables were sex, BMI, hamstring flexibility (both legs), KT1000 (dominant leg), hip flexion peak (dominant leg), medial knee displacement (both legs), vertical ground reaction force (vGRF) (both legs), height, knee flexion at IC (non-dominant leg), ankle flexion at IC (both legs), leg press 1RM, hip flexion moment peak (non-dominant leg), previous injuries of non-dominant knee, knee valgus at IC (dominant leg), knee valgus peak (non-dominant leg), and knee flexion moment peak (non-dominant leg). In the figure, these are the twenty variables with the highest frequency value.

The mean AUC-ROC value for logistic regression models was 0.65 (0.76 for the training data). The AUC-ROC values were higher ($p<0.001$) with real responses than the randomized ones (mean AUC-ROC 0.50).

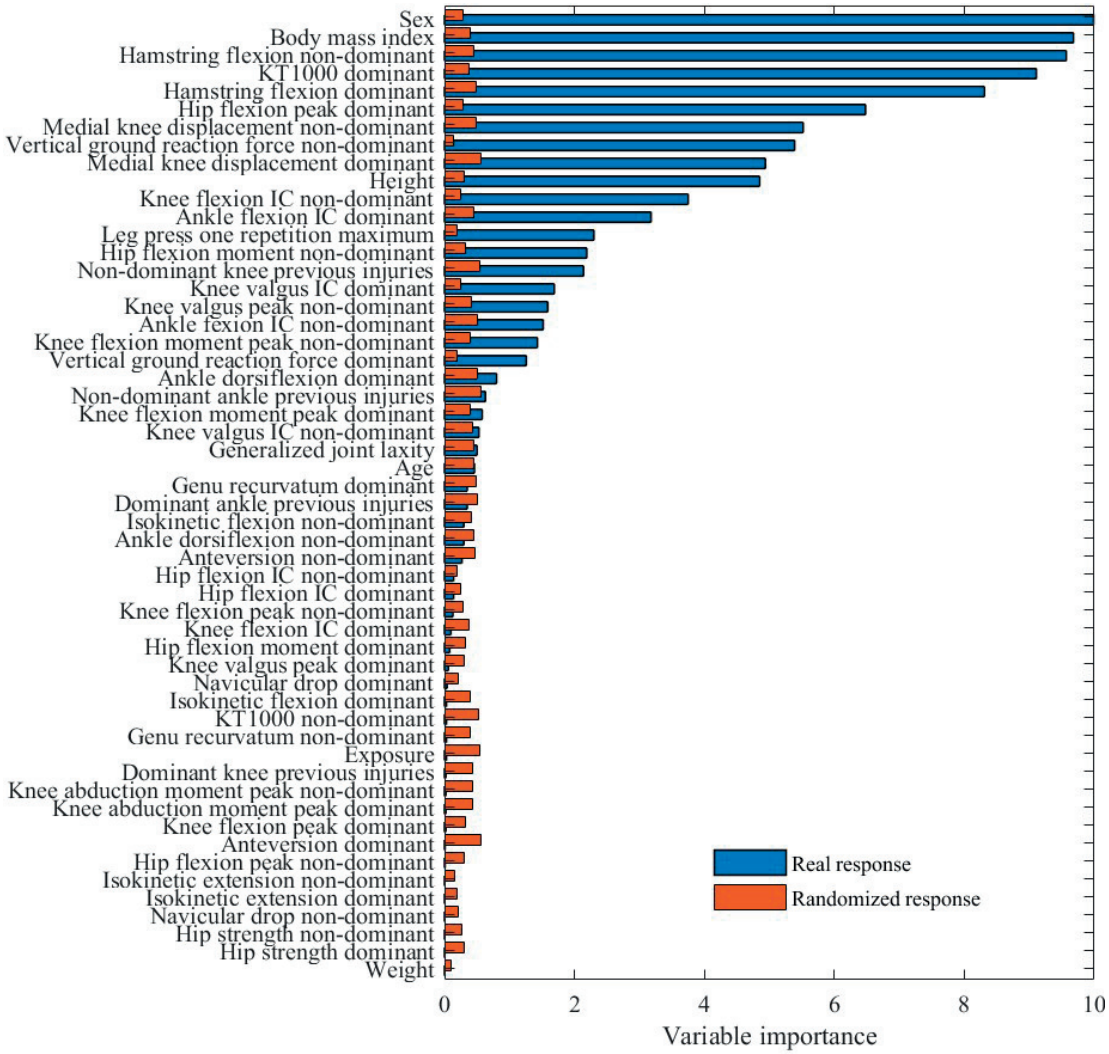


Figure 3. Variable importances for L1-regularized logistic regression. Measured as the number of times each variable was chosen over the ten CV folds. Blue bars correspond to the results with real response, red to the randomized response

3.3. Logistic regression coefficients

Whenever a variable was chosen to the logistic regression model, the direction of the coefficient was extremely consistent, always either positive or negative. Therefore, over all the folds and a hundred runs, the variable always had a similar effect on the prediction, i.e., it either increased or decreased the risk of injury. Directions of variable coefficients for the ten most often selected variables, as well as those that were found by both models, can be seen in Table 1.

Based on the coefficients, female sex contributes to bigger risk than male (male=1, female=2 in data) as well as larger BMI, lower height, and higher leg press 1RM result. Higher hamstring flexibility and vGRF of both legs increase the risk of injury. The higher value of KT1000 of dominant leg as well as higher hip flexion peak and knee flexion at IC of non-dominant leg also contribute to the injury risk. Less ankle plantar flexion (negative values) and larger knee valgus angle (negative values) of the dominant leg contribute to the higher risk. Interestingly for medial knee displacement, the direction was different between the legs. For non-dominant leg, higher medial knee displacement increased the risk but for dominant leg, a lower value increased it.

Table 1. The number of coefficients with positive, negative and zero values over the ten folds and hundred runs.

Variable	Positive	Negative	Zero
Sex	0	999	1
Body mass index	968	0	32
Hamstring flexion non-dominant	957	0	43
KT1000 dominant	911	0	89
Hamstring flexion dominant	831	0	169
Hip flexion peak dominant	648	0	352
Medial knee displacement non-dominant	552	0	448
Vertical ground reaction force non-dominant	539	0	461
Medial knee displacement dominant	0	494	506
Height	0	485	515
Knee flexion IC non-dominant	375	0	625
Ankle flexion IC dominant	318	0	682
Leg press one repetition maximum	230	0	770
Knee valgus IC dominant	0	169	831
Vertical ground reaction force dominant	126	0	874

3.4. Consistent injury predictors chosen by both methods

The following ten variables were suggested as consistent injury predictors ($p<0.01$) by both models: sex, body mass index, hamstring flexibility (non-dominant leg), KT1000 (dominant leg), hamstring flexibility (dominant leg), medial knee displacement (dominant leg), height, ankle (plantar) flexion at IC (dominant leg), leg press one repetition maximum (1RM), and knee valgus at IC (dominant leg).

4. Discussion

The purpose of this study was to utilize predictive machine learning methods to detect variables with predictive power on sport injuries. Multiple injury risk factors have been recognized in previous explanatory studies, but the predictive power of these variables remains unclear until tested on independent data. We presented a framework that detects consistent injury predictors in a data-driven manner and validates their predictive power on independent test data. This approach can be used for finding new hypotheses for injury risk factors as well as confirming the predictive power of previously recognized risk factors. Any new hypotheses should then be confirmed by domain experts in future studies, utilizing explanatory methods as well.

Despite the low predictive accuracy (AUC=0.65), a set of ten consistent injury predictor variables was detected by both models. The obtained AUC score is in line with the previous studies [6, 9, 16, 17] and confirms the difficulty of predicting sport injuries. A recently published predictive analysis study that compared different methods and their injury prediction accuracies, obtained an AUC score of 0.747 when predicting lower extremity muscle injuries in 132 male professional soccer and handball players [9]. A paper by Dower and colleagues [17] utilized time series data and artificial neural networks, achieving AUC scores between 0.75 and 0.80 on average when predicting soft tissue injuries in Australian football players.

Another study found that previously detected risk factors with explanatory power had a very poor predictive performance (median AUC scores 0.57 and 0.52) on hamstring strain injuries in 362 elite Australian footballers [16]. However, this study used a small number of variables in the prediction (three and eight). In addition, previous studies have focused solely on the prediction task, without considering the explainability of the predictive model. Explainable models, assessing, for example, the effect of each variable in prediction, are easier to trust and provide more practical information to the domain experts.

Most of the injury predictor variables suggested in our study are supported by previous research. Our results suggest that female sex, larger BMI, and lower height increased the risk of acute non-contact

knee and ankle injury. Previous explanatory research has detected similar associations with lower extremity sport injuries [2, 5, 27, 28]. For muscle flexibility, there are contradicting findings [5, 29]. Our results propose that increased hamstring flexibility of both dominant and non-dominant leg contribute to larger risk of acute non-contact knee and ankle injury.

Concerning the association between muscle strength and sport injury risk, the findings are conflicting [30, 31]. Our study found higher leg press 1RM to be associated with higher injury risk. This could be, for example, because stronger athletes exert greater forces and moments to the joints and muscles during activity; are more mature; and tend to train more and perform at higher levels. Also, our findings that increased knee laxity (KT-1000) and less ankle plantar flexion at IC of the dominant leg contribute to higher injury risk have been previously recognized [32, 33].

Our results suggest that larger knee valgus and medial knee displacement of non-dominant leg increase the risk of acute non-contact knee and ankle injury. Associations between knee valgus loading and risk of lower extremity injuries have been found previously [34]. However, our results also suggested that smaller medial knee displacement of the dominant leg increased the risk, which is contradictory to the results of the non-dominant leg. In the group of non-injures athletes, the medial knee displacement of dominant leg is notably larger than with the non-dominant leg. In the injured group, there is no such difference (see Supplementary Table 1). This side difference is causing the conflicting regression coefficients inside the framework. However, such side differences were not observed in the knee valgus angles. This might be due to the medial knee displacement being more sensitive towards the athlete rotating during landing. In our data, approximately 74% of the athletes rotated towards the side of their dominant leg during VDJ. Another possible explanation might simply be differences in the use of dominant and non-dominant leg.

Our secondary aim was to assess differences between linear and non-linear methods. In our prediction task, the predictive accuracy of the linear L1-regularized logistic regression was slightly better (AUC=0.65) than the accuracy of the non-linear random forest model (AUC=0.63). The difference is, however, negligible for drawing conclusion of their mutual superiority. The suggested injury risk

factors were largely the same for both models, but logistic regression suggested a larger set of predictors. Generally, we believe it can be beneficial to utilize a combination of methods to detect the most relevant injury risk factors.

The strength of our approach is that with predictive methods and confirmatory analysis, consistent injury predictors can be detected even from data with weak phenomena. For example, with small datasets the approach can help to avoid findings by chance. Thus, it can be useful in other sport science and medicine studies as well, even though the used data does not necessarily possess high predictive power or strong phenomena itself. Another strength is the prospective data collection of a large number of variables from a large cohort of athletes. Predictive methods utilize computational power and thus enable analysis of all relevant data and do not require exclusion based on prior assumptions. In addition, our study uses a well defined prediction outcome of moderate and severe knee and ankle injuries which risk factors have been established in explanatory research previously.

However, there are also limitations related to the used data. After baseline data was collected, the injury follow-up lasted for 12 months. Many of the collected variables might, however, change notably during this period, especially in young athletes [10]. In the future, more comprehensive data that observes short-term changes in variables should be collected as there can be changes, for example, based on the time in season and weekly training and game loads. Wearable technologies, for example, allow continuous monitoring of athletes. It can be expected that time series data from wearable devices combined with applicable predictive methods will increase the prediction accuracy as the study by Dower et al. indicated [17].

To conclude, in order to have practical value in the clinical assessment of injury risk, the predictive accuracy of the presented models that were trained on the prospective data should be improved. The models were, however, able to detect a set of consistent injury predictors. Thus, the approach can be useful for finding new hypotheses for injury risk factors as well as confirming the predictive power of risk factors found in previous explanatory studies. While the achieved predictive accuracy of our study remained relatively low (AUC=0.65), a set of ten consistent injury predictor variables was

detected by both models (sex, body mass index, hamstring flexibility, knee joint laxity, medial knee displacement, height, ankle plantar flexion at initial contact, leg press one-repetition max, and knee valgus at initial contact). The obtained accuracy is in line with previous studies and confirms that predicting sport injuries is a cumbersome task. More research is required to find risk factors that best predict injury and to include more comprehensive data. The obtained performance was similar between the linear and non-linear methods. Future research is needed to assess the suitability and performance of linear versus non-linear methods in sport injury prediction tasks.

Funding

This study was supported by the Finnish Ministry of Education and Culture, and Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (grants 9S047, 9T046, 9U044, 9N053). This work has been carried out in two projects ”Value from health data with cognitive computing” and ”Watson Health Cloud”, funded by Business Finland. Susanne Jauhiainen was funded by the Jenny and Antti Wihuri Foundation (grant 00180121). Jukka-Pekka Kauppi was funded by the Academy of Finland Postdoctoral Researcher program (Research Council for Natural Sciences and Engineering; grant 286019).

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

1 *Jacobsson J, Timpka T, Kowalski J, et al.* Prevalence of musculoskeletal injuries in Swedish elite track and field athletes. Am J Sports Med 2012; 40: 163–169

2 *Emery CA, Rose MS, McAllister JR, Meeuwisse WH*. A prevention strategy to reduce the incidence of injury in high school basketball: a cluster randomized controlled trial. Clin J Sport Med 2007; 17: 17–24

[3] *Myklebust G, Holm I, Mæhlum S, et al.* Clinical, functional, and radiologic outcome in team handball players 6 to 11 years after anterior cruciate ligament injury: a follow-up study. Am J Sports Med 2003; 31: 981–989

[4] *Bahr R, Holme I*. Risk factors for sports injuries -- a methodological approach. Br J Sports Med 2003; 37: 384–392

[5] *Murphy DF, Connolly DAJ, Beynnon BD*. Risk factors for lower extremity injury: a review of the literature. Br J Sports Med 2003; 37: 13–29

[6] *Rossi A, Pappalardo L, Cintia P, et al.* Effective injury forecasting in soccer with GPS training data and machine learning. PLoS One 2018; 13: e0201264

[7] *Bahr R*. Why screening tests to predict injury do not work—and probably never will…: a critical review. Br J Sport Med 2016; 50: 776–780

[8] *Shmueli G*. To explain or to predict? Stat Sci 2010; 25: 289–310

[9] *López-Valenciano A, Ayala F, Puerta JM, et al.* A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms. Med Sci Sports Exerc 2018; 50: 915–927

[10] *Meeuwisse WH, Tyreman H, Hagel B, Emery C*. A dynamic model of etiology in sport injury: the recursive nature of risk and causation. Clin J Sport Med 2007; 17: 215–219

[11] *Bittencourt NFN, Meeuwisse WH, Mendonça LD, et al.* Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. Br J Sport Med 2016; 50: 1309–1314

[12] *Robertson S*. Improving load/injury predictive modelling in sport: The role of data analytics. J Sci Med Sport 2014; 18: 25--26

[13] *Combrisson E, Jerbi K*. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci

Methods 2015; 250: 126–136

[14] *Biran O, Cotton C*. Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). 2017

[15] *Bellazzi R, Zupan B*. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008; 77: 81–97

[16] *Ruddy JD, Shield AJ, Maniar N, et al.* Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers. Med Sci Sports Exerc 2018; 50: 906–914

[17] *Dower C, Rafeli A, Weber J, Mohamad R*. An enhanced metric of injury risk utilizing Artificial Intelligence. In: Proceedings of the 13th annual MIT SLOAN sports analytics conference. 2019

[18] *Pasanen K, Rossi MT, Parkkari J, et al.* Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. BMJ open Sport Exerc Med 2015; 1: e000076

[19] *Harriss DJ, MacSween A, Atkinson G*. Ethical Standards in Sport and Exercise Science Research: 2020 Update. Int J Sports Med 2019; 40: 813–817

[20] *Leppänen M, Pasanen K, Kujala UM, et al.* Stiff landings are associated with increased ACL injury risk in young female basketball and floorball players. Am J Sports Med 2017; 45: 386–393

[21] *Fuller CW, Molloy MG, Bagate C, et al.* Consensus statement on injury definitions and data collection procedures for studies of injuries in rugby union. Br J Sports Med 2007; 41: 328–331

[22] *Boulesteix A-L, Janitza S, Kruppa J, König IR*. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2012; 2: 493–507

[23] *Carey DL, Ong K, Whiteley R, et al.* Predictive modelling of training loads and injury in Australian football. Int J Comput Sci Sport 2018; 17: 49–66

[24] *Breiman L*. Random forests. Mach Learn 2001; 45: 5–32

[25] *Fawcett T*. An introduction to ROC analysis. Pattern Recognit Lett 2006; 27: 861–874

[26] *Krstajic D, Buturovic LJ, Leahy DE, Thomas S*. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014; 6: 10–25

[27] *Vanderlei FM, Bastos FN, Tsutsumi GYC, et al.* Characteristics and contributing factors related to sports injuries in young volleyball players. BMC Res Notes 2013; 6

[28] *Jones BH, Bovee MW, Harris III JM, Cowan DN*. Intrinsic risk factors for exercise-related injuries among male and female army trainees. Am J Sports Med 1993; 21: 705–710

[29] *Boden BP, Dean GS, Feagin JA, Garrett WE*. Mechanisms of anterior cruciate ligament injury. Orthopedics 2000; 23: 573–578

[30] *Yamamoto T*. Relationship between hamstring strains and leg muscle strength. A follow-up study of collegiate track and field athletes. J Sports Med Phys Fitness 1993; 33: 194–199

[31] *Beynnon BD, Renström PA, Alosa DM, et al.* Ankle ligament injury risk factors: a prospective study of college athletes. J Orthop Res 2001; 19: 213–220

[32] *Woodford-Rogers B, Cyphert L, Denegar CR*. Risk factors for anterior cruciate ligament injury in high school and college athletes. J Athl Train 1994; 29: 343–346

[33] *Boden BP, Torg JS, Knowles SB, Hewett TE*. Video analysis of anterior cruciate ligament injury: abnormalities in hip and ankle kinematics. Am J Sports Med 2009; 37: 252–259

[34] *Hewett TE, Myer GD, Ford KR, et al.* Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a

prospective study. Am J Sports Med 2005; 33: 492–501

Supplementary table: Variables used in this study. For more detailed description see (1) *and supplementary material therein. VDJ stands for the 3D motion analysis for vertical drop jump.*

Variable name (unit)	Test	Mean non-injured	Mean injured	Description
Height (cm)	Anthropometric	174.05±9.22	170.15±8.12	Height
Weight (kg)	Anthropometric	64.63±10.77	64.72±8.90	Weight
BMI (kg/m^2)	Anthropometric	21.26±2.62	22.31±2.15	Body mass index
Anteversion dominant (deg)	Joint anatomy	9.05±6.08	9.06±5.61	Femoral anteversion. Measured with Craig's test (2). The athlete lies in prone position while physiotherapist passively flexes the knee to 90°. The hip is passively rotated internally and externally until the most lateral portion of the greater trochanter is palpable. In this position, the angle between the true vertical and the shaft of the tibia is measured to the nearest degree with a universal goniometer (Absolute+Axis™ Baseline® Evaluation Instruments, White Plains, NY, USA). Dominant leg.
Anteversion non-dominant (deg)	Joint anatomy	9.28±6.24	8.81±5.83	Femoral anteversion (see the description above). Non-dominant leg
Knee valgus IC dominant (deg)	VDJ	6.65±8.66	3.39±7.92	Knee valgus at initial contact (negative value refers to valgus alignment and positive value to varus alignment). Dominant leg.
Knee valgus IC non-dominant (deg)	VDJ	7.61±9.09	3.80±7.55	Knee valgus at initial contact (negative value refers to valgus alignment and positive value to varus alignment). Non-dominant leg.
Knee valgus peak dominant (deg)	VDJ	-4.49±8.29	-6.84±8.71	Peak knee valgus during contact (negative value refers to valgus alignment and positive value to varus alignment). Dominant Leg.
Knee valgus peak non-dominant (deg)	VDJ	-3.98±7.85	-5.33±8.23	Peak knee valgus during contact (negative value refers to valgus alignment and positive value to varus alignment). Nondominant.
Knee flexion IC dominant (deg)	VDJ	28.71±9.82	29.84±9.22	Knee flexion at initial contact. Dominant leg
Knee flexion IC non-dominant (deg)	VDJ	28.52±10.59	30.60±10.98	Knee flexion at initial contact. Non-dominant leg
Knee flexion peak dominant (deg)	VDJ	84.19±10.15	85.13±10.44	Peak knee flexion during contact. Dominant leg
Knee flexion peak non-dominant (deg)	VDJ	84.24±10.41	85.38±9.84	Peak knee flexion during contact. Non-dominant leg

Vertical ground reaction force dominant (N)	VDJ	1182.39±330.66	1251.04±384.07	Peak vertical ground reaction force during contact. Dominant leg
Vertical ground reaction force non-dominant (N)	VDJ	1139.01±311.46	1210.93±342.44	Peak vertical ground reaction force during contact. Non-dominant leg
Knee abduction moment peak dominant (N·m)	VDJ	-32.76±20.63	-34.27±21.87	Peak knee abduction moment during contact. Dominant leg
Knee abduct moment peak non-dominant (N·m)	VDJ	-31.06±17.88	-33.29±19.83	Peak knee abduction moment during contact. Non-dominant leg
Medial knee displacement dominant (mm)	VDJ	24.57±20.93	21.93±18.36	Medial knee displacement during contact. Dominant leg
Medial knee displacement non-dominant (mm)	VDJ	17.74±18.73	22.80±20.88	Medial knee displacement during contact. Non-dominant leg
Hip flexion peak dominant (deg)	VDJ	65.50±11.66	69.12±12.52	Peak hip flexion during contact. Dominant leg
Hip flexion peak non-dominant (deg)	VDJ	65.77±11.75	68.77±12.97	Peak hip flexion during contact. Non-dominant leg
Ankle dorsiflexion dominant (deg)	VDJ	43.32±7.07	41.81±7.30	Peak ankle dorsiflexion during contact. Dominant leg
Ankle dorsiflexion non-dominant (deg)	VDJ	42.30±6.82	41.41 ±7.90	Peak ankle dorsiflexion during contact. Non-dominant leg
Ankle flexion IC dominant (deg)	VDJ	-8.53±10.10	-6.96±11.32	Ankle flexion at initial contact (negative value refers to plantar flexion and positive value to dorsiflexion)
Ankle flexion IC non-dominant (deg)	VDJ	-8.63±9.51	-7.59±9.97	Ankle flexion at initial contact (negative value refers to plantar flexion and positive value to dorsiflexion). Non-dominant leg
Hip flexion IC dominant (deg)	VDJ	43.35±10.83	45.70±10.89	Hip flexion at initial contact. Dominant leg
Hip flexion IC non-dominant (deg)	VDJ	43.50±11.42	46.00±10.62	Hip flexion at initial contact. Non-dominant leg
Knee flexion moment peak dominant (N·m)	VDJ	135.95±44.84	140.36±37.29	Peak knee flexion moment during contact. Dominant leg
Knee flexion moment peak non-dominant (N·m)	VDJ	127.40±45.18	134.01±39.59	Peak knee flexion moment during contact. Non-dominant leg
Hip flexion moment dominant (N·m)	VDJ	205.41±76.08	219.07±84.29	Peak hip flexion moment during contact. Dominant leg
Hip flexion moment non-dominant (N·m)	VDJ	205.77±70.05	219.65±72.86	Peak hip flexion moment during contact.Non-dominant leg
Hamstring flexion dominant (degree)	Muscle extensibility	136.45±15.75	146.35±14.91	Hamstring flexibility. The athlete is lying in supine position, while the hip of the testing leg is fixed at 120° flexion. Three landmarks are placed on the leg: lateral fibular malleolus, lateral femoral epicondyle and the greater trochanter of femur. The knee is extended passively with an 8kg load (a fish scale, Salter Super Samson, Taylor Precision Products, Inc., Illinois,

				USA). A goniometer (HiRes, Baseline® Evaluation Instruments, White Plains, NY, USA) is placed to point of knee joint line and flexibility is measured as static range of motion. Dominat leg
Hamstring flexion non-dominant (degree)	Muscle extensibility	136.85±15.57	146.51±14.29	Hamstring flexibility (see the description above). Non-dominant leg
Hip strength dominant (kg)	Strength	13.00±3.49	12.35±2.74	Maximum isometric hip abductor strength, tested with a hand-held dynamometer (Hydraulic Push-Pull Dynamometer, Baseline® Evaluation Instruments, White Plains, NY, USA). Dominant leg
Hip strength non-dominant (kg)	Strength	12.74±3.56	12.23±3.30	Maximum isometric hip abductor strength (see the description above). Non-dominant leg
Isokinetic extension dominant (kg)	Strength	162.71±38.81	158.44±32.88	Maximum isokinetic strength, tested with Biodex Multi-Joint System Pro dynamometer (Biodex System 4, Biodex Medical Systems, Inc., Shirley, NY, USA), extension of dominant leg
Isokinetic extension non-dominant (kg)	Strength	156.94±36.71	156.91±30.60	Maximum isokinetic strength (see the description above), extension of non-dominant leg
Isokinetic flexion dominant (kg)	Strength	98.01±23.15	97.52±20.09	Maximum isokinetic strength (see the description above), flexion of dominant leg
Isokinetic flexion non-dominant (kg)	Strength	96.05±22.66	97.37±20.16	Maximum isokinetic strength (see the description above), flexion of non-dominant leg
Leg press one repetition maximum (kg)	Strength	170.82±50.49	174.52±45.99	One repetition maximum leg press
Navicular drop dominant (mm)		0.69±0.43	0.65±0.38	Navicular drop of dominant leg
Navicular drop non-dominant (mm)		0.71±0.45	0.66±0.36	Navicular drop of non-dominant leg
Exposure (h)	Team diary	202.53±96.88	188.08±85.26	Total exposure time from practices and games (practice and game hours). Collected individually for each participant.
Age (yr)	Baseline Questionnaire	15.68±0.50	15.93±1.99	Age
Genu recurvatum dominant (deg)	Anatomical characteristics	4.93±4.24	5.58±3.41	Genu regurvatum/knee hyperextension, dominant leg. The athlete lies in supine position and a small bolster is placed under the distal aspect of the tibia. The anterior and posterior portions of the lateral knee joint line are palpated and a mark placed at the midpoint in the sagittal plane. The most prominent aspect of the lateral malleolus and the greater trochanter are palpated and marked. A goniometer (HiRes goniometer, Baseline® Evaluation Instruments, White Plains, NY, USA) is used for measurement. The axis of the

				goniometer is positioned over the mark on the joint line, and the angle formed by a line from the lateral joint line to the greater trochanter. A line from the lateral joint line to the lateral malleolus is measured to the nearest degree with a goniometer.
Genu recurvatum non-dominant (deg)	Joint anatomy	5.17±4.08	5.33±3.56	Genu regurvatum (see the description above). Non-dominant leg
KT1000 dominant (mm)	Joint laxity	6.70±2.04	7.35±2.28	Knee joint laxity, dominant leg. The KT-1000 arthrometer (MEDmetric Corp, San Diego, California) is used to measure anterior-posterior (A-P) knee laxity (A-P displacement of the tibia relative to the femur). The athlete is in a supine position and the knee joint space line is marked medially with the knee in slightly flexed position (25° ± 5°). First, posterior-directed forces are applied to the tibia to establish a zero reference point, followed by anterior-directed forces (134 N) to measure anterior knee joint laxity (mm).
KT1000 non dominant (mm)	Joint laxity	6.75±2.12	7.04±2.30	Knee joint laxity (see the description above). Non-dominant leg
		Median non-injured	Median injured	
Generalized joint laxity (points)	Joint laxity	1	2	Generalized joint laxity, measured using the Beighton scale (3). The athlete is measured for excessive joint laxity at the trunk, the fifth fingers, thumbs, elbows, and knees. The score of four points or more on a scale of 0- 9 indicates generalized joint laxity. Two goniometers (HiRes, Baseline® Evaluation Instruments, White Plains, NY, USA) are used to measure the fifth fingers, elbows, and knees.
Dominant knee previous injuries	Baseline Questionnaire	0	0	Number of previous knee injuries in dominant leg
Non-dominant knee previous injuries	Baseline Questionnaire	0	0	Number of previous knee injuries in non-dominant leg
Dominant ankle previous injuries	Baseline Questionnaire	0	1	Number of previous ankle injuries in dominant leg
Non-dominant ankle previous injuries	Baseline Questionnaire	0	0	Number of previous ankle injuries in non-dominant leg
		Non-injured	Injured	
Sex (male-female)	Baseline Questionnaire	138-119	14-43	Sex

1. Pasanen K, Rossi MT, Parkkari J, Heinonen A, Steffen K, Myklebust G, et al. Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. BMJ open Sport Exerc Med. 2015;1(e000076).
2. Ruwe PA, Gage JR, Ozonoff MB, DeLuca PA. Clinical determination of femoral anteversion. A comparison with established techniques. J Bone Joint Surg Am. 1992;74(6):820–30.
3. Beighton PH, Solomon L, Soskolne CL. Articular mobility in an African population. Ann Rheum Dis. 1973;32(5):413.

III

PREDICTING ACL INJURY USING MACHINE LEARNING ON DATA FROM AN EXTENSIVE SCREENING TEST BATTERY OF 880 FEMALE ELITE ATHLETES

by

Susanne Jauhiainen, Jukka-Pekka Kauppi, Tron Krosshaug, Roald Bahr, Julia Bartsch, Sami Äyrämö 2022

The American Journal of Sports Medicine, 50(11), 2917-2924

DOI:10.1177/03635465221112095

Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes

Susanne Jauhiainen,*† MSc, Jukka-Pekka Kauppi,† PhD, Tron Krosshaug,‡ PhD, Roald Bahr,‡ PhD, Julia Bartsch,‡ BSc, and Sami Äyrämö,† PhD
Investigation performed at University of Jyväskylä, Jyväskylä, Finland

Background: Injury risk prediction is an emerging field in which more research is needed to recognize the best practices for accurate injury risk assessment. Important issues related to predictive machine learning need to be considered, for example, to avoid overinterpreting the observed prediction performance.

Purpose: To carefully investigate the predictive potential of multiple predictive machine learning methods on a large set of risk factor data for anterior cruciate ligament (ACL) injury; the proposed approach takes into account the effect of chance and random variations in prediction performance.

Study Design: Case-control study; Level of evidence, 3.

Methods: The authors used 3-dimensional motion analysis and physical data collected from 791 female elite handball and soccer players. Four common classifiers were used to predict ACL injuries (n = 60). Area under the receiver operating characteristic curve (AUC-ROC) averaged across 100 cross-validation runs (mean AUC-ROC) was used as a performance metric. Results were confirmed with repeated permutation tests (paired Wilcoxon signed-rank-test; $P < .05$). Additionally, the effect of the most common class imbalance handling techniques was evaluated.

Results: For the best classifier (linear support vector machine), the mean AUC-ROC was 0.63. Regardless of the classifier, the results were significantly better than chance, confirming the predictive ability of the data and methods used. AUC-ROC values varied substantially across repetitions and methods (0.51-0.69). Class imbalance handling did not improve the results.

Conclusion: The authors' approach and data showed statistically significant predictive ability, indicating that there exists information in this prospective data set that may be valuable for understanding injury causation. However, the predictive ability remained low from the perspective of clinical assessment, suggesting that included variables cannot be used for ACL prediction in practice.

Keywords: predictive methods; machine learning; prediction significance; cross-validation; motion analysis; ACL injury; team sports

Anterior cruciate ligament (ACL) injuries are a major concern in team and cutting sports, making injury prevention essential and prediction alluring.[33,38] However, while multiple potential risk factors have been suggested in the literature, whether a future ACL injury can be predicted is still a matter of controversy. Advances in data collection and storage, as well as computational power, have opened new possibilities, but there are several potential pitfalls and, consequently, also a number of important guidelines to consider to obtain reliable and valid results. The main pitfall is confusion around what is actually considered prediction in sports injury research and the difference between explanatory and predictive analyses.

Sports injury research has mainly been based on traditional statistical inference[43] with a focus on explaining or understanding phenomena of interest in the data sample at hand. This approach is also referred to as explanatory analysis.[5,45] The boundary between explanatory analysis and machine learning (ML) is not at all unambiguous, but in ML, the generalizability of a model usually takes precedence over its explainability. Generalizability means the ability to make accurate predictions on new unseen observations, and this approach is also referred to as predictive analysis.[4,45] Predictive analysis requires testing generalizability on carefully selected independent (test) data (ie, examples not involved in model fitting or selection).

Several injury prediction studies have been conducted in the past using biomechanical data in combination

The American Journal of Sports Medicine
2022;50(11):2917–2924
DOI: 10.1177/03635465221112095

with, for example, anthropometrics and strength measurements.[17,33,34] However, these studies have several limitations, making their validity questionable. First, they predict knee abduction moments as a surrogate for injury based on the assumption that high knee abduction moments predict ACL injury risk. This assumption, however, is based on explanatory analyses of data from a pilot study with <10 injury cases,[16] which is inadequate. Risk factors recognized in explanatory studies only demonstrate a statistical association with injuries but offer no evidence that they have predictive ability.[1,43,45] Moreover, the biomechanical data that these models are based on originate from a vertical drop jump (VDJ) task. Other, much larger studies have shown small or no associations between biomechanics (including knee abduction moments) and injury risk in the VDJ task.[21,47]

Another important pitfall in prediction is inadequate assessment of the generalizability of the predictive models. Many ML methods have practically infinite ability to fit in complex phenomena present in the data, given sufficient computational resources. On the other hand, this high learning capacity risks overfitting, and therefore it is critical to test the generalizability of a predictive model properly before it is implemented into practice.[22] Importantly, the role of chance results should be considered, ensuring that the predictive performance is better than chance and not just a singular random result.[18] This is essential with small and/or high-dimensional (ie, large number of variables) data sets as well as imbalanced data, which often is the case in sports injury prediction. For example, in neuroscience the problem of chance findings has been widely recognized and permutation tests have been suggested for confirming findings.[8] Moreover, the use of cross-validation, the most popular way to estimate model generalization ability in many fields, introduces randomness to the analysis and results can vary widely based on the fold division,[12] as was apparent in a recent hamstring injury prediction study.[44] An example in which these pitfalls were not considered is a recent ACL injury prediction study that did not exclude the possibility of a chance result.[52] While their study uses predictive analysis (ie, independent test data to assess generalizability), the high test accuracy (92%) against notably lower validation accuracy (70%) strongly suggests overfitting to test data either by (unconsciously) repeatedly resampling the test data set or purely by chance.

Obviously, it is also important to consider what types of data are best for injury prediction use.[19] No matter how appropriately the ML process is planned, no method is able to describe phenomena that are not captured in the data in the first place. Sports injury causation is multifactorial, indicating that a large number of variables, covering different properties and their interrelationships, should be considered.[25,28] With modern computational power, ML enables efficient analysis of a large amount of data and variables, including their interactions and nonlinear relationships, and is therefore thought to have potential in most fields, including sports injury research.[40,41] The predictive ability of previously recognized factors needs to be assessed in different settings and populations. However, periodic screening tests might not be sufficient for sports injury prediction,[1] and thus far only a few studies exist and results are variable.[19,25,42,44]

Therefore, the purpose of this study was to investigate the predictive ability of data from a large prospective ACL injury screening study, taking into account the effect of chance results and randomness from cross-validation. We applied a recently published ML approach[19] and extended the ML hypothesis space by applying different methods and preprocessing techniques for handling class imbalance in the data.

METHODS

Participants

The data used in this study were originally collected for a cohort study designed to examine risk factors for noncontact ACL injuries in female elite handball and soccer players.[21,32,35,38,46,49,50] A total of 451 soccer and 429 handball players (age, 21 ± 4 years; height, 170 ± 6 cm, weight, 66 ± 8 kg) were tested between the years 2007 and 2015. For the 2007 season, handball players with a first-team contract who were expected to play in the premier league were eligible for participation. Additionally, new players were invited for preseason testing when new teams advanced to the premier league between 2008 and 2014. From 2009, soccer players from the female premier league were also included. The study was approved by the regional committee for medical research ethics, the South-Eastern Norway Regional Health Authority, and the Norwegian Social Science Data Services, Norway. Players signed a written informed consent form before inclusion (including parental consent for players aged <18 years).

Data Collection

At baseline, each player participated in a comprehensive set of screening tests designed to assess potential

*Address correspondence to Susanne Jauhiainen, MSc, Faculty of Information Technology, University of Jyväskylä, PO Box 35, FI-40014, Jyväskylä, Finland (email: susanne.m.jauhiainen@jyu.fi).

†Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland.

‡Oslo Sports Trauma Research Center, Department of Sports Medicine, Norwegian School of Sport Sciences, Oslo, Norway.

Submitted December 30, 2021; accepted May 19, 2022.

One or more of the authors has declared the following potential conflict of interest or source of funding: S.J. was funded by the Jenny and Antti Wihuri Foundation (grant 00190110) and by the Emil Aaltonen Foundation (personal travel grant). The Oslo Sports Trauma Research Center has been established at the Norwegian School of Sport Sciences through generous grants from the Royal Norwegian Ministry of Culture, the South-Eastern Norway Regional Health Authority, the International Olympic Committee, the Norwegian Olympic Committee & Confederation of Sport, and Norsk Tipping AS. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.

Figure 1. Examples of the conducted tests (hip anteversion, knee joint laxity [KT-1000], hip abductor isometric strength, quadriceps/hamstrings isokinetic strength, leg press, marker-based static anthropometric measures, knee recurvatum, single-leg balance, navicular drop/pronation, vertical drop jump, single-leg squat, star excursion test, single-leg drop stabilization).

demographic, neuromuscular, biomechanical, anatomic, and genetic ACL injury risk factors. The screening tests were conducted at the Norwegian School of Sport Sciences in the preseason, June through August for handball and February through March for soccer. A baseline questionnaire was completed on player characteristics, elite playing experience, and history of any previous injuries to the ACL. Additionally, a variety of tests measuring anthropometrics, strength, flexibility, and balance were conducted (Figure 1). Included variables are described in Appendix Table A1 (available in the online version of this article), and for a more detailed description of the tests see Mok et al[31] and Pasanen et al.[37]

Three-dimensional motion analysis was carried out on VDJ and cutting tasks. The VDJ was performed from a 30-cm box. Players were instructed to drop off the box and perform a maximal jump upon landing with their feet on 2 separate force platforms (LG6-4-1; Advanced Mechanical Technology Inc). For more details on the VDJ protocol and setup see Krosshaug et al.[21] The sidestep cutting task was sport specific (Figure 2); the handball players performed a handball-specific faking maneuver involving a static human defender, while the soccer players performed a sidestep cutting task with a soccer throughpass. For a more detailed description of the cutting protocols see Mok et al.[31] Full-body kinematics were captured with 35 reflective markers attached over anatomic landmarks on the legs, arms, and torso.[20] From 2008 to 2011, 2 additional markers (left and right iliac crest) were used for those players whose markers on the left and right anterior superior iliac spine were occluded. From 2012 and onward, the crest markers were included for all players but only used in cases in which the anterior superior iliac spine markers were occluded. Between 2007 and 2012, eight 240-Hz infrared cameras (ProReflex; Qualisys) were used together with 2 force platforms collecting at 960 Hz. From 2012, an upgraded 16,480-Hz camera system (Oqus 4; Qualisys) was used. Marker trajectories were calculated and tracked with the Qualisys Track Manager. For a more detailed description of the motion data collection and variable extraction, see Krosshaug et al.

We recorded all complete ACL injuries among the tested players through May 2015, primarily through semiannual contact with the participating teams (manager, coach, medical staff). If any acute knee injuries occurring during regular team training or competition were reported, we contacted the injured player by telephone to obtain detailed medical data and a description of the injury situation. All ACL injuries were verified by magnetic resonance imaging and/or arthroscopy. The injury mechanisms were self-reported as contact (ie, direct contact to the lower extremity), indirect contact (ie, contact with other body parts), or noncontact, and these were categorized into 2 groups: noncontact/indirect contact or contact.[36]

Data Preprocessing

All data analyses were performed with MATLAB R2018b (MathWorks Inc) and classifiers run with the *Statistics and Machine Learning Toolbox 11.0*. For the 3-dimensional motion analysis data as well as other variables with multiple trials or measurements (star excursion, hip abduction, navicular drop), a mean of trials was calculated for analyses. For generalized joint laxity, the sum of the 9 tests included was calculated. The variables that had been

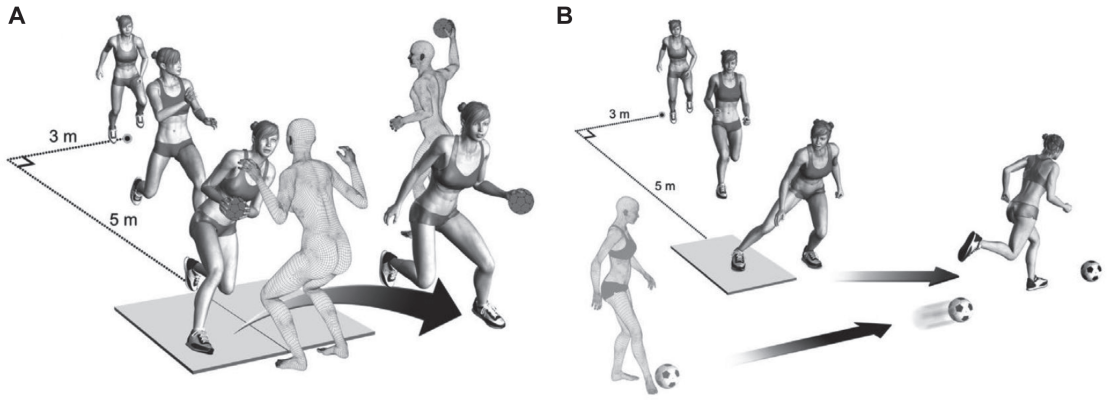


Figure 2. The testing situation of (A) the handball-specific sidestep cutting task and (B) the soccer-specific sidestep cutting task. Reprinted from Mok KM, Bahr R, Krosshaug T. Reliability of lower limb biomechanics in two sport-specific sidestep cutting tasks. *Sport Biomech*. 2017;17(2):157-167.[31] Reprinted with permission of the publisher (Taylor & Francis Ltd, http://www.tandfonline.com).

measured separately for the right and left legs were transformed to dominant (leg used for kicking a ball) and nondominant leg variables, and participants with missing dominance information were dropped (n = 14; 0 injured). Participants with a contact ACL injury were excluded (n = 9) to focus prediction on noncontact and indirect contact. Additionally, players with more than 50% of missing data (n = 66; 5 injured) were excluded, and finally, the data set used for analyses included 791 players with 60 ACL injuries and 283 variables.

To ensure validity of measurements, the most obvious outliers were identified with MATLAB's *isOutlier* function, as those that were >2 scaled median absolute deviations from the variable median (see function documentation for definition). If the function indicated possible outliers, visual confirmation was done to decide whether a value was a clear mistake or measurement error in data. In this case, only that 1 value from the particular participant was discarded. Visual analysis is a common preprocessing approach, and here it ensures as little data as possible are excluded in the cleaning process. Altogether, 47 values (0.0001%) from 16 players were discarded, with only 4 values being from injured players.

After discarding outliers, 9029 missing values (4.01% of total) existed across 478 players. These were imputed with the k-nearest-neighbor (knn) imputation with a k value of 10. Knn imputation works by finding the k most similar (measured with Euclidean distance in this study) observations and imputing the missing value with a summary metric (mean used in this study) from these k similar players. For weight and height, if a measured value was missing, a linear regression approach was used to impute a value based on the self-reported values.

Continuous variables were normalized to have a mean of 0 and SD of 1 for each column, while discrete variables were centered around 0. In addition, variations in data between sport (ie, different cut test in soccer vs handball) as well as different test years, to account for potential minor differences in testing procedures, were considered in normalization by including sport and test year in addition to labels in the stratified cross-validation split and normalizing each test group separately.

Choice of Classifiers

Four commonly used methods, random forest, L2-regularized logistic regression, and support vector machines (SVMs) with both linear and nonlinear kernel, were chosen as binary classifiers in our analyses. Random forest is a nonlinear classification and regression method that has become a standard data analysis tool in different fields such as medicine and bioinformatics[3] and has been used in sports injury research as well.[6,19,25,44] It is based on building an ensemble of multiple decision trees.[4] The model (*TreeBagger* MATLAB function) was trained with a hundred trees,[4] and Bayesian optimization[48] with *bayesopt* function was used to select the minimum number of observations per tree leaf (from 50 to 150) and the number of predictors to sample at each split (from 1 to 100). L2-regularized logistic regression, in turn, is a linear classifier that shrinks regression coefficients by penalizing the model with the L2 norm.[13] Regularization can discard irrelevant variables and possibly increase predictive performance and decrease overfitting of a model.[13] It also works well with highly correlated variables.[27] The model was trained with MATLAB's *fitclinear* function, and the optimal amount of penalization was estimated with Bayesian optimization from the default values.

SVMs are powerful and flexible classifiers[22] trying to find a hyperplane that best separates the classes from each other. They have previously been used to model nonlinear patterns and interactions in sports injury research.[6,44] In this study, we trained the SVM models with the *fitcsvm* function with both linear and nonlinear (*rbf*) kernel to assess both interactions. Hyperparameters for kernel scale (as default values from 0.001 to 1000) as well as box constraint (as default values from 0.001 to 1000) were selected with Bayesian optimization.

Data Imbalance Handling

Data imbalance means that there are clearly more observations from 1 (or more) class (majority class) than the other(s) (minority class). It is a very common and troublesome issue in the ML field,[23] and multiple different approaches to handle data imbalance have been developed and applied, including in the sports injury prediction field recently.[10,25,43,44]

Random undersampling simply means that the majority class is limited by randomly deleting observations from it, resulting in a balanced but smaller data set. Random oversampling works similarly but instead increases the observations in the minority class by randomly duplicating them, thus making the data set larger. The Synthetic Minority Oversampling Technique (SMOTE) can be used to increase the minority class observations in a balanced way.[7] It works by utilizing the existing minority examples as input and creates new observations by combining variables based on the knn algorithm. In cost-sensitive learning, the cost of misclassifying a minority observation is set higher than the cost of misclassifying a majority example. For example, in sports injury prediction (or medicine in general), not identifying an injury can be considered more harmful than incorrectly predicting some healthy athletes as injured. In practice, this is often achieved by providing the trained model a weight vector,[24] in which a higher value is set for observations corresponding to the minority class.

In this study, we experimented with the effect of random undersampling, SMOTE, as well as class weight vector in the training phase on the injury prediction task. For SMOTE, a MATLAB implementation from the MATLAB Central File Exchange[26] based on the original paper by Chawla et al[7] was used. For training class weights, each of the used methods contains an inbuilt hyperparameter option *Weights*, and a 10 times higher cost was set for the minority class.

Validation

In predictive analysis, a model's generalizability to new data has to be assessed with independent test data, that is, data that have not been used in the training of the model.[22] The most common way to do this is by splitting data into separate training and testing data or by cross-validation. K-fold cross-validation is based on randomly splitting the data into K sets and leaving each set at a time for testing while the rest of the sets are used to train a model. In general, k-fold is a common approach when data size is limited, as the complete data can be utilized for training the model.[13] In this study, we used 5-fold cross-validation.[13] Normalization and imputation of the training data were done separately inside each fold, and the test data were then normalized using coefficients estimated from the training data.

In addition, the model performance metric needs to be chosen carefully, especially with imbalanced data sets, which is often the case in sports injury prediction. Accuracy, for example, is not suitable with a class imbalance, as simply assigning all observations to the major class will yield high results. We assessed test performance with area under the receiver operating characteristic curve (AUC-ROC).[11] It is based on both true-positive and false-positive rates, and it can be used with imbalanced class distributions,[11,22] which was the case in our data. The value can be defined as excellent (0.90-1), good (0.80-0.89), fair (0.70-0.79), poor (0.60-0.69), or fail (0.50-0.59).[21,29]

Confirmatory Data Analysis

To avoid singular chance findings and ensure that the achieved results are not just due to some noise or fluctuations in data but actually present patterns significantly above a chance level, permutation tests with multiple repetitions can be utilized.[8] By repeating the analyses, the variation in results by cross-validation can be assessed. In practice, permutation tests are done by training a reference model, randomly shuffling the labels in the training phase, and then comparing it with the actual model trained with true labels. If the true models are consistently better than the random models across repetitions, the results are confirmed not to be observed by chance or just due to some noise in the data. In this study, the analysis was repeated a hundred times for both true and random models, and Wilcoxon signed-rank tests were used for a paired comparison to confirm the significance of achieved predictive performance.[19] The limit of significance was set to $\alpha = .05$, and in each cross-validation run, the fold divisions were kept the same for random and true models to allow fair pairwise comparison. Permutation tests were not run for the data imbalance handling analyses.

RESULTS

The mean AUC-ROC predictive ability was relatively consistent between the various ML methods (Table 1). Linear SVM without any imbalance handling achieved the highest mean AUC-ROC value of 0.63. For all methods, the AUC-ROC values were higher ($P < .001$) with the real responses than with the random models. With all 4 classifiers, there was a notable difference between the minimum and maximum AUC-ROC values achieved across repetitions, caused by the random cross-validation splits.

The training AUC-ROC values were very high with the random forest and SVMs, but with logistic regression, regularization seemed to control overfitting better. The test AUC-ROC values were, however, relatively similar despite differences in the training AUC-ROC. Additionally, preprocessing to handle class imbalances, that is, using SMOTE, class weight, and random undersampling, did not improve the prediction results, but results seemed similar or even slightly worse depending on the technique.

DISCUSSION

Main Findings and Clinical Relevance

This study investigated the predictive ability of a large prospective ACL injury screening data set with 60 injury

TABLE 1
AUC-ROC Values Over the 100 Repetitions[a]

	Logistic Regression	Random Forest	Linear SVM	Nonlinear SVM
Test	0.61 ± 0.02	0.57 ± 0.02	0.63 ± 0.02	0.61 ± 0.03
Min-max, range	0.57-0.65	0.51-0.63	0.55-0.67	0.53-0.69
Permuted	0.58 ± 0.03	0.52 ± 0.04	0.50 ± 0.04	0.49 ± 0.04
Training	0.86 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	0.98 ± 0.02
SMOTE	0.60 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.59 ± 0.02
Weighted	0.61 ± 0.02	0.58 ± 0.03	0.59 ± 0.02	0.60 ± 0.02
Undersampling	0.57 ± 0.03	0.50 ± 0.00	0.57 ± 0.03	0.58 ± 0.03

[a]Data are presented as mean ± SD area under the receiver operating characteristic curve (AUC-ROC), unless otherwise indicated. Permuted row correspond to the values for the random model and training row to the values for the training data. SMOTE, Synthetic Minority Oversampling Technique; SVM, support vector machine.

cases, using 4 common ML algorithms, repeated cross-validation runs, and permutation tests that will ensure reliable, consistent, and confirmed results. The results demonstrate that, even with an extensive data set, including anthropometric, clinical, neuromuscular, genetic, and sophisticated 3-dimensional biomechanical measurements, ACL injury prediction was poor (mean AUC-ROC, 0.63 for the best method). Thus, while statistically significant predictive ability was discovered, it remained too low for use in clinical risk assessment. Importantly, our results indicate that the included variables, even those identified as risk factors in previous explanatory studies, are not able to predict ACL injuries in practice. Nevertheless, associations in this prospective data set may still be valuable for understanding injury causation, but further analysis on variables is outside the scope of this paper.

Methodological Considerations

The wide range of AUC-ROC values across repetitions is notable (Table 1) and demonstrates that the use of a single random cross-validation split can lead to highly varying interpretations based on the same data and analyses, even with the current data set, which is by far the largest prospective cohort study for ACL injury in a team/ball sport. This variability was clearly visible in the results of Ruddy et al[44] as well. As cross-validation is based on randomly splitting the data into k sets, each model is trained on a different, random subsample of data and results vary. Repeated analysis can be used to handle and investigate the variation in results and reach more robust and reliable estimates for the data. Utilizing a sufficient number of repetitions is essential for obtaining a reliable estimate (eg, average AUC-ROC) for the predictive performance. Additionally, noise in data introduces randomness in results as methods might capture the noise in prediction. Noise is inevitable in any real-world data,[15] and assessing the significance of results is especially important with small data sets or with lower-performance results[8] to make sure they reflect a truly present phenomenon. Our results were confirmed with permutation tests as suggested in Combrisson and Jerbi[8] and Jauhiainen et al,[19] and despite relatively low predictive performance, there was predictive ability since the results were significantly above chance level. This confirms the presence of true phenomena, and since these relationships were captured by all models, we can be relatively confident in these results.

Importantly, studies should also report predictive performance estimates for test and/or validation data to make reliable interpretations and rule out chance results. In our study, for 3 of the methods—namely, random forest and SVMs—the training AUC-ROC was noticeably higher than the test AUC-ROC. In general, random forests should be resilient to overfitting, as the combination of multiple decision trees reduces the variance of individual trees.[4] With a hundred trained trees and the minimum leaf size of 50, this training AUC-ROC was surprisingly high, as more trees as well as larger minimum leaf size values should reduce overfitting.[4,14] With the SVMs, the *box constraint* parameter can be used to control overfitting in MATLAB so that larger values lead to fewer support vectors. Looking at the parameter values chosen by optimization, the values seem relatively high (in the level of hundreds from 0.001 to 1000) for both SVMs, meaning the separation between classes remains simpler and smoother instead of overfitting. Thus, parameter selection for all methods seems appropriate despite high training AUC-ROCs. Previous studies show that high or near-perfect training AUC-ROC values do not cause a generalization problem with classifiers used in the current study, that is, random forest and SVM.[2,9] Additionally, regularization seems to acceptably control overfitting of the logistic regression in our results, while the test AUC-ROC values are very similar compared with the other methods. This indicates that the predictive performance of our models was likely not largely affected by the high training AUC-ROC values.

The use of imbalance handling techniques before prediction did not improve the predictive performance. This could possibly be because of existing samples not being separable to begin with, in which case any resampling techniques would naturally not improve prediction. However, our AUC-ROC values were significantly higher than chance, indicating that some class separation is present in the data. In the studies by Ruddy et al[44] and López-Valenciano et al,[25] the use of SMOTE did not improve injury prediction, but random undersampling yielded slightly better results in the study by López-Valenciano et al. It seems that more studies are needed to assess the effect and necessity of imbalance handling in sport injury prediction.

Using ML for Predicting Sport Injuries: Current Status

Recently, there have been a few examples of using ML approaches to predict sports injuries from data. Ruddy et al[44] tested the predictive ability of previously recognized hamstring strain injury risk factors in 2 data sets with 186 and 176 elite Australian footballers and found them to have a failed predictive power (median AUC-ROCs, 0.58 and 0.52). Jauhiainen et al[19] predicted knee and ankle injuries from a data set with 314 young basketball and floorball players and obtained an AUC-ROC value of 0.65. López-Valenciano et al[25] used screening data with personal, psychological, and neuromuscular measures to predict muscle injuries in 122 male professional soccer and handball players and found AUC-ROC values up to 0.747. Their study, however, did not assess the stability of random k-fold division and only reported results from a singular repetition. Considering the randomness from cross-validation, class imbalance (23.7% injured), and extensive testing of different approaches, the possibility of chance findings would be important to consider in their results.[18] Rommers et al[42] achieved both precision (fraction of true injuries among those predicted as injuries) and recall (fraction of injuries that were correctly predicted) of 0.85 when predicting acute and overuse injuries in 734 elite youth soccer players with 20% holdout test data. This study was different from all previous studies in its age range (11.7 ± 1.7 years) as well as the fact that no class imbalance existed with 368 injured players (50.1% of players). They reported that the 5 most important variables that predict injury were anthropometric measures. The results indicate that injuries are possibly easier to predict accurately among teenagers during the growth spurt as well as if a more balanced data set can be collected. Taborri et al[51] predicted "ACL injury risk" (Landing Error Scoring System [LESS] score, >5)[30] with data from inertial sensors and optoelectronic bars and obtained an accuracy and F1 score of 0.96 and 95%, respectively. However, the LESS score has been shown to have no association with ACL injury with biomechanical data,[47] and its validity with wearable data has not been investigated previously. In addition, their study had a small sample size (N = 39) and did not assess the stability of random k-fold division and the possibility of chance results.

Using ML for Predicting Sports Injuries: Future Considerations and Conclusions

Considering the scale of different classification and preprocessing methods investigated in our analyses, it is possible that other tests or variables than the ones we have measured would be better for predicting ACL injuries. It has been suggested that the VDJ test is not a suitable screening test for ACL injury in female soccer and handball players.[21,32,38,49] Additionally, training and match loads were not recorded in our data. It is also possible that 1 single screening test is not suitable for injury prediction, as baseline variables might change during follow-up.[28] However, in the current data set it has previously been reported that changes in landing biomechanics were minor and that the consistency was high 2 years apart.[21,49] It has been suggested that future studies exploit more continuous monitoring of athletes and consider short-term changes in physical variables and training loads.[19] Recent studies indicate that wearable sensors and smartphone applications could be used to replace traditional laboratory motion data collection.[39] Additionally, there are predictive studies showing potential in continuous monitoring and wearable sensors in injury prediction.[10,43] Rossi et al[43] predicted noncontact injuries in the next training session or game based on recent training load measured by wearable sensors in 26 professional male soccer players. They repeated the analysis 10,000 times to assess the stability with respect to fold divisions and achieved an AUC-ROC value of 0.78 ± 0.12. While their results are promising, the study is limited by a relatively small sample size and large class imbalance (in training data, 279 noninjury examples vs 7 injury examples). Dower et al[10] predicted risk of soft tissue injuries in Australian rules football with GPS data. They achieved AUC-ROC values between 0.75 and 0.80 with repeated tests to ensure the stability of k-fold results.

CONCLUSION

Despite analyzing a large prospective data set with extensive anthropometric, clinical, genetic, neuromuscular, and biomechanical measurements, using a variety of ML methods, the predictive ability was too low for ACL injury risk assessment in clinical practice. Therefore, further studies are needed to investigate what type of data and ML approaches should be used for more accurate injury prediction.

REFERENCES

1. Bahr R. Why screening tests to predict injury do not work—and probably never will …: a critical review. *Br J Sports Med*. 2016;50(13):776-780.
2. Belkin M, Hsu DJ, Mitra P. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *arXiv*. Preprint published online October 26, 2018. Accessed December 31, 2021. doi:10.48550/arxiv.1806.05161
3. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2012;2(6):493-507.
4. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
5. Breiman L. Statistical modeling: The two cultures. *Stat Sci*. 2001; 16(3):199-231.
6. Carey DL, Ong K, Whiteley R, Crossley KM, Crow J, Morris ME. Predictive modelling of training loads and injury in Australian football. *Int J Comput Sci Sport*. 2018;17(1):49-66.
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
8. Combrisson E, Jerbi K. Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods*. 2015;250:126-136.
9. Cutler A, Zhao G. PERT—perfect random tree ensembles. *Comput Sci Stat*. 2001;33:490-497.
10. Dower C, Rafeli A, Weber J, Mohamad R. An enhanced metric of injury risk utilizing artificial intelligence. In: *Proceedings of the 13th Annual MIT SLOAN Sports Analytics Conference*. MIT Sloan; 2019.

11. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861-874.
12. Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explor Newsl*. 2010;12(1):49-57.
13. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol 1. Springer Series in Statistics. Springer; 2001.
14. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42.
15. Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: a systematic review. *Procedia Comput Sci*. 2019;161: 466-474.
16. Hewett TE, Myer GD, Ford KR, et al. Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes: a prospective study. *Am J Sports Med*. 2005;33(4):492-501.
17. Hewett TE, Myer GD, Ford KR, Paterno MV, Quatman CE. Mechanisms, prediction, and prevention of ACL injuries: cut risk with three sharpened and validated tools. *J Orthop Res*. 2016;34(11):1843-1855.
18. Hosseini M, Powell M, Collins J, et al. I tried a bunch of things: the dangers of unexpected overfitting in classification of brain data. *Neurosci Biobehav Rev*. 2020;119:456-467.
19. Jauhiainen S, Kauppi JP, Leppänen M, et al. New machine learning approach for detection of injury risk factors in young team sport athletes. *Int J Sports Med*. 2021;42(2):175-182.
20. Kristianslund E, Krosshaug T, den Bogert AJ. Effect of low pass filtering on joint moments from inverse dynamics: implications for injury prevention. *J Biomech*. 2012;45(4):666-671.
21. Krosshaug T, Steffen K, Kristianslund E, et al. The vertical drop jump is a poor screening test for ACL injuries in female elite soccer and handball players: a prospective cohort study of 710 athletes. *Am J Sports Med*. 2016;44(4):874-883.
22. Kuhn M, Johnson K. *Applied Predictive Modeling*. Vol 26. Springer; 2013.
23. Longadge R, Dongre S. Class imbalance problem in data mining review. *arXiv*. Preprint published online May 8, 2013. Accessed December 30, 2021. doi:10.48550/arXiv.1305.1707
24. López V, Fernández A, Garćia S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf Sci*. 2013;250:113-141.
25. López-Valenciano A, Ayala F, Puerta JM, et al. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Med Sci Sports Exerc*. 2018;50(5):915-927.
26. Manohar. SMOTE: synthetic minority over-sampling technique. MATLAB Central File Exchange. Accessed December 30, 2021. www.mathworks.com/matlabcentral/fileexchange/38830-smote-synthetic-minority-over-sampling-technique
27. Marquardt DW, Snee RD. Ridge regression in practice. *Am Stat*. 1975;29(1):3-20.
28. Meeuwisse WH, Tyreman H, Hagel B, Emery C. A dynamic model of etiology in sport injury: the recursive nature of risk and causation. *Clin J Sport Med*. 2007;17(3):215-219.
29. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;8(4):283-298.
30. Mikkelsen M, Peterson A, Rinkenberger T. Using the Landing Error Scoring System (LESS) to predict the risk of lower extremity injuries in athletes. *Phys Ther Sch Proj*. 2019;677.
31. Mok KM, Bahr R, Krosshaug T. Reliability of lower limb biomechanics in two sport-specific sidestep cutting tasks. *Sport Biomech*. 2017; 17(2):157-167.
32. Mørtvedt AI, Krosshaug T, Bahr R, Petushek E. I spy with my little eye … a knee about to go ''pop''? Can coaches and sports medicine professionals predict who is at greater risk of ACL rupture? *Br J Sports Med*. 2020;54(3):154-158.
33. Myer GD, Ford KR, Brent JL, Hewett TE. An integrated approach to change the outcome part I: neuromuscular screening methods to identify high ACL injury risk athletes. *J Strength Cond Res Strength Cond Assoc*. 2012;26(8):2265.
34. Myer GD, Ford KR, Khoury J, Hewett TE. Three-dimensional motion analysis validation of a clinic-based nomogram designed to identify high ACL injury risk in female athletes. *Phys Sportsmed*. 2011; 39(1):19-28.
35. Nilstad A, Petushek E, Mok KM, Bahr R, Krosshaug T. Kiss goodbye to the ''kissing knees'': no association between frontal plane inward knee motion and risk of future non-contact ACL injury in elite female athletes. *Sport Biomech*. Published online April 28, 2021. doi:10.1080/14763141.2021.1903541
36. Olsen OE, Myklebust G, Engebretsen L, Bahr R. Injury mechanisms for anterior cruciate ligament injuries in team handball: a systematic video analysis. *Am J Sports Med*. 2004;32(4):1002-1012.
37. Pasanen K, Rossi MT, Parkkari J, et al. Predictors of lower extremity injuries in team sports (PROFITS-study): a study protocol. *BMJ Open Sport Exerc Med*. 2015;1:e000076.
38. Petushek E, Nilstad A, Bahr R, Krosshaug T. Drop jump? Single-leg squat? Not if you aim to predict anterior cruciate ligament injury from real-time clinical assessment: a prospective cohort study involving 880 elite female athletes. *J Orthop Sport Phys Ther*. 2021;51(7):372-378.
39. Reenalda J, Maartens E, Homan L, Buurke JHJ. Continuous three dimensional analysis of running mechanics during a marathon by means of inertial magnetic measurement units to objectify changes in running mechanics. *J Biomech*. 2016;49(14):3362-3367.
40. Richter C, O'Reilly M, Delahunt E. Machine learning in sports science: challenges and opportunities. *Sport Biomech*. Published April 20, 2021. doi:10.1080/14763141.2021.1910334
41. Robertson S. Improving load/injury predictive modelling in sport: the role of data analytics. *J Sci Med Sport*. 2014;18:25-26.
42. Rommers N, Rössler R, Verhagen E, et al. A machine learning approach to assess injury risk in elite youth football players. *Med Sci Sports Exerc*. 2020;52(8):1745-1751.
43. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernàndez J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One*. 2018;13(7):e0201264.
44. Ruddy JD, Shield AJ, Maniar N, et al. Predictive modeling of hamstring strain injuries in elite Australian footballers. *Med Sci Sports Exerc*. 2018;50(5):906-914.
45. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25(3):289-310.
46. Sivertsen EA, Haug KBF, Kristianslund EK, et al. No association between risk of anterior cruciate ligament rupture and selected candidate collagen gene variants in female elite athletes from high-risk team sports. *Am J Sports Med*. 2019;47(1):52-58.
47. Smith HC, Johnson RJ, Shultz SJ, et al. A prospective evaluation of the Landing Error Scoring System (LESS) as a screening tool for anterior cruciate ligament injury risk. *Am J Sports Med*. 2012;40(3):521-526.
48. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst*. 2012;25:2951-2959.
49. Steffen K, Nilstad A, Kristianslund EK, Myklebust G, Bahr R, Krosshaug T. Association between lower extremity muscle strength and noncontact ACL injuries. *Med Sci Sports Exerc*. 2016;48(11):2082-2089.
50. Steffen K, Nilstad A, Krosshaug T, Pasanen K, Killingmo A, Bahr R. No association between static and dynamic postural control and ACL injury risk among female elite handball and football players: a prospective study of 838 players. *Br J Sports Med*. 2017; 51(4):253-259.
51. Taborri J, Molinaro L, Santospagnuolo A, Vetrano M, Vulpiani MC, Rossi S. A machine-learning approach to measure the anterior cruciate ligament injury risk in female basketball players. *Sensors*. 2021;21(9):3141.
52. Tamimi I, Ballesteros J, Lara AP, et al. A prediction model for primary anterior cruciate ligament injury using artificial intelligence. *Orthop J Sports Med*. 2021;9(9):23259671211027543.

IV

A HIERARCHICAL CLUSTER ANALYSIS TO DETERMINE WHETHER INJURED RUNNERS EXHIBIT SIMILAR KINEMATIC GAIT PATTERNS

by

Susanne Jauhiainen, Andrew J. Pohl, Sami Äyrämö, Jukka-Pekka Kauppi, Reed Ferber 2020

Scandinavian Journal of Medicine and Science in Sports, 30(4), 732-740

DOI:10.1111/sms.13624

MRS SUSANNE JAUHIAINEN (Orcid ID : 0000-0001-8553-8018)

Article type : Original Article

A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns.

Running title: Hierarchical clustering of injured runners

Susanne Jauhiainen[1,*], Andrew J. Pohl[2], Sami Äyrämö[1], Jukka-Pekka Kauppi[1], Reed Ferber[2,3,4].

[1] Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland;[2] Faculty of Kinesiology, University of Calgary, Calgary, Alberta, Canada;[3] Faculty of Nursing, University of Calgary, Calgary, Alberta, Canada;[4] Running Injury Clinic, Calgary, Alberta, Canada

Corresponding Author:

M.Sc. Susanne Jauhiainen,
Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FI-40014, University of Jyväskylä, Finland; Tel. +358408053652; susanne.m.jauhiainen@jyu.fi; Orcid ID: 0000-0001-8553-8018

Acknowledgements

Susanne Jauhiainen was funded by the Jenny and Antti Wihuri Foundation (grant 00180121). Reed Ferber and Andrew Pohl were funded by the Natural Sciences and Engineering Research Council of Canada (NSERC grant 1030390).

Disclosure of interest

The authors report no conflicts of interest

A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns.

Abstract

Previous studies have suggested that runners can be subgrouped based on homogeneous gait patterns, however, no previous study has assessed the presence of such subgroups in a population of individuals across a wide variety of injuries. Therefore, the purpose of this study was to assess whether distinct subgroups with homogeneous running patterns can be identified among a large group of injured and healthy runners and whether identified subgroups are associated with specific injury location. Three-dimensional kinematic data from 291 injured and healthy runners, representing both sexes and a wide range of ages (10-66 years) was clustered using hierarchical cluster analysis. Cluster analysis revealed five distinct subgroups from the data. Kinematic differences between the subgroups were compared using one-way analysis of variance (ANOVA). Against our hypothesis, runners with the same injury types did not cluster together, but the distribution of different injuries within subgroups was similar across the entire sample. These results suggest that homogeneous gait patterns exist independent of injury location and that it is important to consider these underlying patterns when planning injury prevention or rehabilitation strategies.

Keywords: Running, Injury, Kinematics, Unsupervised machine learning

Introduction

Running is a popular sport for developing and maintaining cardiovascular fitness[1,2], despite a relatively high injury rate. Estimates of the annual prevalence of lower extremity running related injuries (RRI) vary between 19.4% to 79.3%[3], while a widely accepted estimate is that 50% of runners experience an RRI annually[1]. Naturally, with such a high injury rate, the etiology of RRIs has received much attention within the gait analysis community to determine injury etiology for specific RRIs.

Multiple studies have suggested that there are similar kinematic gait patterns for runners with similar injury locations. For example, studies involving common knee RRIs, such as patellofemoral pain (PFP)[4–7] and iliotibial band syndrome (ITBS)[8,9] have reported that runners exhibit increased peak hip adduction and peak knee internal rotation. As well, similarities in ankle-related RRIs have been reported in individuals with Achilles tendinopathy (AT)[10] and those with posterior tibial tendon dysfunction[7] suggesting these injured runners exhibit increased time to peak pronation[10]. However, a recent study by Bramah et al.[11] investigated 72 injured runners, with PFP, ITBS, AT, and medial tibial stress syndrome (MTSS), and 36 healthy runners and reported that all injured runners exhibited greater contralateral pelvic drop and forward trunk lean as well as more knee extension and ankle dorsiflexion at initial contact irrespective of injury location. Therefore, further research is necessary to determine if common kinematic gait patterns exist for specific injuries and/or anatomical location of injuries.

One method in which to gain some understanding as to the etiology of RRIs may be to examine whether unsupervised machine learning techniques can identify homogenous subgroups within a large population of injured runners and healthy individuals. Unsupervised machine learning works by discovering underlying patterns and associations in datasets without any a priori information aside from a set of input variables[12]. These statistical methods have demonstrated success in uncovering underlying structure within datasets in previous sport science research[13–17]. For example, race patterns in elite swimmers[13], different golf swing patterns[14], and different gait patters in injured and healthy runners[15–17], have all been successfully identified using unsupervised machine learning methods. Another study used unsupervised machine learning to create a movement profile for healthy controls walking and assessed the deviation of patients with gait problems from this profile[18]. However, to our knowledge very few studies have utilized this approach for a large population of injured and healthy runners. Recognition of distinct gait

patterns can be utilized in injury rehabilitation protocols and in future biomechanical investigations seeking to better understand injury etiology.

Thus, the main purpose of this exploratory study was to investigate whether a large group of injured and healthy runners can be clustered into subgroups of homogeneous gait patterns based on 3D kinematic data. We hypothesized that runners with injuries at similar locations would exhibit similarities in gait patterns and consequently be identified within the same homogenous subgroups. Similarly, we assumed the gait pattern of healthy runners would be distinct from those identified within injured homogenous subgroups. A secondary purpose was to analyze differences in 3D kinematics between the formed subgroups to better understand differences in running kinematics between the subgroups.

Materials and methods

Participants

A sample of 291injured and healthy runners were queried from an existing database of running kinematics[19,20]. Participants were recruited through standard advertisements (e.g., posters, facebook posts) approved by the Ethics Board. In this study, runners were identified for inclusion provided they: (i) had 3D gait analysis performed by the same experienced clinician using the same motion capture system, (ii) were known to be injury free for the 6 months preceding data collection (25 runners) or were suffering some form of lower body injury at the time of data collection (266 runners) and (iii) contained complete kinematic data with no missing values. A standard approach to define RRI based on Yamato et al. was used "*Running-related musculoskeletal pain in the lower limbs that causes a restriction on or stoppage of running (distance, speed, duration, or training) for at least 7 days or 3 consecutive scheduled training sessions, or that requires the runner to consult a physician or other health professional."*[23]. The runners (146 females, age 39.51±11.21 years) were defined either as competitive (n=57) or recreational (n=234)[22] based on age, sex, and most recent race performance (10 km, half-marathon, or marathon) and the World Masters Association Age Grading Performance Tables[1]. Injuries were diagnosed and injury history was collated by a licensed health professional (e.g., physiotherapist, medical doctor, athletic therapist). Injuries were grouped by location with: 72

[1] http://www.usatf.org/Resources-for---/Masters/LDR/Age-Grading.aspx. Accessed May 21, 2019.

knee, 58 ankle/foot, 51 hip/pelvis, 42 thigh, 39 lower leg (shin) identified. Across the 266 injured runners, a wide variety of RRI were reported. However, the main injuries included patellofemoral pain (n=44), iliotibial band syndrome (n=29), achilles tendinopathy (n=15), plantar fasciitis (n=14), and medical tibial stress syndrome (n=12). The occurrence and distribution of these main injuries were consistent with previously reported epidemiological investigations[24]. Other injuries included specific muscle strains (e.g. gastrocnemius, hamstring, hip flexor), tendinopathies (e.g. tibialis posterior tendinopathy, patellar tendinopathy), as well as generalized joint pain. None of the injured participants described any pain during the treadmill running procedure.
Twenty-five individuals were confirmed as injury free for at least six months prior to data collection. Participant characteristics are summarized by injury location in Table 1. Data collection was approved by the University of Calgary's Conjoint Health Research Ethics Board (CHREB: REB15- 0557). Before data collection, all participants provided a written informed consent to participate.
[Table 1 somewhere here]

Data collection

Three-dimensional (3D) marker trajectory data were captured via an 8-camera VICON motion capture system (MX3+, Vicon Motion Systems Oxford, UK) at 200 Hz while participants ran on a treadmill (Bertec Corporation, Columbus, OH). Spherical retro-reflective markers were placed on anatomical landmarks and rigid plates with clusters of 3-4 markers were placed on each of seven lower body segments as per Pohl et al.[25]. Anatomical markers and segmental clusters were placed by a single examiner with over twenty years' experience in clinical gait analysis and physical therapy[26,27]. This marker-set consisted of seven rigid segments and has been reported to produce reliable kinematic waveforms[25]. To allow for unobstructed movement during running, anatomical markers were removed following a one second static trial where subjects stood upon a template with their feet positioned straight ahead and 0.3m apart with arms crossed over their chest.

Following a warmup period of 2-5 minutes, kinematic data were collected for approximately 60 seconds while participants ran at a constant self-selected speed between 1.84 and 3.37 m/s. In order to standardize the footwear condition, each participant wore the same shoes (Pegasus, Nike, Beaverton, USA). It should be noted that while the use of standardized shoes has

advantages, it might also alter running patterns as it is possible that some participants may not be accustomed to the shoes.

Data processing

Key gait events, foot strike and toe off were identified using a Principal Component Analysis (PCA) approach as described elsewhere[28,29]. Joint angles within each movement plane were extracted using 3D GAIT custom software (Running Injury Clinic Inc., Calgary, Alberta, Canada), and time normalized to 100 data points per gait cycle: 35 data points for stance and 65 data points for the swing phase.

Each subject's running pattern was described by the median for each of 62 kinematic (e.g., peak knee flexion and adduction angles, heel strike angle) and functional variables (e.g., step width, vertical oscillation, stride rate and length) extracted from each gait cycle. A minimum of ten gait cycles were included but generally approximately 30–40 consecutive running strides were collected for processing and analysis. All variables were extracted from frontal and sagittal plane motion given the limited reliability of transverse plane angles during motion capture analysis[27,30,31]. A full description of these variables is provided in Table A1 found in the appendices.

A matrix (291 subjects x 62 variables) was created with each column normalized to have a mean of zero and standard deviation of one. A PCA was performed on the data to reduce multi-collinearity between biomechanical variables. A subset of principal components (PCs) were chosen so that 80% of the total variance in the dataset was explained by the selected PCs[32].

Cluster analysis

A hierarchical cluster analysis (HCA) was used to identify subgroups with homogeneous gait patterns. A hierarchical cluster tree, a dendrogram, was formed with the *linkage*-function in *Statistics and machine learning toolbox 11.0* of MATLAB. The function was used with the Ward's linkage method and Euclidean distance. The subgroups were formed in an agglomerative manner, i.e. starting with each observation as their own subgroup and at every step pairing the two closest subgroups together until only one group remains. The final number of subgroups was chosen based on a stopping rule (a large percentage decrease in the coefficient followed by a plateau)[33,34]. The number of subgroups was also confirmed by visual inspection of the

dendrogram[16,17].

Interpretation and comparison of subgroups

After forming the subgroups, a univariate analysis of variance (ANOVA) was used to determine which PCs separated each subgroup from the others[17]. The PCs were then interpreted by calculating the loadings of variables to determine which variables comprised the PCs[35]. Demographic information (height, weight, age, and running speed) of the subgroups were also compared using ANOVA. Normality of variables was tested via a Shaphiro-Wilk test and equal variances with Levene's test and in the cases where assumptions were not met, non-parametric Kruskal Wallis tests were used instead. When significant differences occurred, post-hoc tests were performed using Tukey's test. The proportion of injuries and males/females in subgroups were compared using Chi-squared test. For all tests, a significance limit of α=0.05 was chosen and adjusted with Bonferroni's correction and Cohen's effect size d was calculated where appropriate[36].

The injury distribution of the formed subgroups was assessed with the adjusted Rand index[37]. The Rand index objectively measures the similarity between two different clusterings of the same data. If $X = \{x_1, x_2, ..., x_n\}$ is the set of observations and $P = \{P_1, P_2, ..., P_{K_1}\}$ and $P' = \{P'_1, P'_2, ..., P'_{K_2}\}$ are two different partitions of X, where n is the number of observations in data and K_1 and K_2 are the number of subgroups in partitions P and P' respectively, the Rand Index is calculated by using all possible pairs of observations in X. Defining s as the number of pairs that are clustered to the same subgroup in both P and P', and d as the number of pairs that are not clustered to the same subgroup in either P or P', finally the Rand Index can be written as $R = \frac{s+d}{\binom{n}{2}}$, where the denominator is the total number of pairs. Simply put, the index measures the proportion of similar pairings, over all possible pairs of observations. The index receives a value between 1 and 0, with 1 indicating the clusterings are exactly the same while 0 indicates that clustering do not agree on any parts. The adjusted version works similarly but is corrected for chance. The index was calculated between the clustering labels and the injury class labels with a custom Matlab script[2]. All data processing and analysis were performed on MATLAB R2016b (MathWorks Inc).

[2] https://se.mathworks.com/matlabcentral/fileexchange/49908-adjusted-rand-index. Accessed May 21, 2019.

Results

The first 16 PCs, explaining 80.98% of the total variance, were chosen as input for the HCA method. The dendrogram for the clustering results is outlined in Figure 1. Between four and five subgroups, there was a large decrease in the agglomeration coefficients (21.0% based on min–max normalized linkage distances), followed by a plateau between five and six (6.0%). Therefore, the number of subgroups was set to five and the result was also confirmed by visual inspection of the dendrogram.

[Figure 1 somewhere here]

Despite five distinct subgroups being identified (average linkage distance of 39.32 between subgroups), the population of injured and healthy runners was randomly dispersed amongst the subgroups and this was confirmed with the very low Rand index score of r=0.012 when the cluster partition and the original injury classification were compared. The proportion of injured and healthy runners was not different between the subgroups (X^2=0.53, p=0.99) and similarly there was no evidence to suggest a difference in any of the injury types/locations between the subgroups (X^2=20.20, p=0.251).

The demographics of the subgroups are described in Table 2. The proportion of males and females was different between the subgroups (X^2=53.85, p<0.01) with the proportion of males in S1 (73.7%) being higher (X^2=35.00, p<0.01, d=0.80) compared to other subgroups and similarly, the proportion of females in S5 (63.3%) was higher (X^2=31.03, p<0.01, d= 0.81). There was evidence to suggest differences in weight (X^2=61.80, p<0.01), height (X^2=22.30, p<0.01), and running speed (X^2=56.18, p<0.01), but not in age (F=2.47, p=0.18).

[Table 2 somewhere here]

There were differences between the subgroups in the five first PCs. The amount of variance explained by the individual PCs was 13.44, 12.34, 10.01, 6.53, and 6.06 percent for the first five PCs respectively. PC1 was primarily loaded on frontal plane hip variables and stride rate, vertical oscillation, and swing time. PC2 consisted of frontal plane knee variables, stride length, vertical oscillation, and swing time. PC3 was comprised of ankle and foot frontal plane variables.

PC4 consisted of variables describing ankle eversion, excursion and peak eversion velocity. PC5 consisted of heel strike angle and knee adduction excursion. Clear differences in running biomechanics between the subgroups were found.

Subgroup 1

S1 was separated from the other subgroups by PC2 ($p<0.001$, $F=135.13$, $d=2.08$) Compared to the other four subgroups, S1 had the largest peak knee adduction (-3.16±4.08 deg), the least knee abduction (-8.60±4.32 deg), and exhibited greater knee flexion (-49.01±4.03 deg). S1 also exhibited the second largest stride length (1.97±0.17 m), vertical oscillation (89±14.32 m), and swing time (0.45±0.03 s) compared to the other subgroups. Also, 73.68% of runners in S1 were males.

Subgroup 2

Subgroup S2 was separated from the other subgroups by PC1 ($p<0.001$, $F=109.69$, $d=2.05$) and PC2 ($p<0.001$, $F=33.72$, $d=1.16$). S2 exhibited the smallest knee flexion peak (-44±5.47 deg), second smallest hip adduction (8.87±3.95 deg) and knee abduction (-11.79±4.00 deg). The S2 subgroup also exhibited the highest stride rate (87.12±4.42 strikes/min) as well as the lowest swing time (0.40±0.03 s), stride length (1.66±0.16 m), and vertical oscillation (72.04±10.10 m) compared to the other subgroups.

Subgroup 3

S3 was separated from the other subgroups by PC1 ($p<0.001$, $F=92.27$, $d=2.12$). The S3 subgroup exhibited the second highest hip adduction peak (12.07±4.00 deg), hip adduction excursion (10.13±2.79 mm), and hip abduction velocity peak (160.76±44.96 deg). They also had the lowest stride rate (77.12±3.25 strikes/min), highest swing time (0.47±0.03 s), and the most vertical oscillation (104±13.10 m) compared to the other four subgroups.

Subgroup 4

S4 was separated from the others by PC3 ($p<0.001$, $F=25.44$, $d=1.20$), PC4 ($p<0.001$, $F=32.53$, $d=1.30$), and PC5 ($p<0.001$, $F=20.14$, $d=1.04$). Compared to the other four subgroups, S4 had the largest heel strike angle (17.10±4.90 deg) and largest foot progression angle (-15±4.96 deg) along with the second largest offset (42.27±8.40 % of gait cycle) and onset (13.96±2.67 % of gait cycle) rearfoot eversion.

Subgroup 5

Subgroup S5 was separated by PC1 ($p<0.001$, $F= 30.65$, $d=1.10$), PC2 ($p<0.001$, $F=36.05$, $d=1.17$), and PC3 ($p<0.001$, $F=103.80$, $d=1.75$). S5 exhibited the highest offset (54.95±17.13 % of gait cycle) and onset (15.66±3.86 % of gait cycle) rearfoot eversion, longest time to peak pronation (0.29±0.13 % of gait cycle) as well as the smallest foot progression angle (-8.45±4.54 deg) compared to the other four subgroups. S5 also demonstrated the largest hip adduction velocity peak (173.87±52.67 deg /s), hip adduction excursion (10.61±3.16 mm), and hip adduction peak (13.03±4.28 deg). Also, 78.95% of the runners in S5 were females.

[Figure 2 somewhere here]

Discussion

The primary purpose of our study was to investigate whether distinct subgroups with homogeneous running gait patterns could be identified from a large group of injured and healthy runners using an unsupervised hierarchical cluster analysis. Five subgroups were identified, however, contrary to our initial hypothesis, individuals with similar injuries (or no injury) did not cluster together. Instead, different types of injuries, and healthy control subjects, were evenly distributed across the five subgroups.

These results support previous research that has shown that there are similarities in kinematics between individuals with different injuries[11] and refutes the premise that injury location is related to similarities in gait kinematic patterns. Moreover, the gait pattern of healthy runners was not distinct of that of the injured and suggests that there is not a single ’protective gait pattern’ reducing the likelihood of developing RRI. However, future prospective studies are necessary to support or refute this premise. Regardless, our results also show that in a large group of runners with different injuries, representing both sexes and a wide distribution of ages, exhibit biomechanical running patterns that can be subgrouped into five distinct patterns.

Specific gait patterns can be observed within each subgroup. Specifically, S1 consisted of mostly male runners whose knee collapsed and flexed the most during running and ran at the fastest pace. Runners in S2 exhibited overall limb stiffness, observed as the least amount of peak knee flexion as well as second least amount of hip adduction and knee abduction. Runners in S3 had the second largest hip adduction peak angle, hip adduction excursion and hip abduction peak

velocity. S4 consisted of runners that exhibited a pronounced heelstrike and a large foot progression angle during running. Runners in S5 had the highest hip adduction peak and hip adduction excursion as well as the smallest foot progression angle, as well as the most rearfoot eversion and time to peak pronation. S5 also had a high ratio of females and they ran at the slowest pace.

The results of the current study suggest that it is possible that the traditional method of creating a "cluster" of subjects based on a pre-defined injury does not consider that variance of gait biomechanical patterns exists independent of the injury location/category. Thus, we propose that in order to discover these inter-relationships between movement patterns and injuries better, it is necessary to segment, or sub-type, according to gait patterns as an initial step in developing rehabilitation protocols and with respect to future biomechanical investigations seeking to better understand injury etiology. We also suggest that future prospective studies should employ PCA and HCA approaches for large cohorts of injured and pain-free runners in order to determine whether biomechanical sub-types, or unique homogeneous clusters, are potentially related to higher rates of injury.

Previous studies have suggested that atypical biomechanical patterns can lead to injuries by causing excessive repetitive tissue loading during running[10,38]. In addition, associations between certain injuries and kinematic gait patterns have been detected in multiple studies[4,6,10]. In support of this premise, a study by Braham et al.[11] reported that runners with different injuries all exhibited similar patterns among each other. However, this study[11] only involved 72 injured runners and 36 healthy controls and a simple logistic regression model to determine which kinematic parameters could best separate the two groups. In contrast, the results of the current study used a much larger cohort and employed an unsupervised machine learning approach to reveal that certain running patterns cannot be conclusively linked to injury location and that homogeneous kinematic subgroups exist regardless of injury location.

Our study benefits from a large cohort of injured and healthy runners along with robust data collection procedures. Moreover, all data were collected by a single examiner with over 20 years' experience. This is an especially important point when using unsupervised machine learning methods to identify subgroups, as these methods might pick up patterns originating from subtly different marker placements resulting from different examiners[39,40]. Specifically, Osis et al.[26] reported that a novice examiner, with 6-years of experience and trained by the same expert examiner used in the current study, made improvements in their consistency over the course of

one-year of training. However, systematic differences were apparent in data collected during the end of the year. Thus, future research involving a large cohort should take into consideration the number of people collecting the data and/or use appropriate feedback methods[40,41] to minimize marker placement error.

Limitations in the current study are acknowledged. First, given our data source was created by amalgamating data collected for specific purposes, running speed was not uniformly controlled within this study. Deviations from preferred or self-selected speed have been shown to result in deviations from typical gait patterns in walking[42] and future research should consider this factor. Second, the variables were averaged over the gait cycles, while variability in movement patterns has been associated with injuries in previous studies[43–45]. Future studies could benefit from considering the variability across gait cycles. Second, the current study was retrospective in nature and future research should prospectively follow runners to determine whether similar subgroups exist prior to injury development. In addition, each injury group included several types of injuries, that might have different effects on gait. Lastly, the data for the present study were collected in a laboratory setting whilst running on a treadmill. Previous studies[46,47] have suggested that a laboratory-based setting limits our ability to study the multifactorial nature of RRIs. Therefore, future studies should utilise inertial measurement units (IMUs) to quantify running gait patterns in real-world environments and determine whether these homogenous subgroups exist.

Perspective

This study showed that among a large population of runners with different injuries, representing both sexes and a wide distribution of ages, distinct subgroups exist with homogeneous running gait patterns. Interestingly, these patterns were not related to injury location, but different type of injuries were randomly distributed throughout the subgroups, together with healthy individuals. These results suggest that the location of injury is not related to specific gait kinematic patterns and this should be considered when planning future research studies or when developing rehabilitation and injury prevention strategies. Therefore, we recommend that when performing a clinical examination of an injured runner, individual presentation plays a larger role than attempting to determine whether they are exhibiting a gait pattern previously associated with a specific injury. Finally, based on the results of this study, prediction of injuries, based on whether or not an individual exhibits specific kinematic gait patterns, is not supported.

References

1. Fields KB, Sykes JC, Walker KM, Jackson JC. Prevention of running injuries. *Curr Sports Med Rep*. 2010;9(3):176-182.

2. Van Middelkoop M, Kolkman J, Van Ochten J, Bierma-Zeinstra SMA, Koes BW. Risk factors for lower extremity injuries among male marathon runners. *Scand J Med Sci Sports*. 2008;18(6):691-697.

3. Van Gent RN, Siem D, van Middelkoop M, Van Os AG, Bierma-Zeinstra SMA, Koes BW. Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. *Br J Sports Med*. 2007;41(8):469-480.

4. Luz BC, dos Santos AF, de Souza MC, de Oliveira Sato T, Nawoczenski DA, Serrão FV. Relationship between rearfoot, tibia and femur kinematics in runners with and without patellofemoral pain. *Gait Posture*. 2018;61:416-422.

5. Watari R, Kobsar D, Phinyomark A, Osis S, Ferber R. Determination of patellofemoral pain sub-groups and development of a method for predicting treatment outcome using running gait kinematics. *Clin Biomech*. 2016;38:13-21.

6. Noehren B, Hamill J, Davis I. Prospective evidence for a hip etiology in patellofemoral pain. *Med Sci Sports Exerc*. 2013;45(6):1120-1124.

7. Rabbito M, Pohl MB, Humble N, Ferber R. Biomechanical and clinical factors related to stage I posterior tibial tendon dysfunction. *J Orthop Sport Phys Ther*. 2011;41(10):776-784.

8. Ferber R, Noehren B, Hamill J, Davis I. Competitive female runners with a history of iliotibial band syndrome demonstrate atypical hip and knee kinematics. *J Orthop Sport Phys Ther*. 2010;40(2):52-58.

9. Miller RH, Lowry JL, Meardon SA, Gillette JC. Lower extremity mechanics of iliotibial band syndrome during an exhaustive run. *Gait Posture*. 2007;26(3):407-413.

10. Ogbonmwan I, Kumar BD, Paton B. New lower-limb gait biomechanical characteristics in individuals with Achilles tendinopathy: A systematic review update. *Gait Posture*.

2018;62:146-156.

11. Bramah C, Preece SJ, Gill N, Herrington L. Is there a pathological gait associated with common soft tissue running injuries? *Am J Sports Med*. 2018;46(12):3023-3031.

12. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol 1. Springer series in statistics New York; 2001.

13. Chen I, Homma H, Jin C, Yan H. Identification of elite swimmers' race patterns using cluster analysis. *Int J Sports Sci Coach*. 2007;2(3):293-303.

14. Ball KA, Best RJ. Different centre of pressure patterns within the golf stroke I: Cluster analysis. *J Sports Sci*. 2007;25(7):757-770.

15. Kulmala J-P, Äyrämö S, Avela J. Knee extensor and flexor dominant gait patterns increase the knee frontal plane moment during walking. *J Orthop Res*. 2013;31(7):1013-1019.

16. Watari R, Osis ST, Phinyomark A, Ferber R. Runners with patellofemoral pain demonstrate sub-groups of pelvic acceleration profiles using hierarchical cluster analysis: an exploratory cross-sectional study. *BMC Musculoskelet Disord*. 2018;19(1):120.

17. Phinyomark A, Osis S, Hettinga BA, Ferber R. Kinematic gait patterns in healthy runners: A hierarchical cluster analysis. *J Biomech*. 2015;48(14):3897-3904.

18. Barton GJ, Hawken MB, Scott MA, Schwartz MH. Movement Deviation Profile: A measure of distance from normality using a self-organizing neural network. *Hum Mov Sci*. 2012;31(2):284-294.

19. Phinyomark A, Petri G, Ibáñez-Marcelo E, Osis ST, Ferber R. Analysis of big data in gait biomechanics: Current trends and future directions. *J Med Biol Eng*. 2018;38(2):244-260.

20. Ferber R, Osis ST, Hicks JL, Delp SL. Gait biomechanics in the era of data science. *J Biomech*. 2016;49(16):3759-3761.

21. Phinyomark A, Hettinga BA, Osis ST, Ferber R. Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PLoS One*.

2014;9(8):e105246.

22. Clermont CA, Osis ST, Phinyomark A, Ferber R. Kinematic gait patterns in competitive and recreational runners. *J Appl Biomech*. 2017;33(4):268-276.

23. Yamato TP, Saragiotto BT, Lopes AD. A consensus definition of running-related injury in recreational runners: a modified Delphi approach. *J Orthop Sport Phys Ther*. 2015;45(5):375-380.

24. Taunton JE, Ryan MB, Clement DB, McKenzie DC, Lloyd-Smith DR, Zumbo BD. A retrospective case-control analysis of 2002 running injuries. *Br J Sports Med*. 2002;36(2):95-101.

25. Pohl MB, Lloyd C, Ferber R. Can the reliability of three-dimensional running kinematics be improved using functional joint methodology? *Gait Posture*. 2010;32(4):559-563.

26. Osis ST, Hettinga BA, Macdonald SL, Ferber R. A novel method to evaluate error in anatomical marker placement using a modified generalized Procrustes analysis. *Comput Methods Biomech Biomed Engin*. 2015;18(10):1108-1116.

27. Osis ST, Hettinga BA, Macdonald S, Ferber R. Effects of simulated marker placement deviations on running kinematics and evaluation of a morphometric-based placement feedback method. *PLoS One*. 2016;11(1):e0147111.

28. Osis ST, Hettinga BA, Leitch J, Ferber R. Predicting timing of foot strike during running, independent of striking technique, using principal component analysis of joint angles. *J Biomech*. 2014;47(11):2786-2789.

29. Osis ST, Hettinga BA, Ferber R. Predicting ground contact events for a continuum of gait types: an application of targeted machine learning using principal component analysis. *Gait Posture*. 2016;46:86-90.

30. Reinschmidt C, Van Den Bogert AJ, Nigg BM, Lundberg A, Murphy N. Effect of skin movement on the analysis of skeletal knee joint motion during running. *J Biomech*. 1997;30(7):729-732.

31. Kadaba MP, Ramakrishnan HK, Wootten ME, Gainey J, Gorton G, Cochran GVB. Repeatability of kinematic, kinetic, and electromyographic data in normal adult gait. *J Orthop Res*. 1989;7(6):849-860.

32. Jolliffe I. *Principal Component Analysis*. Springer; 1986.

33. Hair JF, Anderson Jr RE, Tatham RL, Black WC. Multivariate data analysis 7th Ed.(Global Edition). 2009.

34. Kinsella S, Moran K. Gait pattern categorization of stroke participants with equinus deformity of the foot. *Gait Posture*. 2008;27(1):144-151.

35. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459.

36. Cohen J. Statistical power analysis for the behavioural sciences. 1988.

37. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*. 1971;66(336):846-850.

38. Mousavi SH, Hijmans JM, Rajabi R, Diercks R, Zwerver J, van der Worp H. Kinematic risk factors for lower limb tendinopathy in distance runners: A systematic review and meta-analysis. *Gait Posture*. 2019.

39. Gorton III GE, Hebert DA, Gannotti ME. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait Posture*. 2009;29(3):398-402.

40. Osis ST, Kobsar D, Leigh RJ, Macaulay CAJ, Ferber R. An expert system feedback tool improves the reliability of clinical gait kinematics for older adults with lower limb osteoarthritis. *Gait Posture*. 2017;58:261-267.

41. Macaulay CAJ, Osis ST, Clermont C, Ferber R. The use of real-time feedback to improve kinematic marker placement consistency among novice examiners. *Gait Posture*. 2017;58:440-445.

42. Chung M-J, Wang M-JJ. The change of gait parameters during walking at different

percentage of preferred walking speed for healthy adults aged 20--60 years. *Gait Posture*. 2010;31(1):131-135.

43. Brown C, Bowser B, Simpson KJ. Movement variability during single leg jump landings in individuals with and without chronic ankle instability. *Clin Biomech*. 2012;27(1):52-63.

44. Hamill J, Palmer C, Van Emmerik REA. Coordinative variability and overuse injury. *Sport Med Arthrosc Rehabil Ther Technol*. 2012;4(1):45.

45. Stergiou N, Decker LM. Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Hum Mov Sci*. 2011;30(5):869-888.

46. Ahamed NU, Kobsar D, Benson LC, Clermont CA, Osis ST, Ferber R. Subject-specific and group-based running pattern classification using a single wearable sensor. *J Biomech*. 2019;84:227-233.

47. Benson LC, Clermont CA, Bošnjak E, Ferber R. The use of wearable devices for walking and running gait analysis outside of the lab: A systematic review. *Gait Posture*. 2018;63:124-138.

Tables

Table 1: Characteristics of participants in each injury group.

	Knee	Ankle/foot	Hip/Pelvis	Thigh	Lower leg	Healthy	Other injuries
Male-female	38-34	24-34	20-31	24-18	21-18	14-11	4-0
Age (years)	37.06±12.95	41.10±11.74	40.00±10.42	39.83±8.98	39.46±9.67	40.52±11.94	45.25±8.85
Height (cm)	173.53±9.31	170.35±9.70	171.07±13.34	172.95±7.78	171.54±9.58	172.87±9.14	178.08±10.23
Weight (kg)	69.80±12.27	71.78±17.12	70.53±13.32	71.00±13.23	71.24±11.46	70.81±10.68	78.15±11.22
Running speed (m/s)	2.49±0.26	2.46±0.31	2.49±0.31	2.55±0.24	2.52±0.28	2.59±0.25	2.67±0.22

*Table 2: Characteristics of the five identified subgroups. Significant differences identified: * p<0.05, **p<0.01. Mean ± standard deviation for continuous variables, count (portion in that cluster) for the injury locations.*

Subgroup	S1	S2	S3	S4	S5
Size	95	60	32	28	76
Male/Female	70-25**	22-38	21-11	15-13	16-60**
Age (years)	40.01±10.29	42.48±13.11	37.03±10.93	40.72±9.15	37.14±11.06
Height (cm)	176.45±7.74**	166.17±12.36**	177.32±8.14**	171.84±7.54	169.29±8.70**
Weight (kg)	73.13±12.17**	65.29±11.58**	74.67±12.37**	74.75±13.45	69.50±15.09
Speed (m/s)	2.65±0.25**	2.41±0.26**	2.63±0.23**	2.50±0.27	2.37±0.25**
Healthy	11 (11.6%)	6 (10.0%)	0 (0.0%)	3 (10.7%)	5 (6.6%)
Knee	25 (26.3%)	10 (16.7%)	14 (43.8%)	4 (14.3%)	17 (22.4%)
Ankle/foot	16 (16.8%)	18 (30.0%)	3 (9.4%)	7 (25.0%)	16 (21.1%)
Hip/pelvis	12 (12.6%)	10 (16.7%)	8 (25.0%)	7 (25.0%)	13 (17.1%)
Thigh	18 (19.0%)	7 (11.7%)	4 (12.5%)	4 (14.3%)	10 (13.1%)
Lower leg	12 (12.6%)	7 (11.7%)	3 (9.4%)	3 (10.7%)	14 (18.4%)

Figure legends

Figure 1: Dendrogram of the hierarchical cluster analysis. Linkage distance on the y-axis and individual runners on the x-axis. The five identified subgroups are identified by color. For clarity, not all runners are plotted on the dendrogram and the x-axis labels are omitted.

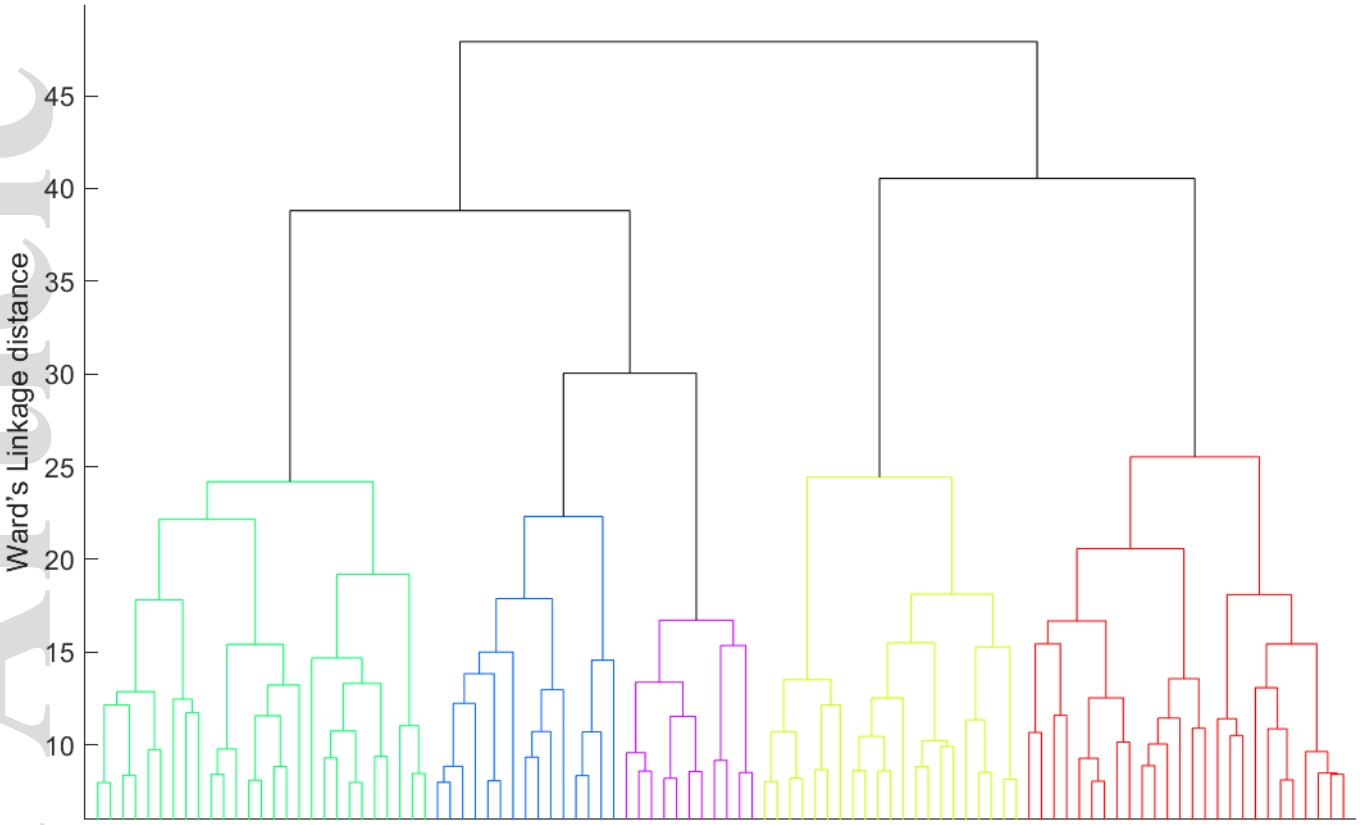
Figure 2: Boxplots highlighting the kinematics for subgroup 1 (left) to subgroup 5 (right). The vertical dashed line corresponds to the average of the study population and values have been normalized to have a mean of zero and standard deviation 1.

Appendices

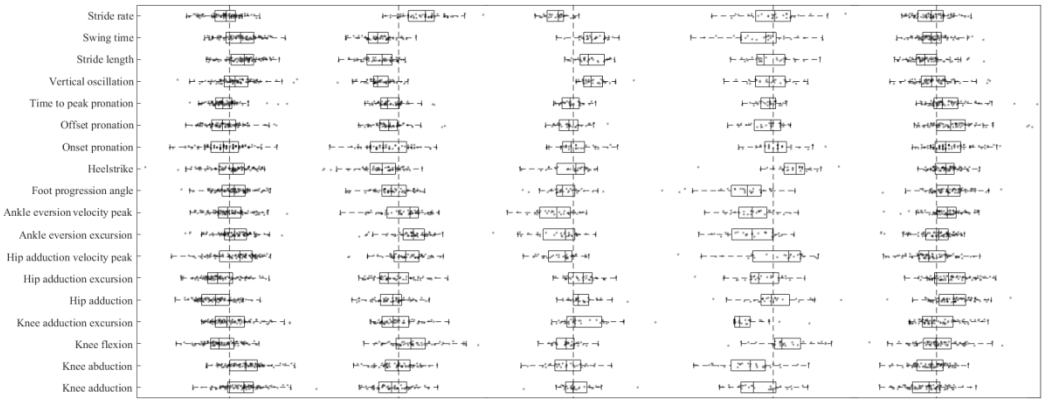
Table A1: Kinematic and functional variables used to describe each subject's running pattern:

	Joint	Variable	Description
Functional	Left/Right Side	Step Width	Side-to-side distance between left and right footsteps (m)
		Stride Rate	Number of foot strikes per minute (strikes per min)
		Stride Length	For-aft distance between left and right footfalls (m)
		Swing Time	Length of time (s) subject spent during the swing phase of gait.
		Stance Time	Length of time (s) subject spent during the stance phase of gait.
		Maximum Heel Whip	Difference between foot external rotation (deg) from toe-off to the point of maximal external rotation during swing phase.
		Vertical Oscilliation	Vertical oscillation of the center of mass (m) during complete gait cycle.
Kinematic	Left/Right Foot	Progression Angle	Angle of foot relative to direction of movement (deg) during stance phase of gait cycle.
		Heel Strike angle	Sagittal plane angle of foot at heel strike (deg).
	Left/Right Ankle	Peak Dorsiflexion	Maximum dorsiflexion angle (deg) experienced during complete gait cycle.
		Peak Eversion	Maximum eversion angle (deg) experienced during complete gait cycle.
		Time to peak pronation	The amount of time (% gait cycle) to reach peak pronation
		Eversion excursion	Difference between eversion angle at toe-off to peak eversion angle (deg).
		Peak eversion velocity	Maximum eversion angle (deg/s) subject experienced during complete gait cycle.
		Onset of Pronation	Point at which the foot reaches a pronated position (% of gait cycle)
		Offset of pronation	Point at which the foot leaves a pronated position (% of gait cycle)
	Left/Right Knee	Peak flexion Angle	Maximum knee flexion angle (deg) experienced during complete gait cycle.
		Peak adduction angle	Maximum knee adduction angle (deg) experienced during complete gait cycle.
		Adduction excursion	Distance (mm) of knee adduction excursion.
		Peak adduction velocity	Maximum knee adduction velocity (deg/s) experienced during complete gait cycle.
		Peak Abduction angle	Maximum knee abduction angle (deg) experienced during complete gait cycle.
		Abduction excursion	Difference between minimum and maximum knee abduction during stance (deg).
		Peak abduction velocity	Maximum knee abduction angle (deg) experienced during complete gait cycle.
	Left/Right Hip	Peak extension angle	Maximum hip extension angle (deg) experienced during complete gait cycle.
		Peak adduction angle	Maximum hip adduction angle (deg) experienced during complete gait cycle.
		Adduction excursion	Distance (mm) of hip adduction during gait cycle.

		Peak abduction velocity	Maximum hip adduction velocity (deg/s) experienced during complete gait cycle.
		Peak adduction velocity	Maximum hip adduction velocity (deg/s) experienced during complete gait cycle.
	Left/Right Pelvis	Peak Pelvic Drop	Maximum frontal plane angle of pelvis segment relative to horizontal (deg) experienced during complete gait cycle.
		Pelvic Drop Excursion	Difference between minimum and maximum pelvic drop during stance phase (deg).
		Peak Pelvic Drop Velocity	Maximum pelvic drop angle velocity (deg/s) experienced during stance.



sms_13624_f1.tif



sms_13624_f2.tif