UNIVERSITÀ
DEGLI STUDI
DI MILANO

SCUOLA DI SCIENZE MOTORIE

Dipartimento di Scienze Biomediche per la Salute

Corso di Dottorato in Ricerca Biomedica Integrata

XIX ciclo – Triennio accademico 2014-2016

PREDICTIVE MODELS IN SPORT SCIENCE: MULTI-DIMENSIONAL ANALYSIS OF FOOTBALL TRAINING AND INJURY PREDICTION

Tesi di Dottorato di ricerca di:

Dott. Alessio Rossi

Matricola: R10430

Docente tutor:

Prof. Giampietro Alberti

Coordinatore del Corso di Dottorato:

Prof.ssa Chiarella Sforza

“La matematica non è una scienza esatta.
Il sogno della scienza - e il suo intimo sfogo - è sempre stato quello di prevedere ciò che accadrà,
di poter calcolare con precisione assoluta, una serie di cause ed effetti che capiterà da qui in poi.
Tutti i più grandi scienziati si sono chiesti se esiste un modo per prevedere con assoluta certezza il
nostro futuro e un giorno Werner Heisenberg ha trovato la risposta.
E la risposta è: non lo sapremo mai.”

--- M. Venier ---

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank who helped me in these three years of PhD program.

First thanks are devoted to my parents and my brother for their support and encouragement without which I would not have had any hope of succeeding.

Special thanks are reserved to my PhD supervisor Prof. Giampietro Alberti who constantly believed in me providing the opportunity to undergo a PhD course three years ago. In addition, I would like to thank Prof. Andrea Caumo who transmitted me his passion for data analysis and for his invaluable contribution to my research projects. Moreover, many thanks are dedicated to Prof. Giovanni Michielon for his dear support in all parts of my PhD course.

During the third year of PhD course, I had the opportunity to spend a research period at the University of Pisa thanks to Prof. Marcello Fedon Iaia who helped me find a university that could improve my statistical skills. In Pisa I started to be interested in Machine Learning approaches thanks to my supervisors Dino Pedreschi, Luca Pappalardo and Paolo Cintia who supported me in several research projects. Special thanks are needed also for them.

I would also like to thank all of the friends met during my PhD course. In particular, a special thanks to Damiano, Athos, Luca C., Luca B. and Enrico. Thank you so much for your precious help.

Dulcis in fundus, I want to say tanks to my close friends (i.e. Nicolò, Valentina, Luca, Silvia, Stefano, Davide, Enrico, Alberto, Gabriele, Alice and Etien) and my girlfriend Margherita who endured me over these years.

PUBLICATIONS

Scientific publications I have published in my three years of PhD course are listed below:

1. Formenti D., **Rossi A.**, Calogiuri G., Thomassen T.O., Scurati R., Weydahl A. Exercise intensity and pacing strategy of cross-country skiers during a 10 km skating simulated race. Research in Sports Medicine (DOI:10.1080/15438627.2015.1005298). 2015
2. Calogiuri G., **Rossi A.**, Formenti D., Weydahl A. Sleep recovery in participants after racing in the finnmarkslop - Europe's longest dog-sled race. The Journal of Sports Medicine and Physical Fitness (PMID: 26558838) 2015
3. **Rossi A.**, Formenti D., Vitale J.A., Calogiuri G, Weydahl A. The effect of chronotype on psychophysiological responses during aerobic self-paced exercises. Perceptual and Motor Skills (DOI:10.2466/27.29.PMS.121c28x1). 2015
4. Trecroci A., Formenti D., **Rossi A.**, Esposito F., Alberti G. Acute effects of kinesio taping on a 6-s maximal cycling sprint performance. Research in Sports Medicine (DOI:10.1080/15438627.2016.1258644). 2016
5. Trecroci A. Milanović Z., **Rossi A.**, Broggi M., Formenti F., Alberti G. The effectiveness of SAQ training on physical performance and reactive agility in young soccer players. Research in Sports Medicine (DOI:10.1080/15438627.2016.1228063). 2016
6. Formenti D., Ludwig N., **Rossi A.**, Trecroci A., Alberti G., Gargano M., Merla A., Ammer K., Caumo A. Skin temperature evaluation by infrared thermography: comparison of two image analysis methods under non steady-state condition. Infrared Physics & Technology (DOI: 10.1016/j.infrared.2016.12.009). 2016

Scientific publications under review are listed below:

1. **Rossi A.**, Formenti D., Chirico M., Iaia F.M., Alberti G. The effect of slow strip set resistance training on muscular strength, power performance, and sport specific movement. Science & Sports (UNDER REVIEW)
2. Trecroci A., Formenti D., **Rossi A.**, Esposito F., Alberti G. Short-term delayed effects of kinesio taping on maximal cycling sprint. Research in Sports Medicine (UNDER REVIEW).
3. Cavaggioni L, **Rossi A.**, Adamo G., Mocciola S., Vago P., Alberti G. Functional movement screen as indicator for sedentary behaviour in a young population. Journal of School Nursing (UNDER REVIEW)
4. **Rossi A.**, Incognito O., Castellucci A., Gigante A., Colombo A., Alberti G. Well-being as a coping factor against excess in sports practice. Journal of School Health (UNDER REVIEW)
5. Alberti G., **Rossi A.**, Roi G.S. Reference centile curves for screening body mass index and body postural stability in north Italian school children and adolescents aged 6-18 years. Perceptual and Motor Skills (UNDER REVIEW)
6. **Rossi A.**, Formenti D., Alberti G., Caumo A. Detection of the best-fit nonlinear regression model in physiological responses: a practical application. Research in Sports Medicine (UNDER REVIEW)
7. **Rossi A.**, Pappalardo L., Cintia P., Savino M., Alberti G., Iaia F.M. Injury prediction in elite football players by machine learning process. PlosOne (UNDER REVIEW)

Scientific publications I have presented as oral presentations at congresses are listed below:

1. **Rossi A.**, Perri E., Trecroci A., Savino M., Alberti G., Iaia F.M. Characterization of in-season elite football trainings by GPS features – The identity card of a short-term football training cycle. Paper accepted at ICDM Barcelona 2016.

Scientific publications I have presented as poster presentations are listed below:

2. Formenti D., **Rossi A.**, Thomassen T.O., Weydahl A. Intensity pacing strategy and downhill strategy during a cross country ski race. Abstract ECSS Barcellona 2013 (poster).
3. **Rossi A.**, Calogiuri G., Formenti D., Vitale J.A., Weydahl A. The chronotype can influence the perceived exertion during self-paced exercise performed at different times of day. Abstract Sismes Pavia 2013 (poster).
4. Chirico M., **Rossi A.**, Formenti D., Alberti G. Resistance training with slow movement in wing chung martial artists. Abstract ECSS Amsterdam 2014 (poster).
5. **Rossi A.**, Romanò V., Boccolini G., Alberti G. Influence of reaction time in table tennis players. Abstract ECSS Amsterdam 2014 (poster).
6. **Rossi A.**, Chirico M., Formenti D., Trecroci A., Alberti G. The slow strip set resistance training. Abstract Sismes Napoli 2014 (poster).
7. Trecroci A., **Rossi A.**, Formenti D., Esposito F., Alberti G. Effects of a task-specific warm-up on a single-sprint cycling performance. Abstract Sismes Napoli 2014 (poster)
8. Alberti G., Boccolini G., **Rossi A.** CMJ 2.1: functional test during post injury period in football players. Abstract Isokinetic Londra 2015 (poster).
9. **Rossi A.**, Boccolini G., Alberti G. Valuation of football players level by lower limbs strength using CMJ 2.1. Abstract ECSS Malmo 2015 (mini-oral).

10. Scurati R., Formenti D., **Rossi A.**, Invernizzi P.L., Michielon G. Acute effects of low-intensity resistance training with slow movement in swimming: a pilot study. Abstract ECSS Malmo 2015 (poster).
11. **Rossi A.**, Formenti D., Alberti G. Effects of three different training programmes on instep kick in preadolescent football players. Abstract Isokinetic Londra 2016 (poster).
12. Alberti G., **Rossi A.**, Roi G.S. Reference centile curves for screening body mass index and body postural stability in football players aged 8-18 years. Abstract Isokinetic Londra 2016 (poster).
13. Munerati P., **Rossi A.**, Cavaggioni L., Formenti D., Alberti G. Relationship between knee angle in squat test and defence or reception efficacy in amateur volleyball players. Abstract ECSS Vienna 2016 (poster).
14. Riva G., **Rossi A.**, Bonfanti L., Formenti D., Alberti G. The influence of knee joint angle and time of force application on vertical jump height during volleyball commit-block. Abstract ECSS Vienna 2016 (poster).
15. Donghi F., **Rossi A.**, Bonfanti L., Formenti D., Alberti G. Validation of the microsoft kinect™ for evaluating knee joint angle in squat exercise. Abstract ECSS Vienna 2016 (poster).
16. **Rossi A.**, Formenti D., Trecroci A., Roi G.S., Alberti G. Relationship between vertical jump and body postural stability in males and females aged 6-18 years. Abstract ECSS Vienna 2016 (poster).
17. Formenti D., Ludwig N., Trecroci A., **Rossi A.**, Fernandez-Cuevas I., Gargano M., Caumo A., Alberti G. Has kinesio tape a thermal effect on sprint cycling performance? a thermographic study. Abstract QIRT Gdansk 2016 (poster).
18. Ludwig N., Formenti D., **Rossi A.**, Trecroci A., Gargano M., Alberti G. Assessing Facial Skin Temperature Asymmetry with Different Methods. Abstract QIRT Gdansk 2016 (poster).
19. **Rossi A.**, Cavaggioni L., Formenti D., Raimondi M., Alberti G. Comparison between unilateral and bilateral lower limb strength trainings. Abstract Sismes Roma 2016 (poster).

20. **Rossi A.**, Pappalardo L., Cintia P., Pedreschi D., Iaia F.M., Alberti G. The importance of GPS features to describe elite football training. Abstract Sismes Roma 2016 (poster).

ABSTRACT

Due to the fact that team sports such as football have a complex multidirectional and intermittent nature, an accurate planning of the training workload is needed in order to maximise the athletes' performance during the matches and reduce their risk of injury. Despite the evaluation of external workloads during trainings and matches has become more and more easier thanks to the advent of the tracking system technologies such as Global Position System (GPS), the planning of the best training workloads aimed to obtain the higher performance during the matches and a lower risk of injury during sport stimuli, is still a very difficult challenge for sport scientists, athletic trainers and coaches. The application of machine learning approaches on sport sciences aims to solve this crucial issue. Hence, the combination between data and sport scientists' peculiarities could maximize the information that can be obtained from the football training and match analysis. Thus, the aim of this thesis is to provide examples of the application of the machine learning approach on sport science. In particular, two studies are provided with the aim of detecting a pattern during in-season football training weeks and predicting injuries.

For these studies, 23 elite football players were monitored in eighty in-season trainings by using a portable non-differential 10 Hz global position system (GPS) integrated with 100 Hz 3-D accelerometer, a 3-D gyroscope, and a 3-D digital, Northern Ireland compass (STATSports Viper). Information about non-traumatic injuries were also recorded by the club's medical staff. In order to detect a pattern during the in-season training weeks and the injuries, Extra Tree Random Forest (ETRFC) and Decision Tree (DT) Classifier were computed, respectively.

In the first study it was found that the in-season football trainings follow a sinusoidal model (i.e. zig-zag shape found in autocorrelation analysis) because their periodization is characterized by repeated short-term cycles which are constituted by two parts: the first one (i.e. trainings long before the match) is constituted by high training loads, and the second one (i.e. trainings close to the match) by low ones. This short-term structure appears to be a strategy useful both to facilitate the decay of

accumulated fatigue from high training loads performed at the beginning of the cycle and to promote readiness for the following performance. As a matter of fact, a patter was detected through the in-season football training weeks by ETRFC. This machine learning process can accurately define the training loads to be performed in each training day to maintain higher performance throughout the season. Moreover, it was found that the most important features able to discriminate short-term training days are the distance covered above 20 $W \cdot kg^{-1}$, the acceleration above 2 $m \cdot s^{-2}$, the total distance and the distance covered above 25.5 $W \cdot Kg^{-1}$ and below 19.8$Km \cdot h^{-1}$. Thus, in accordance with the results found in this study, athletic trainers and coaches may use machine learning processes to define training loads with the aim of obtaining the best performance during all the season matches.

Players' training loads discrepancy in comparison with the ones defined by athletic trainers and coaches as the best ones to obtain enhancement in match performance, might be considered an index of individuals' physical issue, which could induce injuries. As a matter of fact, in the second study presented in this thesis, it was found that it is possible to correctly predict 60.9% of the injuries by using the rules defined by DT classifier assessing training loads in a predictive window of 6-days. In particular, it was found that the number of injuries that the player suffered through the season, the total number of Acceleration above 2 $m \cdot s^{-2}$ and 3 $m \cdot s^{-2}$, and the distance in meters when the Metabolic Power (Energy Consumption per Kilogramme per second) is above the value of 25.5 W/Kg per minute, are the most important features able to predict injuries. Moreover, the football team analysed in this thesis should keep under control the discrepancy of these features when players return to the regular training because of the numerous fall-backs into injuries that have been recorded. Thus, this machine learning approach enables football teams to identify when their players should pay more attention during both trainings and matches in order to reduce the injury risk, while improving team strategy.

In conclusion, Machine Learning processes could help athletic trainers and coaches with the coaching process. In particular, they could define which training loads could be useful to obtain enhancement in sport performance and to predict injuries. The diversities of coaching processes and

physical characteristics of the football players in each team do not permit to make inferences on the football players' population. Hence, these models should be built in each team in order to improve the accuracy of the machine learning processes.

Keywords – Machine Learning process; training loads; injury prediction; GPS.

PART 1: INTRODUCTION

1.1. THE ERA OF BIG DATA

The term 'Big Data' reflects the enormous size of dataset that we are now producing worldwide. We are on the edge of an era where a large amount of data are obtained from sensors ubiquitously present in the world and which produce a fingerprint of what surround us (Hilbert and López 2011). Analyses of these large amounts of data are so complex that traditional statistics is considered inadequate. As a matter of fact, the term 'Big Data' is often meant as predictive analytics or to other advanced data mining methods useful to examine huge datasets (Cavanillas, Curry, and Wahlster 2016). The innovative technology of the Big Data offers a new way to extract information from the tsunami of data that is sweeping us. The ability to manage and extract information from big datasets permits to have advantages on competitors. In particular, accuracy in the predictive model created from historical evidence may lead to take more focused decisions on future actions.

Several experts of diverse research fields have already embarked on the Big Data analysis. Healthcare, public care, finance and insurance are the first sectors that have an impact from big datasets ("Community Cleverness Required" 2008; Reichman, Jones, and Schildhauer 2011). The impact of Big Data has gone beyond these sectors. The explosion of available data is increasing in all research fields producing what is call Data Science (Hey, Tansley, and Tolle 2009). Data Science aims to extract knowledge from data by interdisciplinary fields such as mathematics, statistics and computer science (i.e. probability models, machine learning, statistical learning, data mining, pattern recognition, artificial intelligence and high performance computing) (Choi 2014). The development of machine learning applied on big datasets has enhanced the growth and importance of Data Science.

Machine learning is focused on pattern recognition and computational learning theory in artificial intelligence. This methodology tries to create algorithms able to learn from and make predictions on Big Data (Flach 2012). In particular, Peter Flatch defines machine learning as something "…all about

using the right features to build the right models that achieve the right task" (Flach 2012). In essence, Flatch tries to explain that by using the correct language (i.e. features) it is possible to solve a problem (i.e. task) through a specific machine learning algorithm (i.e. model).

Hence, this methodology allows researchers, data scientists, engineers, and analysts to predict results and to make decision through the learning of historical relationships and trends in the data. For this reason, all research fields can obtain advantages by using machine learning in order to make inferences on data.

1.1.1. BIG DATA

John Mashey coined the term Big Data in 1990s. It was defined as large datasets so large and complex that it is hard to analyse using "standard" data processing and most common statistical software (Haughton et al. 2015; Snijders, Matzat, and Reips 2012). Hence, Big Data requires specific techniques and technologies able to reveal insights from huge datasets that are diverse, complex, and of a massive scale (Hashem et al. 2015).

Volume, Velocity, Variety and Value are the 4 V that characterized Big Data (Gantz and Reinsel 2011) (Figure 1.1). This categorization is widely recognized because it highlights the meaning and necessity of big data. In detail:

- *Volume* refers to all type of data that can be generated from diverse sources growing continuously in size (O'Leary 2013).
- *Velocity* refers to the speed at which data is created, transferred and processed. Having Big Data, the speed in which data can be accessed, analysed and presented is crucial (Berman 2013; O'Leary 2013).
- *Variety* refers to different types of data collected via sensor (e.g. Global Position System, Accelerometer and Gyroscope), smartphones (e.g. pictures, contacts and

games) or social networks (e.g. Facebook, Instagram and Twitter). These data could include video, image, text and number, in either structured (e.g. numeric, currency, alphabetic, name, date and address) and unstructured (e.g. photos, graphic images, videos, webpages, PDF files, PowerPoint presentations, emails, and word processing documents) format (O'Leary 2013).

- *Value* is the most important part of Big Data. It refers to process that permits to detect information and patterns hidden into the records (Hashem et al. 2015; O'Leary 2013).

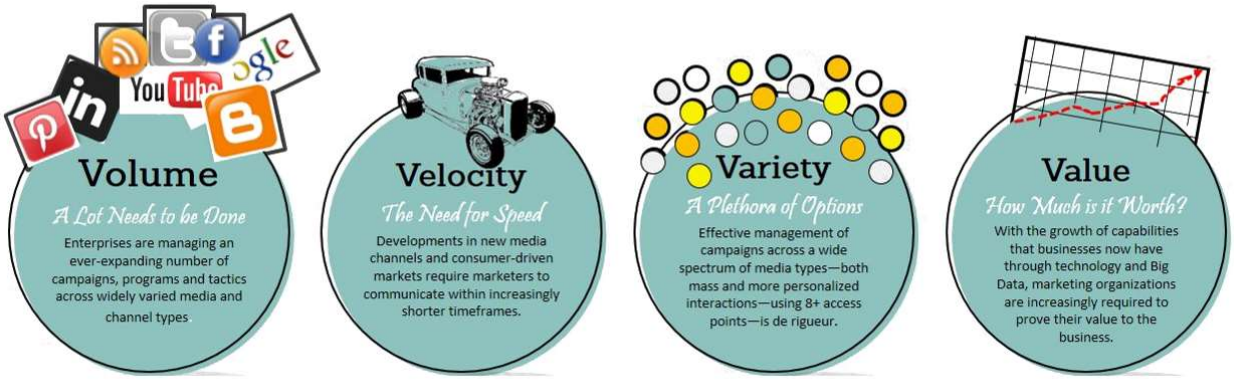

Figure 1.1. 4V of Big Data (Dunn and Coffe 2013)

In order to better understand their characteristics, Big Data can be classified into different categories. This classification is important due to the fact that there are different kinds of large-scale data in the cloud (Hashem et al. 2015). Five different aspects were taken into consideration to classify Big Data: data sources, content format, data stores, data staging, and data processing. Hence, it is possible to record data from several devices (e.g. Global Position System, questionnaire and sensors) that are organized on different formats (i.e. Structured, semi-structured and unstructured; Table 1.1) and are stored with different layouts (see Data Stores in Table 1.1). In this way, data can be organized (i.e. Cleaning, transform and normalization; Table 1.1) in order to make inferences and predictions on a specific problem to solve.

Table 1.1. Various categories of big data (Hashem et al. 2015)

	Classification	Description
Content Format	*Structured*	Structured data are often managed Structured Query Language, a programming language created for managing and querying data in RDBMS. Structured data are easy to input, query, store, and analyse. Examples of structured data include numbers, words, and dates.
	Semi-structured	Semi-structured data are data that do not follow a conventional database system. Semi-structured data may be in the form of structured data that are not organized in relational database models, such as tables. Capturing semi-structured data for analysis is different from capturing a fixed file format. Therefore, capturing semi-structured data requires the use of complex rules that dynamically decide the next process after capturing the data.
	Unstructured	Unstructured data, such as text messages, location information, videos, and social media data, are data that do not follow a specified format. Considering that the size of this type of data continues to increase through the use of smartphones, the need to analyse and understand such data has become a challenge.
Data Stores	*Document-oriented*	Document-oriented data stores are mainly designed to store and retrieve collections of documents or information and support complex data forms in several standard formats, such as JSON, XML, and binary forms (e.g., PDF and MS Word). A document-oriented data store is similar to a record or row in a relational database but is more flexible and can retrieve documents based on their contents (e.g., MongoDB, SimpleDB, and CouchDB).
	Column-oriented	A column-oriented database stores its content in columns aside from rows, with attribute values belonging to the same column stored contiguously. Column-oriented is different from classical database systems that store entire rows one after the other.
	Graph database	A graph database, such as Neo4j, is designed to store and represent data that utilize a graph model with nodes, edges, and properties related to one another through relations.
	Key-value	Key-value is an alternative relational database system that stores and accesses data designed to scale to a very large size. Dynamo is a good example of a highly available key-value storage system. Similarly, a scalable key-value store supports transactional multi-key access using a single key access supported by key-value for use in G-store designs. A scalable clustering method to perform a large task in datasets. Other examples of key-value stores are Apache Hbase, Apache Cassandra, and Voldemort. Hbase uses HDFS, an open-source version of Google's BigTable built on Cassandra. Hbase stores data into tables, rows, and cells. Rows are sorted by row key, and each cell in a table is specified by a row key, a column key, and a version, with the content contained as an un-interpreted array of bytes.
Data Staging	*Cleaning*	Cleaning is the process of identifying incomplete and unreasonable data.
	Transform	Transform is the process of transforming data into a form suitable for analysis.
	Normalization	Normalization is the method of structuring database schema to minimize redundancy.

1.1.2. MACHINE LEARNING

Herbert Simon, who won the Nobel Prize in Economics in 1978, defined Machine Learning as "…computer programs that automatically improve their performance through experiences". As a matter of fact, it is a subfield of Computer Science focused on the study of pattern recognition and computational learning theory in artificial intelligence (Flach 2012).

As explained before, Machine Learning refers to everything about using the right features to build the right models that achieve the right tasks (Flach 2012) (Figure 1.2). Fundamentally, features are the individual measurable proprieties of the phenomenon being observed. The accurate choice of features typologies (i.e. dependent and independent) is crucial to create a useful algorithm in pattern recognition, classification and regression. In addition, the task is the problem we aim to solve using features, and the model is the output produced by the machine learning algorithm applied on training data. Hence, the selection of the correct model to solve our task is the second crucial aspect to define a machine learning algorithm.

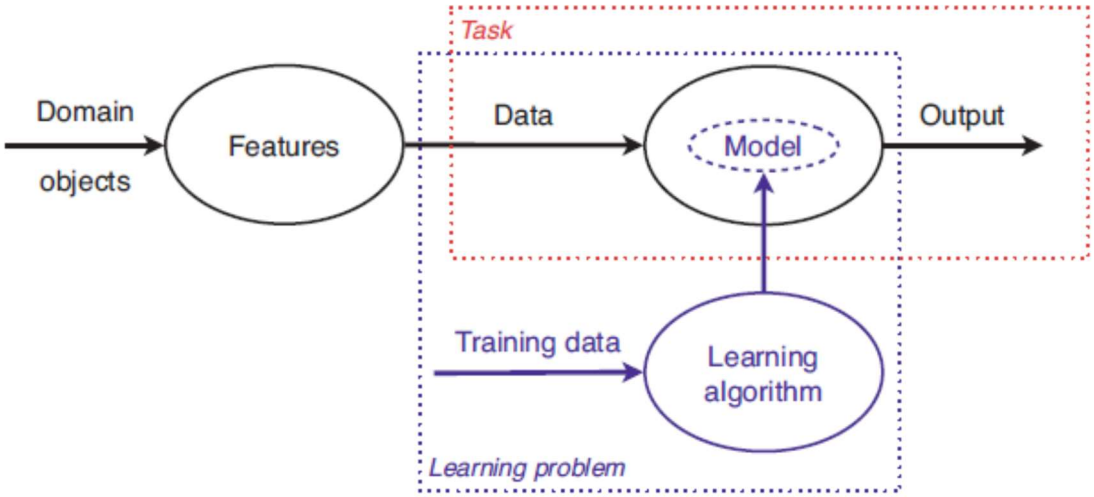

Figure 1.2. An overview of how machine learning works to solve problems. A task (red box) requires an appropriate model to produce outputs. Obtaining outputs from features is the aim of the learning problem (blue box) (Flach 2012).

Machine learning algorithms are classified into two categories in accordance with the task that we want to solve (Figure 1.3):

1. Supervised machine learning algorithms aim to map function able to predict output features (y) based on input ones (x):

$$y = f(x)$$

They are called supervised because the process created to define an algorithm learns from a labelled dataset. In particular, these machine learning processes create algorithms able to discriminate labels (y) from the training datasets (x).

Supervised machine learning problems could be further classified into two subgroups (Figure 1.3):

- Regression problems, when the output features are continuous, such as blood oxygenation, lactate, power or one maximum repetition.
- Classification problems when the output features are categorical such as injury/no-injury, goal/no-goal or male/female.

2. Unsupervised machine learning algorithms aim to create models from the distribution of input data (x) to obtain information about them. As a matter of fact, unsupervised dataset has only input features (x) without corresponding output one (y).

   Like supervised machine learning problems, Unsupervised machine learning problems could be further classified into two subgroups (Figure 1.3):

   - Clustering problems, when it is necessary to discover subgroups in the dataset, such as grouping athletes by training/physiological characteristics.
   - Association rules learning problems, when it is necessary to discover rules describing a large part of the dataset, such as athletes that when performing a particular training (e.g. high intensity resistance training) also perform a particular stretching modality (e.g. ballistic stretching).

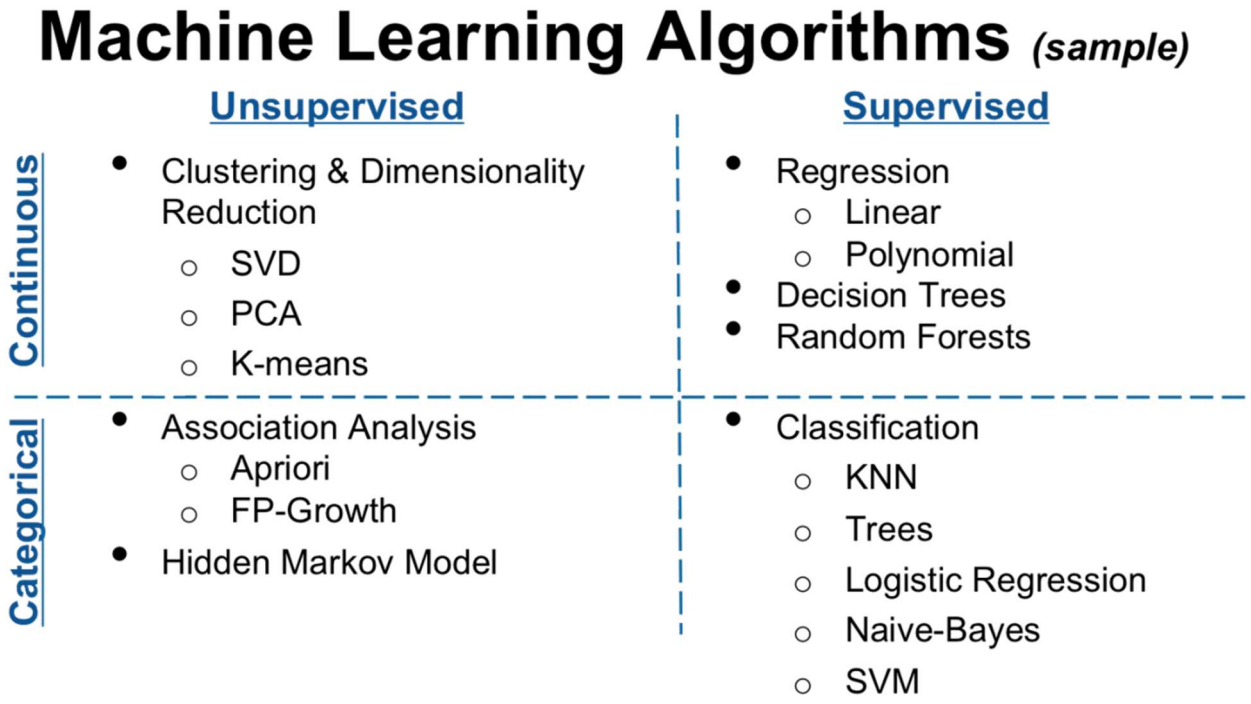

Figure 1.3. Machine learning problems classification and useful algorithms.

1.1.3. DATA ANALYSIS USING PYTHON

The dimensionality of datasets is one of the criterions that should orient us toward a specific analysis software. Another important criterion to choose the more appropriate software to analyse a particular dataset is the statistical analysis required to make inference about it. Python, R and C++ are some of the programming languages that often meet these two criterions. Due to the dimension of the dataset and the statistical analysis required, Python was chosen as the most useful programming language to analyse data in this thesis.

Python is a programming language conceived in the late 1980s by Guido van Rossum at Centrum Wiskunde & Informatics (Nederlands). From that moment it has been widely used by programmers to write programs on both small and large scales. The success of python was due to the fact that it is designed to create readability syntax which allows to express concepts in fewer lines of

code (Kuhlman 2009) supporting multiple models of data computation (programming paradigm) such as object-oriented (computer programs designed to make interaction among different variables, data structure and functions), imperative (computer programs designed to make programming instructions that have to be performed by the computer. It is focused on describing how a program operates) and functional programming (computer programs designed to build mathematical function of computer programs) (Kuhlman 2009).

Data scientists have recently become interested in Python's programming language to analyse both small and large datasets. This interest towards Python is rising because it is open search, it has an awesome online community, it is easy to learn and it is readable. Python community provides numerous libraries focused on data analysis. The most important libraries useful to this aim are:

- *NumPy* is a fundamental package for scientific computing. It is useful to analyse a powerful multi-dimension array and to create sophisticated functions such as linear algebra, Fourier transform and random number capabilities. In addition to scientific uses, *NumPy* can be an efficient package to analyse multi-dimensional generic data.
- *SciPy* is a collection of packages addressing a number of different standard problem domains in scientific computing. In particular, it provides many user-friendly and efficient models for numerical integration, optimization, interpolation, image processing, and other tasks in science and engineering fields.
- *Pandas* is a package that provides high-performance, easy-to-use data structures and data analysis tools. In particular, it provides sophisticated indexing functionality to make it easy to reshape, slice and dice, perform aggregations, and select subsets of data. Moreover, *Pandas* is focused on the creation of models such as linear and panel regressions.
- *Scikit-learn* is a package for data mining and data analysis. In particular, it is focused on various classifications, regression and clustering algorithms such as support vector

machines, random forests classifier, gradient boosting, *k*-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

- *Matplotlib* is the most used package for 2-dimension graphics. It is useful with data visualization providing high-quality figures in many formats and providing a comfortable interactive environment for plotting and exploring data. It is possible to easily generate scatter plots, histograms, power spectra, bar charts, etc. using just a few lines of code. Every aspects of a figure can be customized using *Matplotlib*. In particular, the plots created by *Matplotlib* are interactive, thus allowing the user to zoom on a section of the plot and pan around the plot by using the toolbar in the plot window.

Due to the characteristics listed before, Python could be a useful programming language for data analysis in several research fields providing accurate inference and prediction about the datasets in faster and readable ways by using only a few Python libraries. These same characteristics make Python well suited to researchers and experts of different fields without a computational background. In addition to these reasons, it is worth noticing that biomedical scientists are increasingly becoming computationally oriented. In particular, this is due to the fact that the evolution of measurement instruments such as electronic medical records and digital public health registries require computational tasks to store and analyse data (Chapman and Irwin 2015).

1.2. BIG DATA ANALYSIS IN SPORT SCIENCES

Exercise scientists and sports experts have always been interested in reading statistics about players, teams and performances in order to improve competitions. The vast usage of statistics in sports makes Big Data the perfect methodology to study sport performance because large amounts of data are recordable through multiple channels. Therefore, with the advent of the ability to analyse Big Data, the data extracted from competitions gain more and more importance in the enhancement of the performance thus changing the face of sport analysis. What is more, sports scientists are starting to apply this successful approach in sports such as football (Tenga et al. 2010; Gréhaigne, Bouthier, and David 1997), squash (Y et al. 1996), basketball (Ben Abdelkrim et al. 2010; Csataljay et al. 2009), hockey (Spencer et al. 2004), rugby (Deutsch, Kearney, and Rehrer 2007), volleyball (Hughes and Daniel 2003) and Australian football (Dawson et al. 2004).

Sport teams and analytic companies invest several millions per year to buy cameras and sensors able to collect more and more data of the performance of their players. One of the most significant tasks sport scientists and coaches want to accomplish by using these data, is to understand which are the parts of the team sport competition's pitch which should be dominated by a specific player thus being able to determine the team formation more likely to obtain the best performance during the match. In order to do that, the analysis of game-play data such as movement and position of each player in relation to the teammate, ball and opponent is fundamental (Gudmundsson and Wolle 2010). Moreover, despite the training loads are one of the most important topics in training periodization (V. Issurin 2008), Big Data analysis has not become popular in training analysis yet even though large amounts of data have been recorded in each training. Having this huge quantity of data, sport scientists could have a complete overview of the training, which would allow them to detect the best loads able to provide enhancement in sport performance. Furthermore, another aspect that has a crucial importance in sport sciences is the injuries prevention because players' injuries have a negative impact on both financial and tactical factors of teams (Korkmaz et al. 2014).

Thus, the big challenge that Sport Science has chosen to face is to use large datasets to gain a competitive advantage in real time during the competition and to help coaches and athletic trainers to plan trainings programs able to the performance and reduce the risk of injuries. The aim of this chapter is to provide a brief review of what can be found in literature about match and training analyses by using Big Data focusing on the football game. In addition, the second aim of this chapter is to provide the rationale for which Big Data should be useful to analyse and predict sport performance, training loads and injury risk.

1.2.1. PERFORMANCE ANALYSIS IN TEAM SPORTS COMPETITIONS

The ability to process large amounts of data recorded by motion tracking during team sport competitions has markedly improved in the last 20 years (Barris and Button 2008). As a matter of fact, instruments such as global positioning system (GPS), high-speed videos and accelerometers have become popular in the measurement and the assessment of human physical activity. Thus, the quantitative analysis of players and team activities is increasingly becoming an important aspect of the coaching process (Eom and Schutz 1992). The detection of tactical models in offensive and defensive situations can be helpful for coaches and researchers to identify match regularities and random events during the competition (Garganta 2009). The information obtained from match analyses is obliviously crucial to achieve individual and team efficacy during the match and it constitutes a basic criterion for training program as well (Garganta 2009). The tactical modelling in team sports was initially investigated by several researchers (McGarry et al. 2002; Lames and McGarry 2007; Schöllhorn 2003). They used competition modelling to detect patterns during the match able to predict the success or failure of the team. In particular, such data could be used to establish performance profiles in successful competitions to be compared to the unsuccessful ones (Barris and Button 2008).

Sports analytics has already consolidated in team sports like basketball and baseball, where coaches and experts exploit data on players' and team's movements and events to understand tactics, with the aim to define efficient winning strategies. For example, in National Basketball Association league the player's efficiency rating introduced by Hollinger (Hollinger 2005) is a widely used measure to assess players' performance by combining the large amounts of data gathered during each match (pass completed, shots achieved, etc.). Moreover, in baseball, Smith et al. (2007) proposed a Bayesian classifier able to predict baseball awards in the US Major League Baseball, with an accuracy of 80% (Smith, Lipscomb, and Simkins 2007).

In the last decade, a rapid improvement of technologies and machine learning processes applied on sport performance have been observed. From this quick evolution a reasonable question arises: is it possible that computers might one day replace a coach? In some way, they already do. Coaches take more and more into consideration the results of competition analysis in order to improve the performance of their teams and to reduce the risk of injuries of their players. However, the not yet perfect accuracy detected in the performance prediction suggests that the team coach will always be indispensable. The motivational aspect provided by coaches (Charbonneau, Barling, and Kelloway 2001) is difficult to be replaced by the computer. Nevertheless, this might change in the future.

1.2.1.1. MATCH ANALYSIS IN FOOTBALL

Due to the great complexity governing the football competition, the statistical analysis of the most popular game in the world has been fascinating sport scientists, coaches and field experts for a long time. Sixty years ago, Charles Reep started to record football performance data by hand with the aim of investigating the patterns existing in several aspects of the football competition (e.g. accurate passes, ball lost). He has been the first to demonstrate that several aspects of football could be described by mathematical functions (Reep and Benjamin 1968; Reep, Pollard, and Benjamin 1971).

Specifically, he affirmed that the probability of losing the ball during the match increases with the number of consecutive passes. These results gave him the possibility to define the famous long ball theory. In 1973, Valeriy Lobanovskyi provided an early example of an analytic approach to football competition when he became coach for Dynamo Kyiv. He defined tactics and schemes for his football team by using computer and statistical approach (Kilpatrick, 2011), bringing both Dynamo Kyiv and Soviet Union national team to the highest level of European football.

From these preliminary studies, the technology rapidly grew permitting to easily record a large number of football performance features during a match. Opta e Prozone are the two most specialized companies collecting football data from every match, generally for commercial purpose. On the one hand, Opta provides data describing all events occurred during the competition with details about the position of the event inside the football pitch, the type of the event and the player(s) involved. On the other, Prozone provides spatio-temporal data describing all players' movements in the field during the competition. Based on these two types of data, a vast amount of researches has been produced in the last few years covering different aspects of football performance.

Several data mining processes were used by different authors to analyse football competitions. Borrie at al. (Borrie, Jonsson, and Magnusson 2002) suggested using analysis of time-based event records and real-time behaviour records on football performance known as T-pattern detection in order to find similarity in passes sequence during the competition. Otherwise, Gudmunsson and Wolle (Gudmundsson and Wolle 2010) tried to detect which players' movements are the most repeated. To these aims, they used Frechet distance (i.e. measure of similarity between curves which takes into account the location and ordering of the points along the curves) between trajectories to detect similarity in players' movement. Moreover, mine frequent motifs from team passing sequences was investigated by Gyarmati et al. (Gyarmati, Kwak, and Rodriguez 2014) in order to classify the team playing style. They found that FC Barcelona used a unique passing strategy and playing style compared to the other team, thus identifying and describing the famous "Tiki-taka" strategy. In

addition, Voronoi spatial classification (Berg et al. 2000) is a geometrical model widely used to investigate football competitions as well. Taki and Hasegawa were among the first to use the Voronoi diagram in order to divide the football pitch into cells owned by the football players (Taki and Hasegawa 2000). This model was subsequently improved by Fujimura and Sugihara who defined a more efficient approximation model for region computations (Fujimura and Sugihara 2005). In this study, the area on the football field that a player can reach before any other was considered the player's dominate region defined by the machine learning process. As a matter of fact, the dominant regions of the players were computed on the basis of their current positions, speeds and constant accelerating abilities. By using the Voronoi diagram, researchers and field experts detected players' passing and movement options (Gudmundsson and Wolle 2010) in order to obtain enhancement in football competition.

Another approach to the problem of data analysis of football competition is based on network theory. Players can be easily identified as nodes of a network where a pass between two players represents a link between the respective nodes. Based on the network analysis, it was found that network metrics can be a powerful tool to assess connections between players and the strength of players links (Clemente et al. 2015). Therefore, the results of this analysis could help coaches and athletic trainers to support decisions during the competition. As a matter of fact, Cintia et al. (Cintia, Rinzivillo, and Pappalardo 2015) set a pass-based performance indicator by using the team's passing network in four national championships. They replaced the real competition results by simulated outcomes based on the pass-based indicator obtained from each team. In particular, it was set that a team wins when his pass-based indicator value is higher compared to the opponent. Moreover, two teams draw when their indicators are close. According to researchers, the more the season goes by the more the association between the real ranking and the simulated one increases.

1.2.2. ANALYSIS OF TEAM SPORTS TRAININGS

The sport performance has arisen from well-defined training (Gamble 2006) and not only by match features. In point of fact, the training (single session or short-term cycle) causes fatigue in athletes' work capability which promotes a physiological adaptation after a well-defined recovery time (V. B. Issurin 2010). In accordance with Selye's General Adaptation Syndrome (Selye 1946) training stimuli lower than optimal ones are not enough to produce adaptation. On the contrary, training stimuli higher than optimal ones may lead both to overtraining and to an increased risk of injury (Tim J. Gabbett and Ullah 2012). Hence, coaches and athletic trainers have to adjust and balance training stimuli, competition stimuli and recovery time properly in order to obtain the best performance during the competition (Tim J. Gabbett and Domrow 2007).

Until now, the prescription of the specific training loads at a specific time for the enhancement in sport performance has been largely instinctive, resulting from years of personal experience (Borresen and Lambert 2012). It is generally believed that the higher are the training loads the higher is the improvement in sport performance. However, as explained before, even if this theory is widely accepted, several studies demonstrated that a random increase in training volume, intensity and frequency may increase the probability of injury and overtraining (Urhausen and Kindermann 2002; Williams and Eston 1989; Halson and Jeukendrup 2004). Since sport is a very complex field to analyse because of the several variables coming into play (e.g. physiological, psychological, environmental), there are only a few studies about machine learning process usage able to define specific training loads (Rygula 2005; Novatchkov and Baca 2013; Bartlett et al. 2016). Thus, the role of Big Data analysis in this process is progressively becoming essential.

1.2.2.1. TRAINING PERIODIZATION IN TEAM SPORTS

It is well known that training periodization is crucial to optimize training responses. The training loads variation within the periodization is a key function in successful training prescriptions (Fleck 1999). Training periodization permits to plan a systematic variation of the training loads to obtain the physiological adaptation required from a specific sport (Brown and Greenwood 2005; Plisk and Stone 2003). However, the application of training periodization schedule in team sports is very complex due to the intra-players' differences, the variety of training goals and the extended competition period characterized by frequent competitions. The latter point does not permit to use the traditional periodization model – a division of the entire seasonal program into smaller periods and training units aimed to reach the best work capability two/three times per season – for planned training variations (Plisk and Stone 2003; Gamble 2006; Daniel Baker 1998). As a matter of fact, it is characterized by a gradual progressive increase in intensity (i.e. linear model) that is not useful for multi-peak performance. On the contrary, what represents a useful method to train top-level team sports, which involves frequent competitions during a competitive season, is the so-called block periodization model (Issurin, 2010). It is characterized by drastic variations of intensity within the weekly and daily program (i.e. non-linear model or undulating periodization), and it is hence useful to this aim (V. B. Issurin 2010). The employment of block periodization is essential in team sports such as football, rugby, basketball, etc. because they are characterized by 20-35 weeks with 1 or 2 competitions per week (Gamble 2006). Indeed, it was found that the application of training design following traditional planning precepts in team sports periodization leads to reductions in body mass, maximal strength, maximal anaerobic power and maximal speed (Astorino et al. 2004; Kraemer et al. 2004; Häkkinen 1993). The use of the traditional model is still realistic for junior and low-level athletes only, whose competition phases are relatively short and can be considered similar to those of individual sports (V. B. Issurin 2010).

An example of annual training periodization in terms of duration, dominant training targets, and load level is shown in Figure 1.4. This is just an example of one annual training cycle because of the impossibility to standardize a universal schedule model due to a great variability among team sport calendars. For team sports' annual periodization, it was suggested to perform the traditional training model in the off-season and pre-season phases, and the undulating periodization model in the in-season one, which is characterized by frequent competitions (Gamble 2006). The traditional model performed in the off-season and pre-season phases facilitates the acquisition of general and specific sport abilities, which could be maintained in the in-season phase through the undulating periodization model.

PHASES	OFF-SEASON	PRE-SEASON	IN-SEASON	POST-SEASON BREACK
TARGETS	Active recovery Metabolic conditioning General strenght	Tecnique perfection Sport-specific strenght and power Maximal speed Sport-specific endurance	Metabolic conditioning Sport-specific endurance Techno-tatical skills	Active recovery Psychological recovery
LOADS	LOW TO MEDIUM	MEDIUM TO HIGH	HIGH TO VERY HIGH	LOW
DURATION	3-4 WEEKS	6-20 WEEKS	15-35 WEEKS	1-4 WEEKS

Figure 1.4. Schematic representation of an annual training periodization in team sports (V. B. Issurin 2010).

The importance of a correct training periodization in team sports cannot be underestimate. The long series of stressful competitions often lead to harmful consequences such as pronounced catabolic responses (Kraemer et al. 2004; Carli et al. 1982), musculoskeletal disorders and a high risk

of injuries (Gamble 2006). Thus, a well-defined annual training periodization that avoids discordant physiological responses could facilitate the maintenance of sport-specific abilities and prevent a decrease of relevant physiological capabilities (Newton et al. 2006; D. Baker 2001).

1.2.2.2. TRAINING ANALYSIS IN FOOTBALL

The physical and technical demands in football competitions have noticeably increased over the past decade (Barnes et al. 2014). Thus, football coaches and athletic trainers need to implement training loads in order to increase the probability of success during competitions. To date, the major challenge in football training is to provide accurate training loads boundaries of what the players can achieve without exceeding what their bodies can tolerate (Piggott 2008). Therefore, monitoring and understanding football training loads has become essential to assess the optimal ones able to increase positive training adaptations and reduce the risk of injuries (Rogalski et al. 2013).

The quantification of external loads in multi-direction sport game such as football is extremely difficult to achieve. Thus, traditional approaches focused only upon the duration and frequency of the training stimuli are not adequate to quantify specific training loads (Brink et al. 2010). So far, the evolution of global positioning systems (GPS) has provided the opportunity to record valid and reliable estimates of distance covered by each player during different physical activities which require diverse physiological demands (Waldron et al. 2011; Portas et al. 2010; Varley, Fairweather, and Aughey 2012). Indeed, the assessment of external loads by using a GPS has focused on the evaluation of the distance covered at specific velocities and the time necessary to do it, with a particular attention upon the volume of high-speed activities, which are important in football competition (Aughey 2011; Di Salvo et al. 2009; Iaia, Rampinini, and Bangsbo 2009). However, these studies did not take into account the accelerations and decelerations derived from GPS to describe the external loads. The employment of these variables in addition to the distance covered and the time spent at specific

velocities could increase the precision of the football energy cost demand estimation. As a matter of fact, it was found that the traditional approach underestimates the total energy cost in football activity (Osgnach et al. 2010; Cavagna, Komarek, and Mazzoleni 1971; di Prampero et al. 2005). In light of these results, a new approach was recently introduced: combining acceleration and deceleration with the traditional estimation of the energy cost during specific running speed to assess the overall energy cost of a specific activity (di Prampero et al. 2005). This new approach permitted to Osgnach et al. (Osgnach et al. 2010) to detect that the energy cost of the high-intensity football activity was underestimated of 2-3 times when assessed using the traditional approach.

Therefore, by an overall point of view, the assessment of external load is required for team sports such as football. The large amount of features easily recorded by GPS technology should be used to precisely define the training loads required to obtain enhancement in football competitions and to reduce the risk of injuries of football players'. The machine learning process could be useful to this aim. However, to the best of our knowledge no study has been performed to quantify training loads by using Big Data analysis in order to have an overall point of view on football training loads periodization.

1.2.3. INJURY PREDICTION

In addition to performance and training analysis, the injury prediction is one of the major topics that in the last decade have started to be investigated. This increasing interest towards injury prevention and prediction derived from the fact that injuries could impair team's performance caused by the forced absence of essential players to crucial matches.

In the last decade, it was demonstrated that any injuries which could potentially be considered 'training load-related' are commonly viewed as 'preventable' (Tim J. Gabbett 2016). Thus, the common aim of sport scientists, doctors and physiotherapists has become to keep players free from

injury. To this aim, hypothetical relationships between training, injury, fitness and performance (Figure 1.5) (Orchard 2012) verified by several investigation have been proposed (Tim J Gabbett 2004; Tim J. Gabbett and Jenkins 2011; Tim J. Gabbett and Ullah 2012; Rogalski et al. 2013). In particular, it was found that the higher is the training loads the higher is the risk to have an injury. In addition, Gabbet (Tim J Gabbett 2004) also showed that reductions in training loads lead to a decrease of injury risk without affecting sport-related physical skills. In accordance with these encouraging results, sport scientists and researchers have become interested in defining an accurate estimation of training loads to obtain the maximum enhancement of performance keeping an acceptable injury rates (Tim J. Gabbett and Ullah 2012).

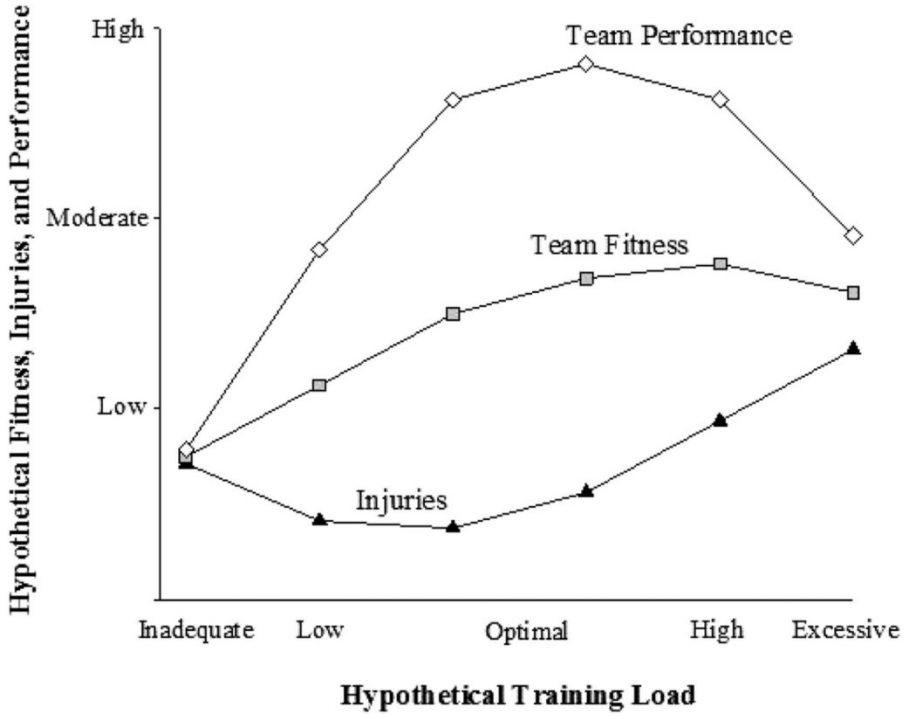

Figure 1.5. Hypothetical relationship between training loads, fitness, injuries and performance (Orchard 2012).

Despite these previous reassuring results based on players' loads, one of the first investigations that tries to predict injuries using machine learning process was conducted by Talukder et al. (Talukder et al. 2016). They created a model able to predict with accuracy when an NBA player

is most likely to get hurt. They used a random forest classifier as machine learning process to predict the injury risk from the data recorded during the match performed in the two previous seasons (e.g. minutes, competitions played and speed, and distance covered per player). They stated that a 7-days predictive window is able to accurately forecast NBA players' injury. In addition, they stated that the most important features that are able to discriminate injury risk are the average speed during the competitions, the number of competitions played to date in the season, the average distance covered during competitions, the number of minutes played to date in the season and the average field goals attempted. This machine learning process enables coaches and athletic trainers to identify the best time for a team to rest players and consequently reduce the risk of long injuries.

Hence, in accordance with the results found in previous studies, it is licit to suppose that the analysis of a large size of variables (i.e. players' load during both training and match and the data derived from the performance) could improve the ability to accurately predict injuries in team sports. Future investigations should be scheduled on the basis of these speculations.

PART 2: EXPERIMENTAL STUDIES

2.1. AIM OF THE THESIS

The importance of machine learning process to analyse big data in football trainings is the major topic of this thesis. The aim of this work is to provide an overall point of view about the in-season football trainings through Big Data analysis. In particular, the presence of a weekly pattern through in-season trainings (i.e. short term cycle), the importance of the features to describe football trainings and the probability of injuries prediction have been investigated.

To this aim, eighty in-season trainings of twenty-six Italian elite football players were recorded by using GPS technology. In accordance with the introduction of this thesis, it was possible to hypothesize that GPS features are able to provide an accurate description of the in-season training pattern. From this overall point of view, it was possible to detect the most important features that should be taken into account to provide enhancement in football competitions. Moreover, thanks to this investigation it is licit to suppose that the players who performed training loads different from the optimal ones are also those who were subjected to a higher risk of injury. Thus, in light of the above assertions, this study aims to provide a valid method to define accurate training loads in in-season periodization. Individuals' load discrepancy to the required one detected by this method may be useful to obtain enhancement in performance and to detect injury risks.

The *Experimental Studies* part consists of two main sections. The first one tries to detect a pattern through the in-season football training. The undulated model permits to assert that the in-season football training was periodized in a short-term cycle in order to obtain the best performance several times during the football season. On the basis of these results, the second section provides the development of an algorithm able to predict the injury risk.

2.2. STUDY 1: CHARACTERIZATION OF IN-SEASON ELITE FOOTBALL TRAININGS BY GPS FEATURES: THE IDENTITY CARD OF A SHORT-TERM FOOTBALL TRAINING CYCLE.

Rossi A., Perri E., Trecroci A., Savino M., Alberti G., Iaia FM

School of Exercise Sciences, Department of Biomedical Sciences for Health, Università degli Studi di Milano

2.2.1. ABSTRACT

Football training periodization is widely recognized as crucial to obtain the best performance throughout the matches and to reduce the risk of injuries. Thus, the aim of this study is to detect the in-season short-term training cycles in an Italian elite football team. 80 trainings of 26 elite football players were monitored during 23 in-season weeks by a global position system (GPS). Machine learning process and autocorrelation analyses were performed in order to detect patterns within the in-season football trainings. Extra tree random forest classifier (ETRFC) was used to create a supervised machine learning process able to describe the football trainings cycle. This analytical model allows us to produce reliable decisions and results learning from historical relationships and trends in the data. In addition, the autocorrelation analysis allows us to detect similarity of observation among the data. On the basis of these analysis, it was found that the in-season football trainings are characterized by a series of short-term cycles. This kind of periodization follows a sinusoidal model because the short-term cycle detected in the in-season trainings is composed of two parts with different training loads. In particular, in the days long before the match football players perform higher training loads than in the close ones. To enhance performance and reduce the risk of injuries, it would be essential to provide correct stimuli in each short-term cycle per day. Thus, developing a valid method able to define the correct training loads in each training day may be central for coaches and athletic trainers to periodize correctly the football trainings.

Keywords — Data analysis; Short-term cycle; Training loads

2.2.2. INTRODUCTION

In the last decade, sport and data scientists have been analysing big datasets of individuals' and teams' performance in order to verify existing sport theories and develop new ones. Despite football is one of the most popular sports in the world, Big Data analysis in football has become popular only recently. Sixty years ago, Charles Reep recorded football performance data by hand with the aim of investigating patterns that exist in several aspects of the football game (e.g. accurate passes, lost ball) (Reep and Benjamin 1968; Reep, Pollard, and Benjamin 1971). From these preliminary studies, the technology rapidly grew permitting to easily record a large number of football performance features. This large amount of data provides new opportunity of collaborations between data and sport sciences to maximise the machine learning potentials for predicting football match performance. Since Reep's studies, more and more investigations have been published in sport science journals in order to detect patterns able to describe the football match using wide datasets (Bialkowski et al. 2014; Gudmundsson and Wolle 2010; Tamura and Masuda 2015; Taki and Hasegawa 2000). The football performance has arisen from well-defined training loads (Gamble 2006) and not only by match features. Thus, football training investigation is equally important as the match analysis in order to predict football performance. This study wishes to be the pioneer of the Big Data analysis application on football training.

The quantification of training loads is a crucial aspect of football. The combination of factors that can be manipulated for planning a training session are various. The training (single session or short-term cycle) causes fatigue in athletes' work capability, which promotes a physiological adaptation after a well-defined recovery time (V. B. Issurin 2010). Physiological adaptations are due to the combination between external loads and recovery (Gamble 2006). Thus, if football players performed a particular training in a specific time before the match, they could be more likely to attain the best performance while playing.

The traditional training periodization – a division of the entire seasonal program into smaller periods and training units aimed to reach the best work capability two/three times per season – is not suitable for team sports as football. In fact, football players have to reach excellent results throughout the football season once or twice a week. The plateau of the performance needed in the football season is obviously in contradiction with traditional periodization (Bialkowski et al. 2014) and this is why block periodization represents a useful method to train top-level sports (Gudmundsson and Wolle 2010). To the best of our knowledge, no study investigated which is the pattern useful to training elite football players in order to reach the plateau of performance during the in-season period. Our study wants fix the gap found in literature in order to provide the training pattern used by football trainers to train their athletes.

The global position system (GPS) provides several indices of external training loads able to describe the football training (Cummins et al. 2013). However, even though there is a great amount of indicators it is very difficult for coaches and athletic trainers to periodize the trainings because of the multidimensional characteristics of football performance. The development of a valid method for assessing training load is essential in football since excessive training responses may lead to training maladaptation and injury (Ehrmann et al. 2016). Previous studies have already detected a short-term training loads periodization but they used only a few GPS features (Akenhead, Harley, and Tweddle 2016; Malone et al. 2015; Scott et al. 2013). Thus, the major novelty of this study is the training analysis on the multidimensional factors that could be able to provide an overall picture of the short-term football training cycle during the in-season period.

Thus, the aim of this study is to describe: i) an in-season short-term football training cycle; ii) the importance of the features provided by the GPS; iii) the overall periodization of the training sessions. For these aims, eighty in-season trainings of twenty-six Italian elite football players were recorded by using GPS. In accordance with the above cited literature, it is licit to suppose that the twelve features extracted from the GPS could be able to characterize a pattern through in-season football trainings. Should this supposition be verified, it would be possible to precisely define training

loads in the in-season periodization and the most important features able to describe football trainings. As a matter of fact, a sinusoidal model was found in this study. Indeed, a supervised classifier and the autocorrelation analysis detected a pattern through the in-season block. Thus, in light of the above assertion, this study wants to provide a valid method to define accurate loads in in-season trainings. The fundamental issue in this paper is that the football match (i.e. the moment when the players perform the higher load) is not taken into account during the analysis. However, the GPS is not possible to dress during the match until now. In the 2016-2017 season it will be allowed to use the GPS also during the match. Thus, future investigations are scheduled in order to define the best training loads in accordance with the match loads in order to obtain enhancement in performance. Moreover, a more precise definition of training loads could be useful to predict injury risks by the individuals' load discrepancy to the required one.

To investigate our aims, the paper is organized as follows. In session 2 we describe the subjects, the instrument and the procedure used to record the football trainings, the features extracted from GPS, and the analyses used to detect patterns inside in-season football trainings. In session 3 we objectively describe the results obtained by our analyses. In particular, in Figure 2.1 we provide an identity card of the short-term cycle that could be useful for football coaches and trainers to have a clear idea of the loads performed during the short-term cycle. In conclusion, in session 4 and 5 we provide comments and explanations about our results suggesting the utility to use data mining process to schedule the right training loads.

2.2.3. MATERIALS AND METHODS

Subjects: Twenty-six professional football players competing in the Italian Serie B (age = 26 ± 4 yrs; height = 179 ± 5 cm; body mass = 78 ± 8 kg) took part in the study during the 2013-2014 in-season competition period (23 weeks). Six central backs, three fullbacks, seven midfielders, eight wingers and two forwards were recruited. Goalkeepers were not included in the study.

Procedure: Players' physical activity during each training session was monitored using a portable non-differential 10 Hz global position system (GPS) integrated with 100 Hz 3-D accelerometer, a 3-D gyroscope, a 3-D digital compass (STATSports Viper, Northern Ireland). The interaction among these devices is useful to create several GPS features able to describe the sport performance (Duncan, Badland, and Mummery 2009). Each player wore a tight vest, and the receiver was placed between their scapulae. All devices were always activated 15-min before the data collection both to allow acquisition of satellite signals in accordance with manufacturer's instructions and to avoid that inter-unit error players wore the same GPS device for each training session. After recording, data were downloaded to a computer and twelve features were automatically computed by using the software package Viper Version 2.1 (STATSports 2014). If players did not complete the training session or the GPS have not correctly recorded the players' position, these data was not taken into account during the analysis (30% of the individuals' training sessions were deleted from the dataset). The Sport Club gave the permission to use the data for the purpose of research according to privacy policy.

Data description: A total of 80 team training sessions and 2080 individual trainings were recorded. The short-term training days analysed were:

- Two days after the previous match (MD+2);
- Four days before the following match (MD-4);
- Three days before the following match (MD-3);
- Two days before the following match (MD-2);
- The day before the following match (MD-1).

The Training Load features recorded were:

(1) Total distance: distance (m) covered by each player;

(2) High Speed Running Distance: distance (m) covered by a player when his speed is above 19.8 Km·h-1 (5.5 $m \cdot s^{-1}$);

(3) Metabolic Distance: distance (m) covered at metabolic power (di Prampero et al. 2005) above 20 $W \cdot Kg^{-1}$;

(4) High Metabolic Load Distance: distance (m) covered by a player when his metabolic power is above 25.5 $W \cdot Kg^{-1}$. This value indicates when a player is running at a constant speed of 19.8 $Km \cdot h^{-1}$ (5.5 $m \cdot s^{-1}$) on grass or when he is performing significant acceleration or deceleration activity (> 2 m·s-2), and it is identified as indicator of high-intensity distance covered (Gaudino et al. 2013) [16];

(5) Explosive Distance: distance (m) covered by a player when his metabolic power is above 25.5 $W \cdot Kg^{-1}$ and the speed is below 19.8 $Km \cdot h^{-1}$;

(6) High Metabolic Load Distance per Minute: average distance covered by a player per minute when his metabolic power is above 25.5 $W \cdot Kg^{-1}$;

(7-8) Accelerations > 2 $m \cdot s^{-2}$ and accelerations > 3 $m \cdot s^{-2}$: acceleration activity measured on the basis of the change in GPS speed data is defined as a change in speed for a minimum period of 0.5 s with a maximum acceleration in the period of at least 0.5 $m \cdot s^{-2}$. The acceleration is considered ended when the player stops accelerating. The classification of acceleration by zone is based on the maximum acceleration reached in the acceleration period;

(9-10) Decelerations > 2 $m \cdot s^{-2}$ and decelerations > 3 $m \cdot s^{-2}$: deceleration activity measured on the basis of the change in GPS speed data is defined as a change in speed for a minimum period of 0.5 s with a maximum deceleration in the period of at least 0.5 $m \cdot s^{-2}$. The deceleration is considered ended when the player stops decelerating. The classification of

deceleration by zone is based on the maximum deceleration reached in the deceleration period;

(11) Dynamic stress load: total of the weighted impacts that the player endures during the training;

(12) Fatigue Index: ratio between DSL and Speed Intensity (the latter being the sum of the product between the time and the index of a specific range of velocity). It is a measure of the stress during the training on football players.

The dataset was composed as <player, short-term training day, match number, f1, f2, f3 … fn> where f indicates the GPS features.

Statistical Analysis: The dataset was normalized using min-max standard scaler. The reason why this scaling was used is to improve the robustness of the features (very small standard deviation) and get the values distribution close to a Gaussian shape. Data were standardized for each subject in order to reduce intra-subject variability. Moreover, clustering analysis required data normalization in order to standardize the distance through the features thus improving the classification accuracy.

Extra tree random forest classifier (ETRFC) was performed to create a machine learning process able to predict the short-term training cycle. We used ETRFC because it is a very robust model that fits a number of randomized decision trees on various sub-samples of the dataset improving the predictive accuracy and controlling the possible over-fitting as compared to the decision tree classifier and the random decision forest. Gini impurity was used as a criterion to develop the machine learning process because it is able to minimize misclassification. Cross validation on diverse test sets was performed to assess the machine learning process. Precision, recall and F1-score were computed to estimate the accuracy of the algorithm as well. In addition, feature importance analysis was computed according to ETRFC algorithm to detect the most important features able to delineate the elite football training.

In order to estimate the ability of ETRFC algorithm of characterizing football short-term training cycle, a dummy classifier (DC) algorithm was computed using a stratified strategy. This classifier is a baseline machine learning process created from the dataset by simple rules. This simple process was compared to ERTFC in order to detect which it the best process able to describe the dataset. Thus, if the DC algorithm is able to discriminate the short-term training cycle similarly or better than ETRFC algorithm, the latter cannot be considered able to discriminate short-term training cycle better than the baseline one.

In order to detect similarity on training time series (i.e. detection of repeating patterns inside time series) an autocorrelation analysis was performed. This analysis allows us to detect a periodicity in football training throughout the season.

2.2.4. RESULTS

The ETRFC algorithm is able to characterize the training inside the short-term cycle with an accuracy of 63.6%, while the stratify DC algorithm is able to designate 21.9% of the short-term cycle. The precision, recall and F1-score for each training day is provided in Table 2.1 and Table 2.2 for ETRFC and dummy classifier, respectively. In addition, figure 2.1 provides the identity card built using the class defined by ETRFC classifier. In the figure, data are presented as mean and coefficient of variation in each training day. The importance of the features to define football trainings by ETRFC algorithm is provided in table 2.3. The autocorrelation plot does not show significant autocorrelation for any features as showed in Figure 2.2. Table 2.3 shows that the distance covered above 20 W·kg-1 (metabolic distance zonal) and the acceleration above 2 m·s-2 are the two most important features to characterize the short-term training cycles using ETRFC process. The ellipse graph showed in figure 2.3 describes the two most important features variability through the in-season block of each short-term cycle day. The proximity of the ellipses permits us to aggregate the short-term cycle days in days long before (i.e. MD+2, MD-4 and MD-3) and close (i.e. MD-2 and MD-1) to the match.

Table 2.1. Precision, recall and F1-score of ETRFC algorithm used to predict short-term training.

Short-term day	Precision	Recall	F1-score	Support
MD+2	0.79	0.47	0.59	81
MD-4	0.53	0.38	0.44	171
MD-3	0.55	0.72	0.62	212
MD-2	0.72	0.56	0.63	218
MD-1	0.68	0.84	0.75	279
Average	0.64	0.64	0.63	
SD	0.11	0.19	0.11	

Table 2.2. Precision, recall and F1-score of DC algorithm.

Short-term day	Precision	Recall	F1-score	Support
MD+2	0.13	0.14	0.13	81
MD-4	0.19	0.18	0.18	171
MD-3	0.21	0.21	0.21	212
MD-2	0.21	0.22	0.21	218
MD-1	0.29	0.29	0.29	279
Average	0.21	0.21	0.20	
SD	0.06	0.06	0.06	

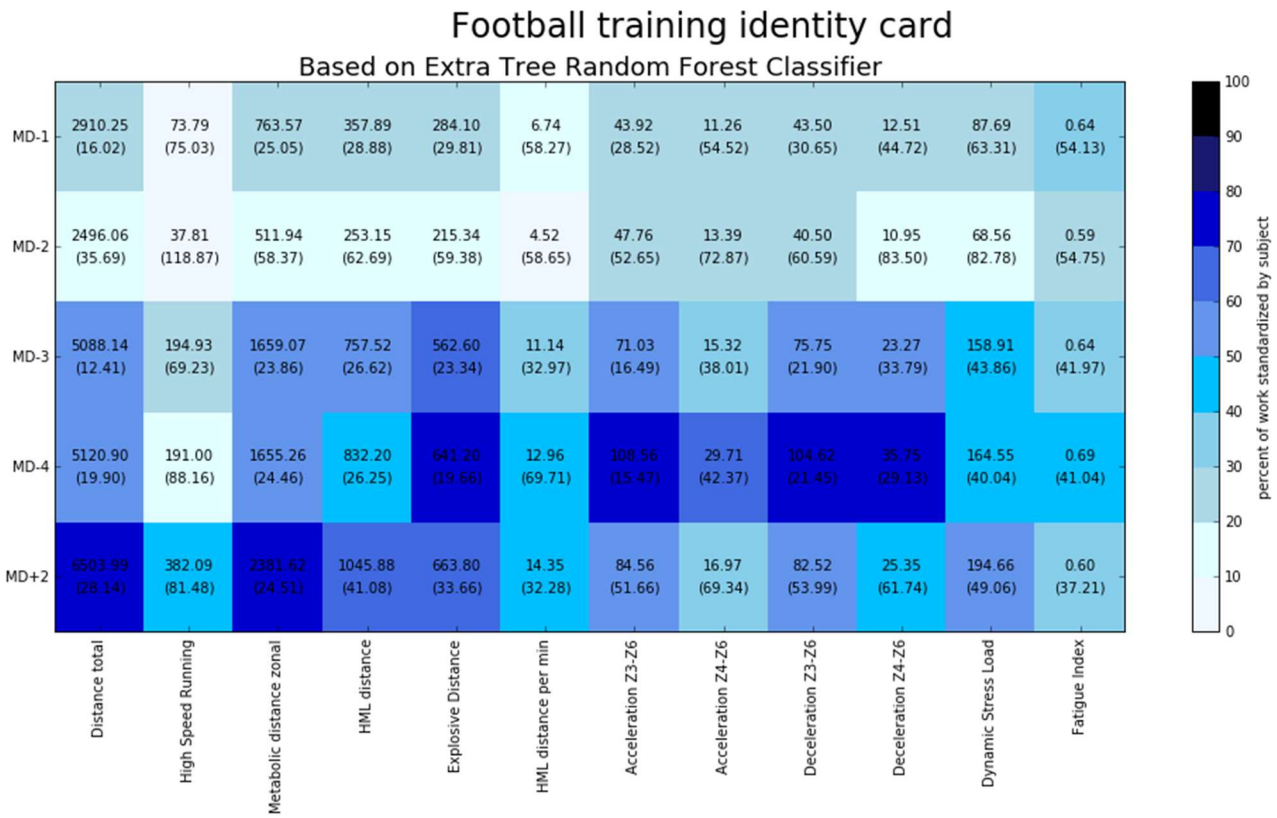

Figure 2.1. Football training identity card described on the basis of ETRFC as mean and coefficient of variation (the latter reported in brackets). The colour indicates to the mean of the percentage of maximal training performed by each player. On the x-axis are presented the GPS features and on the y-axis the short-term football training cycle.

Table 2.3. Percentage of feature importance based on ETRFC algorithm.

Features	Importance (%) -short-term cycle-	Importance (%) -aggregate training-
Metabolic distance zonal	11.22	13.06
Acceleration Z3-Z6	10.71	11.11
Distance total	10.40	11.53
Explosive distance	9.57	12.58
Deceleration Z3-Z6	9.53	10.49
HML distance	8.54	10.04
Deceleration Z4-Z6	7.85	7.68
Dynamic Stress Load	7.07	7.29
HML distance per min	6.77	5.34
High speed running	6.59	4.24
Acceleration Z4-Z6	6.50	3.75
Fatigue Index	5.25	2.89

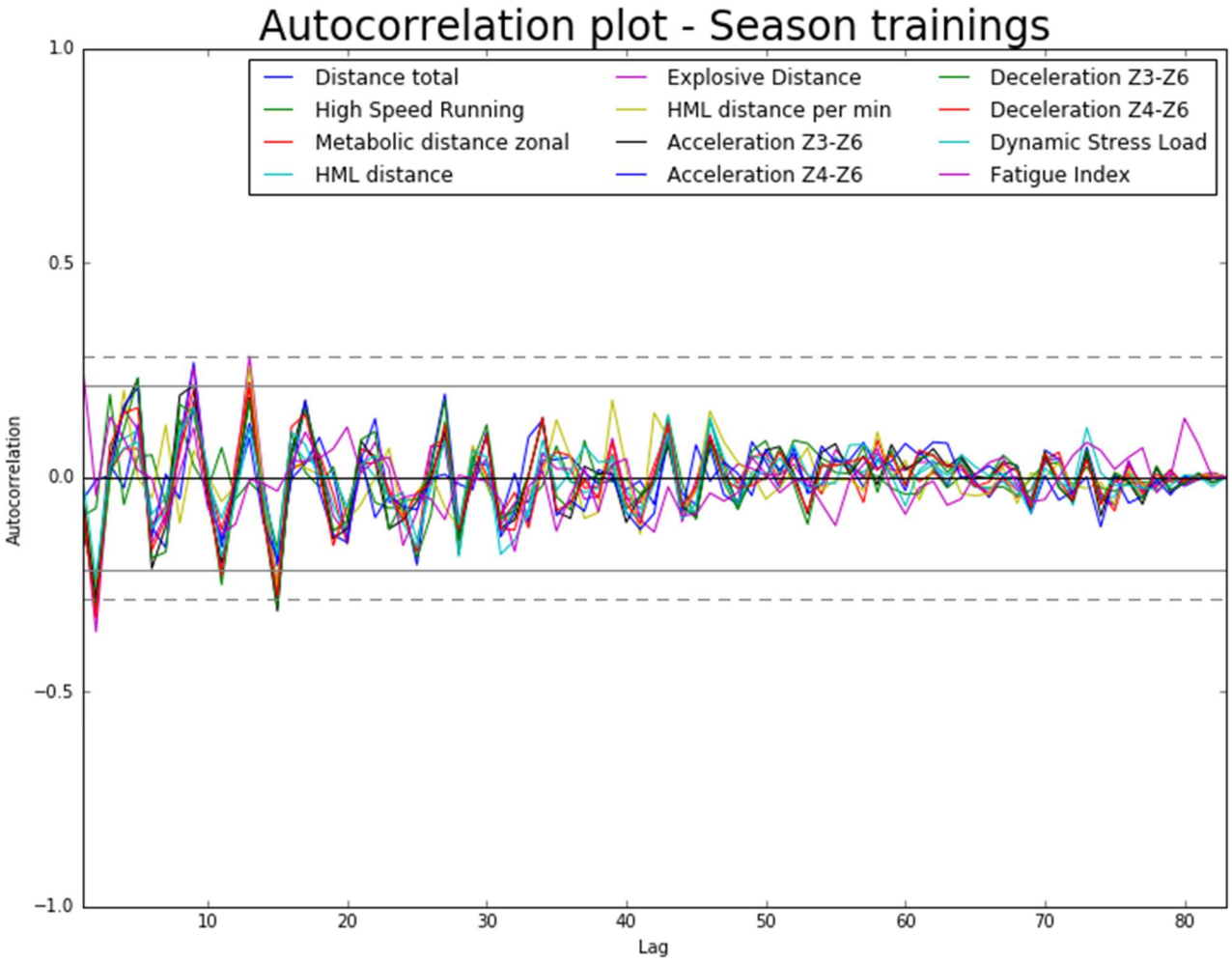

Figure 2.2. Autocorrelation plot of the features time series. The x-axis shows the trainings and the y-axis shows the value of autocorrelation. Different lines represent the GPS features.

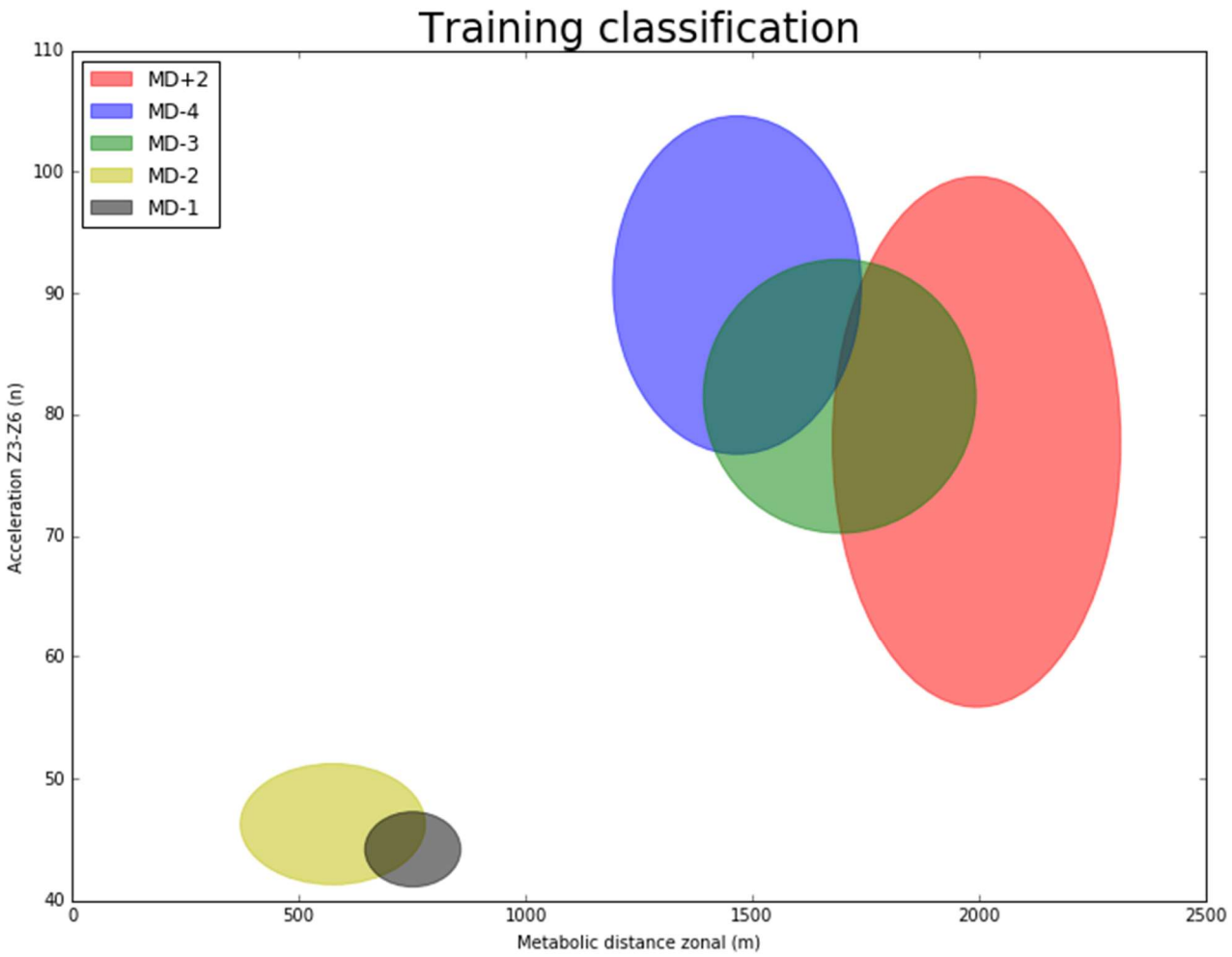

Figure 2.3. Ellipse plot describing the training day in the short-term cycle. The x-axis presents the most important features able to describe the short-term football training cycle and the y-axis indicates the second most important ones. Different colours represent the short-term football training cycle days.

The ETRFC algorithm is applied to aggregate training days in order to delineate the aggregate football trainings (i.e., training day long before or close to the match). This algorithm is able to classify the trainings in the two different classes with an accuracy of 90%. The DC algorithm correctly characterizes only 52.7% of the football training. Thus, the ETRFC is 37.3% better than the DC to discriminate the trainings during the in-season period. The precision, recall and F1-score for each training day is provided in Table 2.4 and Table 2.5 for ETRFC and dummy classifier, respectively. The autocorrelation plot performing on aggregate training days does not show statistically significant autocorrelation as showed in Figure 2.4. Table 2.3 shows the importance of the features for the aggregate training days.

Table 2.4. Precision, recall and F1-score of ETRFC algorithm used to predict short-term training.

Day label	Precision	Recall	F1-score	Support
Long before the match	0.90	0.89	0.90	464
Close to the match	0.90	0.91	0.90	497
Average	0.90	0.90	0.90	
SD	0.00	0.01	0.00	

Table 2.5. Precision, recall and F1-score of DC algorithm.

Day label	Precision	Recall	F1-score	Support
Long before the match	0.51	0.55	0.52	464
Close to the match	0.54	0.50	0.52	497
Average	0.53	0.53	0.52	
SD	0.02	0.04	0.00	

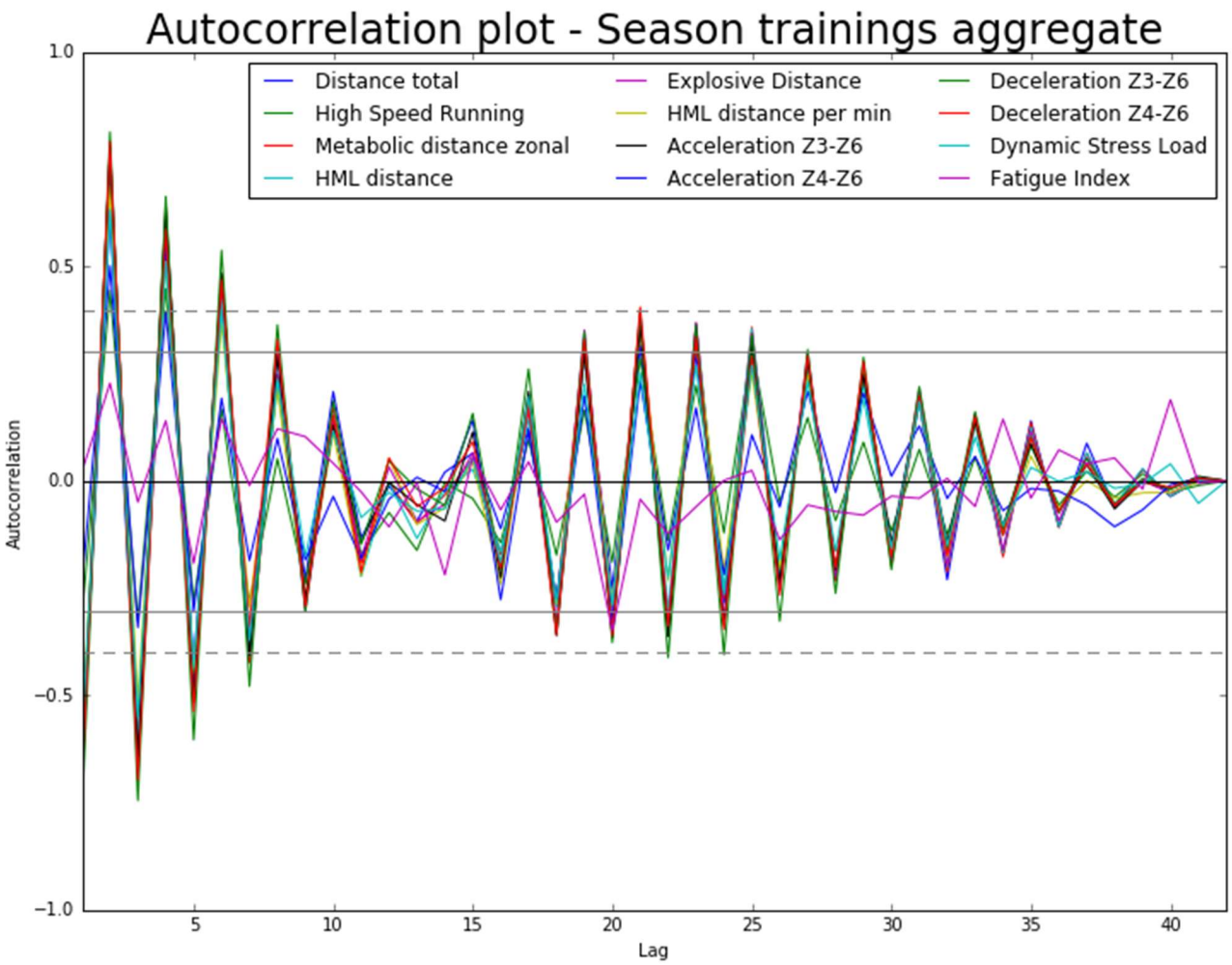

Figure 2.4. Autocorrelation plot of the features time series for aggregate training. The x-axis shows the training and the y-axis shows the value of autocorrelation. Different lines represent the GPS features

2.2.5. DISCUSSION

Progressive manipulation of the training loads during a short-term training cycle is a process of physiological adaptations where enhancements in performance are achieved (Manzi et al. 2010). Therefore, to define accurate training loads during a short-term training cycle is essential to prescribe an efficient training stimuli. As showed in the football training identity card (Figure 2.1), each day of the short-term training cycle is characterized by a well-defined features pattern. The ETRFC is useful to define accurate training loads in each training day due to the fact that this machine learning process is able to predict 41.7% better than the DC. Thus, coaches and athletic trainers should detect correct training loads by using the ETRFC to provide accurate stimuli to their players in each short-term training day thus reducing the risk of training maladaptation and injuries. Moreover, the high training loads variability in each short-term cycle day during the in-season football trainings – detected by using the coefficient of variation showed in Figure 2.1 – explained the inability of the supervised algorithm (36.4%) to perfectly predict the training day in the short-term cycle. In addition, this variability also explains the lack of autocorrelation (Figure 2.2) found through the in-season football trainings. According to our results, the short-term football training cycle is hard to be detected into the in-season training due to the intra-training variability.

Several studies investigated the training loads in the short-term football training cycle focusing only on specific features (Akenhead, Harley, and Tweddle 2016; Malone et al. 2015; Scott et al. 2013). Despite differences in absolute training loads were found for several features, the same short-term cycle structure was detected. Higher training loads were performed at the beginning of the short-term cycle and lower ones in the days prior to the match. This short-term structure appears to be a common strategy both to facilitate the decay of accumulated fatigue from high training loads performed at the beginning of the cycle and to promote readiness for the next performance (Akenhead, Harley, and Tweddle 2016). In accordance with the results found in previous studies, in figure 2.3 - ellipse plot of the short-term training day based on the more important GPS features (i.e. metabolic

distance zonal and the acceleration above 2 $m \cdot s^{-2}$) – it is noticeable that the football players performed the higher training loads in the days long before the match (i.e., MD+2, MD-4 and MD-3) than in those close to it. Thus, supervised machine learning process on two labels (i.e., training either close or not to the match) is needed in order to develop a more accurate and precise algorithm. We found that the ETRFC on aggregate training days was able to classify the trainings (90%) better than the machine learning process used to predict the non-aggregate ones (63.6%). Despite no significant autocorrelation was detected, the zig-zag shape detectable in figure 2.4 could support the fact that the data derived from a sinusoidal model. Indeed, the structure of the in-season football training follows this kind of model because it is composed by repeated short-term cycles characterized by two parts: the first one (i.e. long before the match) is constituted by high training loads and the second one (i.e. close to the match) by low ones. Thus, it is reasonable to obtain the zig-zag autocorrelation shape among trainings during the in-season block because the autocorrelation plot detects an alternate correlation (i.e. positive and negative) while the trainings go by.

Due to the fact that a pattern was found in short-term football training cycles, it is reasonable to assume that players performing training loads distant to the short-term cycle average do not achieve physiological adaptations and subsequently are unable to improve their performance. As a result, the inability to cope with the physiological demand of elite football match leads to an increased risk of injury (Ehrmann et al. 2016). As a matter of fact, Ehrmann et al. (2016) found that players stopped by an injury are the ones that performed higher training loads – GPS features such as total distance, high-intensity running distance and very-high-intensity running distance – in the week before injury compared to the fixed training loads. Thus, future investigations are scheduled in order to create an algorithm able to predict the injuries through the most important GPS features detected in this study.

The four most important GPS features able to characterize short-term football training cycles (Table 2.3) were the distance covered above 20 $W \cdot kg^{-1}$ (metabolic distance zonal), the acceleration above $2 \cdot s^{-2}$, the total distance and the distance covered above 25.5 $W \cdot Kg^{-1}$ and below 19.8$Km \cdot h^{-1}$ (explosive distance). No study in literature investigated the GPS features importance in football

training. The studies found in literature investigated at least a few features in each study (Akenhead, Harley, and Tweddle 2016; Malone et al. 2015; Scott et al. 2013). Thus, we believe that the most important features detected in our study are the best ones to describe the football training and could be useful to define the best training loads for obtaining performance enhancements. In conclusion, future investigations are necessary in order to detect the needed training loads in each short-term cycle day to the aim of achieving the best performance during the match.

2.2.6. CONCLUSIONS

The experiment proposed in this study detects a short-term cycle through the in-season training block by using machine learning process and autocorrelation analysis. The football trainings cycle detected is composed of two kinds of trainings: high and low intensity training loads performed in the days long before and close to the match, respectively. It is crucial for coaches and athletic trainers to define correct training loads in each short-term cycle day in order to control the training loads performed by the football players. The correct loads-recover ratio is useful to obtain the best performance on the match day and to reduce the risk of injury in players who do not perform trainings close to specific loads. Sport experts may use machine learning process to correctly predict the right training loads through in-season trainings. The data-driven approach aims to reduce the intra-training variability on the training periodization induced by coaches' and athletic trainers' intuitions and personal experience. Thus, the main practical application of this study is the capability of this machine learning process to provide an objective evaluation of the weekly training workload. In this way, coaches and athletic trainers could evaluate the fitness state and the effect of training loads on their players.

Due to the fact that this is only a preliminary study, a future work is planned in order to assess the influence of different short-term training cycles on football match performance. In particular, the aim of this study will be to provide a universal periodization strategy that could positively affect the performance in the match day.

2.2.7. CRITICAL ASPECTS

This study has some critical aspects that have to be highlight in order to avoid misunderstandings.

This study provides a method able to predict training loads that an athletic trainer wants to subject his football players to in each training in order to reach excellent results throughout the in-season football period. The results provided are specific to the football team investigated. The machine learning process provided in this study should be used by all the football teams in order to provide in each training day the training loads which the athletic trainers and coaches believe to be the most suitable to obtain the greatest performance during the match. Obviously, this process does not indicate which are the best training loads for football players to obtain enhancements in performance. It should be used only for standardizing training loads in each short-term training day through the in-season football period.

This is only a preliminary study aimed to detect a pattern into the football training week through the in-season period. Due to the fact that this aim is verified, it is possible to schedule future researches with the purpose of detecting the best training loads in each training day in accordance to the performance required during the match and predicting the injury risk based on the training loads discrepancy to the required ones.

2.3. STUDY 2: INJURY PREDICTION IN ELITE FOOTBALL PLAYERS BY MACHINE LEARNING PROCESS.

Rossi A.[1], Pappalardo L.[2], Cintia P.[2], Savino M.[1], Alberti G.[1], Iaia FM[1]

[1]School of Exercise Sciences, Department of Biomedical Sciences for Health, Università degli Studi di Milano

[2]Deparment of Computer Science, Univeristà di Pisa

2.3.1. ABSTRACT

The injury prediction and prevention are important matters for football teams. As a matter of fact, injuries could prejudice team's performance because of forced absence of essential players during the matches. Thus, the aim of this study is to provide a predictive model able to prevent injuries. To this aim, eighty training sessions of twenty-six elite football players were recorded by using STATSports Viper Global Position System (GPS). In order to predict the players' injuries, a Decision Tree Classifier (DT) was used. This machine learning process is based on the most important features detected by Linear Support Vector Classification (LSVC) features selection. In particular, LSVC selected the number of injuries that the player had succumbed through the season, the number of Acceleration above 2 $m \cdot s^{-2}$ and 3 $m \cdot s^{-2}$ and the distance in meters when the Metabolic Power (Energy Consumption per Kilogramme per second) is above the value of 25.5 W/Kg per minute as the most important features to the injury prediction aim. Moreover, a six-days predictive window was taken into account to create the machine learning process because it was found to be the best observational period to detect features discrepancy between players who get injured and those who do not. This DT demonstrates strong accuracy whether a player will get injured (AUC=0.88). In particular, DT showed that the first days after a player's return to the regular training program are those with the higher risk of injuries. In conclusion, this approach enables football teams to identify when their players need to pay more attention during trainings and matches in order to reduce the injury risk, while improving team strategy.

Keywords — GPS; Training Loads; Injury risk

2.3.2. INTRODUCTION

In last 20 years, the use of machine learning process and statistical analysis approaches in sports science have become popular (Pinder et al. 2011). Thanks to these approaches data scientists are being able to process large amounts of data recorded during sport performances (Barris and Button 2008). The development of several instruments such as global positioning system (GPS), gyroscope, accelerometer and high-speed video have allowed to record huge volumes of data able to provide insights of individuals' physical activity (an example of GPS data view is provided in Figure 3.1). Thus, the employment of machine learning process and statistical analysis approaches applied on large amounts of data permits to have more information on athletes providing intelligent decisions during coaching process.

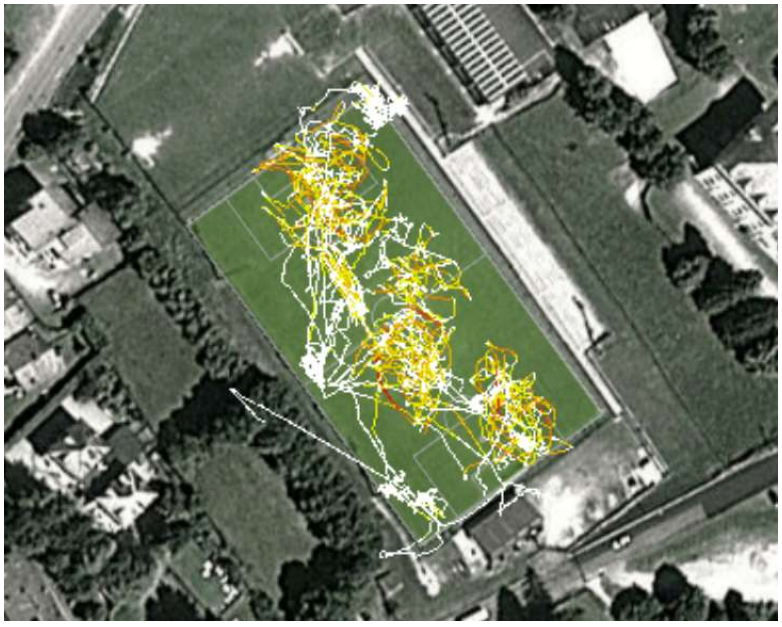

Figure 3.1. View of the data recorded by GPS on football field.

One of the most critical aspect in the coaching process is the reduction of injuries. This is due to the fact that injuries could impair team's performance caused by the forced absence of players to crucial matches. Thus, the injury prediction and prevention is becoming one of the most important

topic for sport researchers. As a matter of fact, in the last decade, it was demonstrated that any illness related to training loads is commonly viewed as 'preventable' (Tim J. Gabbett 2016). Therefore, the assessment of the players' loads during both trainings and match could be useful to provide an accurate model able to detect the injury risk. Actually, several investigations have found a significant connection between training loads and the injury rates (Tim J Gabbett 2004; Tim J. Gabbett and Jenkins 2011; Tim J. Gabbett and Ullah 2012; Rogalski et al. 2013). In particular, they have found that the higher are the training loads the higher is the likelihood to incur an injury. Moreover, Gabbet (Tim J Gabbett 2004) showed that reductions in training loads lead to a decrease of injury risk without affecting sport-related physical skills. Hence, an accurate estimation of training loads is required in order to obtain the maximum enhancement of physical activity yet keeping an acceptable injury rates (Tim J. Gabbett and Ullah 2012).

However, one of the first machine learning processes able to predict injuries was focused on match performance (e.g. minutes, competitions played and speed, and distance covered per player) (Talukder et al. 2016). The model which was created could predict 19% of the total injuries occurred on NBA basketball league (accuracy of 92%). Additionally, Talukder et al. (Talukder et al. 2016) stated that the most important features able to predict players' injuries by their machine learning process are the average speed during the competitions, the number of competitions played to date in the season, the average distance covered during competitions, the number of minutes played to date in the season and the average field goals attempted. Thus, this algorithm allows coaches and athletic trainers to provide smart decisions in order to prevent players' injuries on the basis of individuals' match performance. However, one of the most important limitation of the Talukder et al. (Talukder et al. 2016) study is the fact that they did not take into account the players' loads performed during both trainings and matches which could increase the ability of the machine learning process to predict injuries.

The purpose of this study is to provide a predictive model to prevent injuries on the basis of training loads. In this study, a supervised machine learning process is proposed in order to solve this

high-dimensional unbalanced binary classification problem. The major difficulty to solve this kind of problem is that the dataset contains only a few information about the minority class (only the 2% of events in Dataset were injuries). To solve this unbalanced problem, eighty in-season trainings of twenty-six Italian elite football players were recorded by using GPS. Additionally, non-traumatic injuries were recorded by the football team's medical staff. Despite the difficulty to solve this high-dimensional unbalanced binary classification problem - yet in light of the previous assertions - it is licit to suppose that different training loads patterns are detectable in players that are more likely to get injured and those who are not. The model provided in this study has strong performance in terms of accuracy and is able to provide a clear prediction about when players should be put at rest in accordance with their risk of injury. In particular, in order to predict injuries, it is highlighted the importance of the first days after an injury when players return to play. Moreover, to avoid misunderstanding, it is necessary to emphasize that this study provides a machine learning process able to predict injuries on a specific football team. Thus, to accurately predict injury, the algorithm presented in this study should be customized in accordance with training's loads performed by each football team.

2.3.3. MATERIALS AND METHODS

Dataset and measure – Twenty-six elite football players from an Italian Serie B club (age = 26 ± 4 yrs; height = 179 ± 5 cm; body mass = 78 ± 8 kg) took part in the study during the 2014-2015 season. Six central backs, three fullbacks, seven midfielders, eight wingers and two forwards were recruited. Goalkeepers were not included in the study. Participants gave their written informed consent to participate after having received an explanation of the study aim and methods. The study protocol was approved by the Ethical Committee of University of Milan.

Players' physical activity of 23 training weeks were monitored using a portable non-differential 10 Hz global position system (GPS) integrated with 100 Hz 3-D accelerometer, a 3-D gyroscope, and a 3-D digital, Northern Ireland compass (STATSports Viper). The interaction among these devices is

useful to create several GPS features able to describe the sport performance (Duncan, Badland, and Mummery 2009). Each player wore a tight vest, and the receiver was placed between their scapulae. All devices were always activated 15-min before the data collection both to allow acquisition of satellite signals in accordance with manufacturer's instructions. In order to avoid that inter-unit error players wore the same GPS device for each training session. After recording, data were downloaded to a computer where the software package (Viper Version 2.1, STATSports 2014) computed a set of features for every player and training session. If players did not complete the training session or the GPS have not correctly recorded the players' position, these data were not taken into account during the analysis (30% of the individuals' training sessions were deleted from the dataset). Moreover, the club's medical staff recorded non-traumatic injuries after each training and match day. The Sport Club gave the permission to use the data for the purpose of research according to privacy policy.

A total of 80 team training sessions and 2,080 individual trainings were recorded. Training Load features computed by the softer package based on the GPS data. In addition to GPS data, players' characteristics (i.e., age, weight, height and role), the number of injuries occurred before each training session (Previous Injuries), play time in the previous match, the number of match played by each player were taken in consideration to predict players' injury. The description of each feature was provided in Table 3.1.

The dataset was composed as follows:

$$< \textit{Player ID, Training Sequential Index, Injury, f1, f2, f3…fn} >$$

where $f_1,\ldots f_n$ are the independent features and Injury is the dependent variable. The latter is characterized by two values (binary classification problem): 1 when the players had an injury in the following training/match day and 0 otherwise.

Table3.1. Description of the GPS and players' characteristics features.

GPS features description	
Total Distance	Distance (meters) covered during the training session
High Speed Running Distance	Distance (meters) covered above 5.5m/s
Metabolic Distance	Distance (meters) covered at metabolic power (di Prampero et al. 2005)
High Metabolic Load Distance	Distance (metres) travelled by players when their Metabolic Power (Energy Consumption per Kilogramme per second) is above the value of 25.5 W/Kg. This value of 25.5 corresponds to when a player is running at a constant speed of 5.5m/s on grass or when they are performing a significant acceleration or deceleration activity, for example if they are accelerating from 2 to 4m/s over 1 second (Gaudino et al. 2013)
Explosive Distance	Distance (meters) covered above 25.5 W/Kg and below 19.8 Km/h
High Metabolic Load Distance per minute	Average of High Metabolic Load Distance covered in each minute
Acceleration above 2m/s2	Number of accelerations measured on the basis of the change in GPS speed data using established statistical methods. To count as an acceleration, the increase in speed must take place for at least half a second with maximum acceleration in the period at least 2m/s/s
Acceleration above 3m/s2	Number of accelerations measured on the basis of the change in GPS speed data using established statistical methods. To count as an acceleration, the increase in speed must take place for at least half a second with maximum acceleration in the period at least 3m/s/s
Deceleration above 2m/s2	To count as deceleration, the decrease in speed must take place for at least half a second for an activity to be counted as deceleration. Also, the maximum deceleration in the period must be at least 2m/s/s
Deceleration above 3m/s2	To count as deceleration, the decrease in speed must take place for at least half a second for an activity to be counted as deceleration. Also, the maximum deceleration in the period must be at least 3m/s/s
Dynamic Stress Load	Dynamic Stress Load is the total of the weighted impacts, which is based on accelerometer values of magnitude above 2g. Impacts are a mixture of collisions and step impacts while running. In sports such as Football Dynamic Stress Load is very much dominated by running step peak impacts. In this way the Dynamic Stress Load can, as a measure of internal physiological response, measure the level of fatigue of a player.
Fatigue index	Ratio between Dynamic Stress Load and Speed Intensity. Speed Intensity is calculated by assigning an intensity value to the players speed on the basis of the Speed Intensity Weighting Function (i.e., $\sum W_i DT_i$ where i is the number of time points, Wi is the speed intensity weighting for time point -exponential function- and DT is the difference between two consecutive time points)
Players' Characteristics features description	
Age	Players' age
BMI	Body Mass Index (ratio between weight expressed in kg and the square of height expressed in meters)
Role	Players' role (i.e., central backs, fullbacks, midfielders and wingers)
Previous Injuries	Number of injuries that players have occurred before each training session
Play time	Minutes of play in the previous match
Match played	Number of matches that the player have played before each training session

Experiments – Since the players' probability of non-traumatic injuries depends on the individual's recent workload history, we considered the *n* most recent training session of every player. We aggregated the *n* most recent values for every measures by using exponential weighted moving average (EWMA). In our experiments we varied *n*=1…10 to select the value of *n* leading to the best classification results (see Appendix Section 1).

We first performed a features selection task to select the relevant features for the injury prediction. This process also reduces the dimensionality of the features space to overcome the risk of overfitting and it makes the machine learning model easier to be interpreted by researchers and field experts (James et al. 2013). We used Linear Support Vector Classification (LSVC) for the feature selection because it is robust to the presence of sparse and noisy data (James et al. 2013) (see Appendix Section 2). In addition, we performed Decision Tree (DT) and the Extra Tree Random Forest (ETRFC) classifiers to predict injuries on the complete dataset (CD), on the dataset created by LSVC feature selection (FSD) and on the dataset with Previous Injuries as unique features (PID) varying *n*=1…10. All the classification tasks were performed using Stratified-Kfolds cross-validation (k=10). We compared the classifier with two baselines which predict injury class in two ways: (i) by respecting the distribution of classes (DCs); and (ii) by selecting always the most frequent class (DC_{mf}). We built a third baseline classifier based just on the Pervious Injury feature (DCpi) to assess the contribution of fall back into the injury prediction process. This classifier predicted injuries when players had had a previous injury. We evaluated the quality of the classification tasks by four measures: precision, recall, F1-score and area under the ROC curve (AUC) (see Appendix Section 3). Finally, Mann-Whitney U test was performed in order to detect differences between No-Injury and Injury predicted events. No parametric unpaired t-test was performed because the assumption of normal data distribution was not met using Shapiro-Wilks' Normality test. The level of significance was set at $p<0.05$. Python 2.7 was used to perform predictive machine learning algorithm and the statistical analysis.

In addition, twenty-three training-test sets were used in order to assess the accuracy of machine learning process to predict injuries as the season goes by (Real Scenario). In particular, the learning process was performed through subsequent training weeks (i.e. from one to twenty-three) to predict injuries in the next one (figure 3.2). In each training-test set, DT classifier on the feature selected by LSVC feature selection was performed to predict injuries. All the training tasks were performed using Stratified-Kfolds cross-validation (k=10). In addition, the classification report (i.e. precision, recall and F1-score) was computed to detect the accuracy of the machine learning process when the season went by.

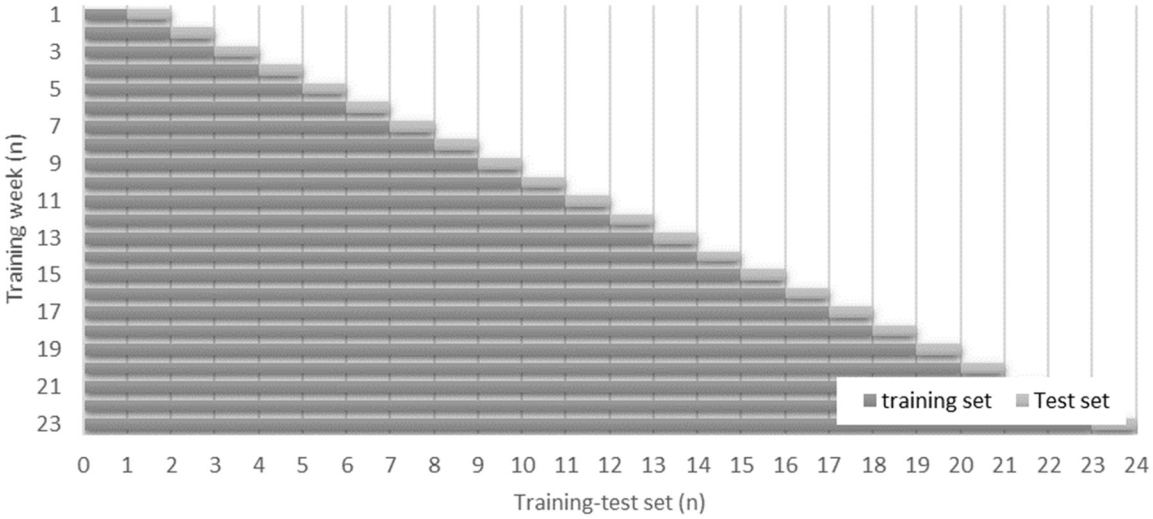

Figure 3.2. Organization of training and test sets to test the accuracy of machine learning process to predict injuries.

2.3.4. RESULTS

Table 3.2 shows the classification reports of machine learning processes and dummy classifiers for the moving average with *n*=6 which it is found to be the best moving average to predict injuries (Figure 3.3). Moreover, it was found that the DT classifier performed on FSD is the best machine learning process to predict injuries (60.9% of the injury was detected. See Table 3.2 and Figure 3.4). The features selected by LSVC feature selection method are Previous Injuries, Acceleration above 2 $m \cdot s^{-2}$, High Metabolic Load Distance per Minute and Acceleration above 3 $m \cdot s^{-2}$. ROC curves and AUC of the machine learning process that was able to better predict injuries (i.e.

DT classifier) are provided in Figure 3.5. In addition, 82.13%, 7.41%, 7.12% and 3.34% are the features importance percentages to predict injuries of Previous Injuries, Acceleration above 2 $m \cdot s^{-2}$, Acceleration above 3 $m \cdot s^{-2}$ and High Metabolic Load Distance per Minute, respectively. Means, Standard Deviations and the results of the statistical analyses of the features selected grouped in accordance to the injury prediction performed by DT Classifier are provided in Table 3.3 and Figure 3.6. In addition, the Decision Tree process to predict injuries is showed in Figure 3.7. Moreover, in Table 3.4 the different injury scenarios detected by the DT is resumed. This table provides the variables thresholds that are able to discriminate players who got injured, the probability of injury, and the number of injuries observed and predicted by the DT. Figure 3.8 shows the classification report and the number of injuries predicted when the machine learning process has been performed through the football season. It is noticeable that this process was be able to correctly predict 9 of 21 injuries if it was performed when the season go by.

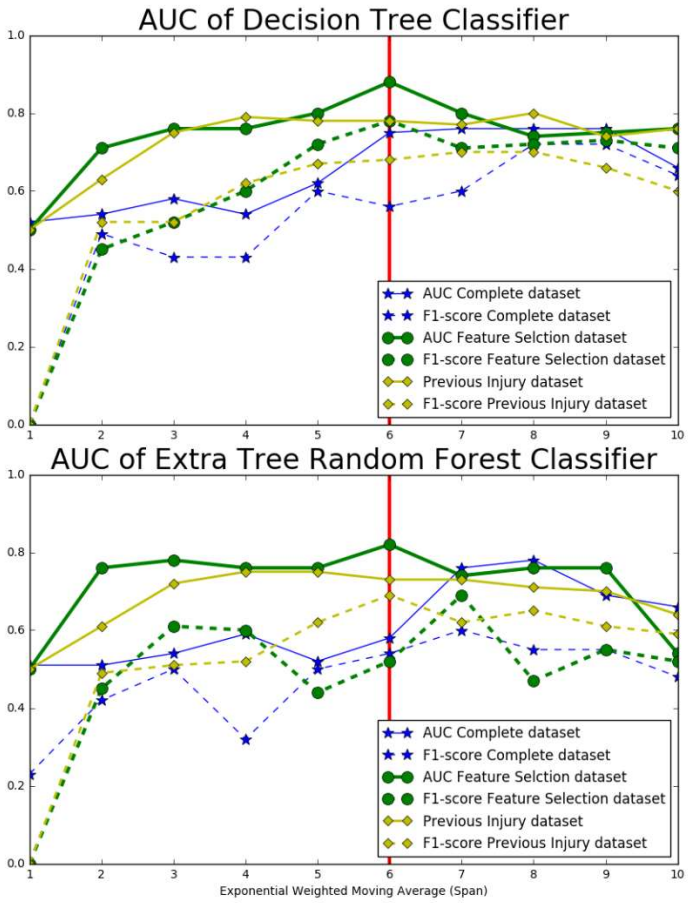

Figure 3.3. AUC average values of Stratified-Kfolds cross-validation for different Exponential Weighted Moving Average span on Complete and Feature Selection datasets.

Table 3.2. Classification report of DT, ETRFC and DC using n=6 moving average.

			Precision	Recall	F1-score	AUC
Complete Dataset	DT	No injury	0.98	1.00	0.99	**0.81**
		Injury	0.67	0.10	0.17	
	ETRFC	No injury	0.99	1.00	0.99	0.75
		Injury	0.64	0.52	0.56	
	DCs	No injury	0.98	0.97	0.98	0.51
		Injury	0.00	0.00	0.00	
	DCmf	No injury	0.98	1.00	0.99	0.49
		Injury	0.00	0.00	0.00	
	Dcpi	No injury	1.00	0.65	0.79	0.80
		Injury	0.06	0.91	**0.49**	
Feature Selection Dataset	DT	No injury	0.99	0.99	0.99	**0.88**
		Injury	0.80	0.76	**0.78**	
	ETRFC	No injury	0.98	1.00	0.99	0.86
		Injury	1.00	0.19	0.32	
	DCs	No injury	0.98	0.97	0.98	0.49
		Injury	0.04	0.05	0.04	
	DCmf	No injury	0.98	1.00	0.99	0.50
		Injury	0.00	0.00	0.00	
	DCpi	No injury	1.00	0.65	0.79	0.80
		Injury	0.06	0.91	0.49	
Previous Injuries Dataset	DT	No injury	0.99	0.99	0.99	**0.81**
		Injury	0.65	0.62	**0.63**	
	ETRFC	No injury	0.99	1.00	1.00	0.76
		Injury	1.00	0.22	0.39	
	DCs	No injury	0.98	0.98	0.98	0.49
		Injury	0.09	0.10	0.09	
	DCmf	No injury	0.99	1.00	0.99	0.50
		Injury	0.00	0.00	0.00	
	DCpi	No injury	1.00	0.65	0.79	0.80
		Injury	0.06	0.91	0.49	

Table 3.3. Means and Standard Deviations and statistical differences of the feature selected grouped by the injury prediction. Boxplots of these results are provided in Figure 3.3.

	Previous Injuries ***	Acceleration above 2 $m \cdot s^{-2}$ ***	High Metabolic Load Distance per Minute ***	Acceleration above 3 $m \cdot s^{-2}$ ***
No-injury	0.55 ± 0.89	65.60 ± 16.37	8.91 ± 3.25	16.42 ± 6.66
Injury	0.91 ± 0.79	60.56 ± 12.48	7.96 ± 1.76	16.95 ± 5.28

Mann-Whitney U test statistical significance: *p<0.05, **p<0.01, ***p<0.001

Table 3.4. Decision Tree thresholds for 7 different injury scenarios.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**	**Scenario 6**	**Scenario 7**
Previous Injuries	<=0.40	<=1.36	>=2.05	<=1.51	<=1.36	<=1.08	>=1.08
Acceleration above 2 m·s^{-2}	---	<=53.96	<=57.96	<=77.16	<=57.93	<=49.20	<=78.42
HML per min	---	---	---	---	>=7.53	<=6.80	---
Acceleration above 3 m·s^{-2}	---	---	>=16.67	---	---	---	---
Probability of injury	6%	80%	33%	9%	50%	33%	12%
Number of injuries observed	11	3	2	2	1	1	1
Number of injuries predicted	8	1	1	1	1	1	0

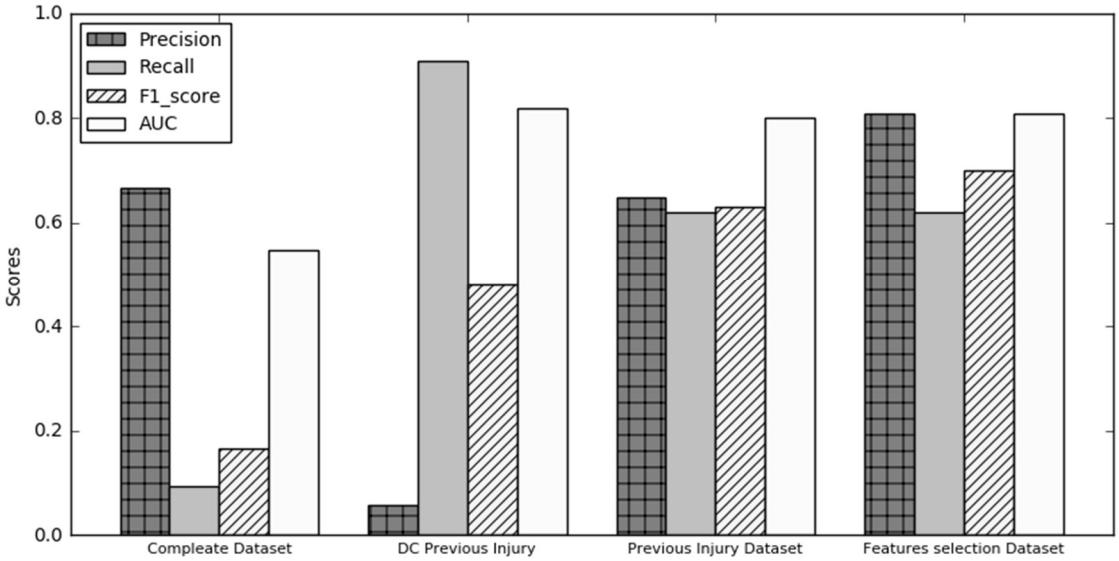

Figure 3.4. Classification report of the Decision tree by different datasets ordered by F1 score

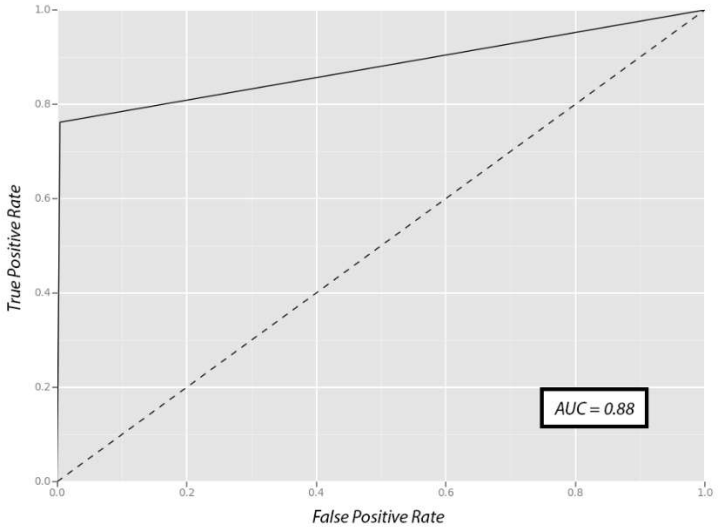

Figure 3.5. ROC curves of Decision Tree Classifier and associated Area Under the Curve (AUC)

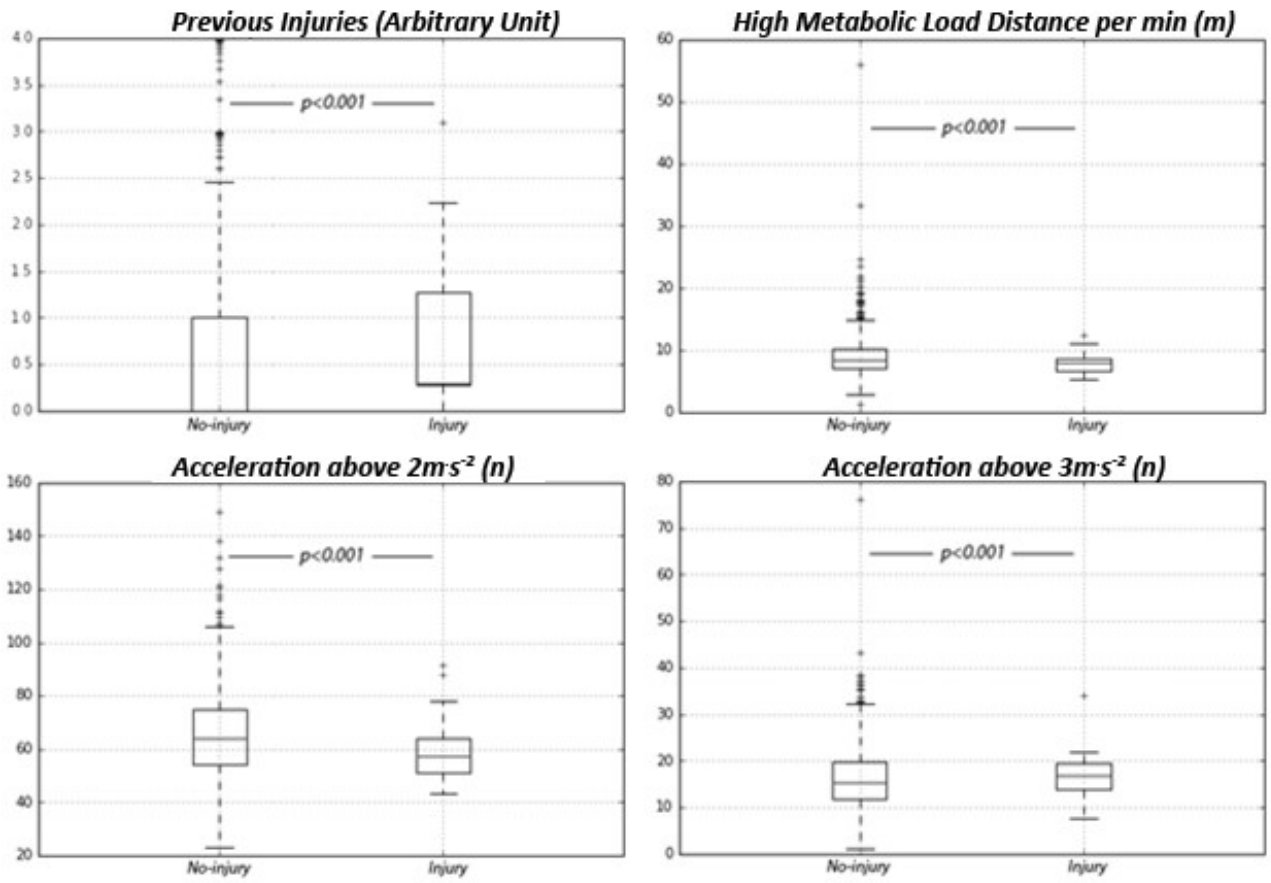

Figure 3.6. Boxplots of Previous Injuries, Acceleration above 2 $m \cdot s^{-2}$, High Metabolic Load Distance per Minute and Acceleration above 3 $m \cdot s^{-2}$ grouped by no-injury/injury predicted using Decision Tree Classifier. Numerical means and standard deviations are provided in Table 3.3.

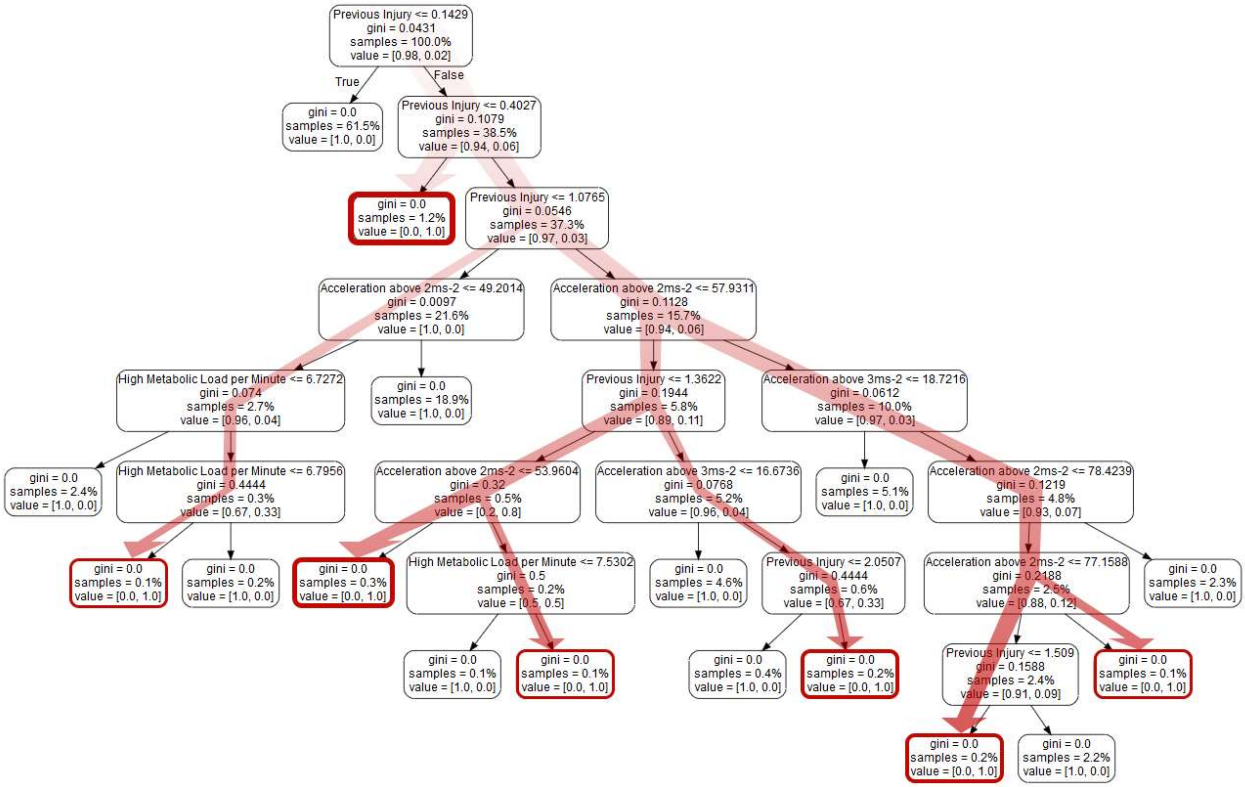

Figure 3.7. Decision Tree Classifier to predict injuries. The red line shows the different scenarios that describe injuries. The bigger is the line, the higher is the number of injuries that the scenario is able to describe. The value scores reports the percentage of no-injury/injury risk that players have in each leaf.

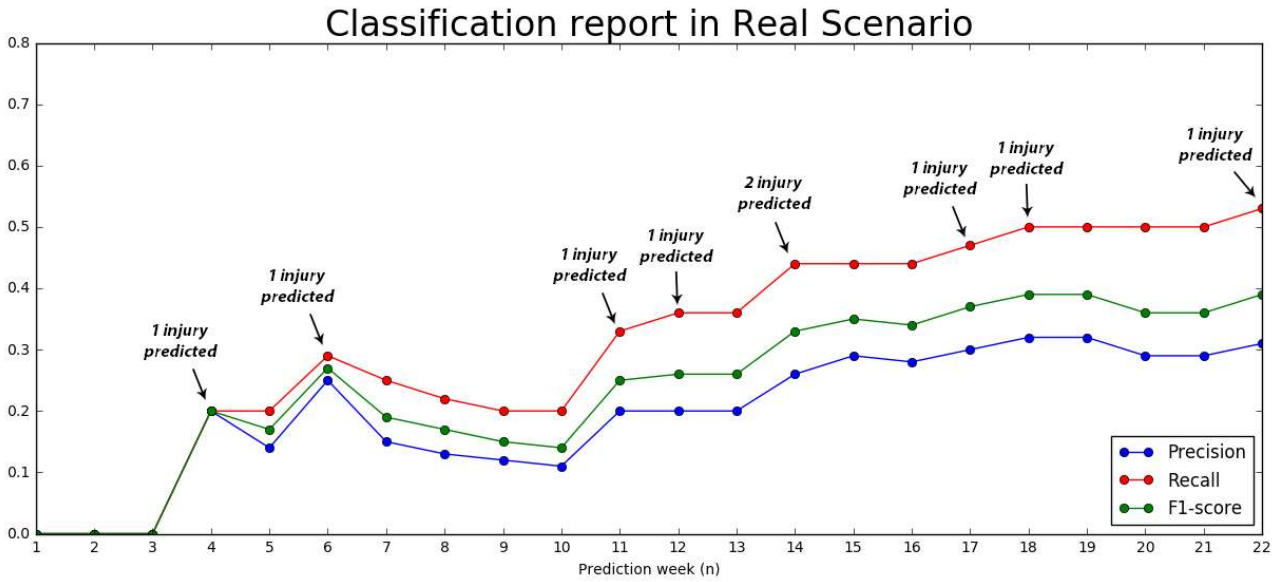

Figure 3.8. Classification report of each prediction week in the real scenario. It reports the number of injuries correctly predicted when the season went by. The DT classifier could detect 9 injuries.

2.3.5. DISCUSSION

This is one of the first studies which tried to predict injuries by using machine learning processes. The novelty of this study is to provide a mathematical method based on players' training loads able to detect when a player is going to be injured. The growing interest towards this topic is due to the fact that injuries could impair team's performance caused by the forced absence of essential players during crucial matches. Hence, an accurate prediction of injuries could have beneficial effects on team performance and consequently on football teams' finance.

Our results showed that it is possible to predict injuries by using machine learning process. As a matter of fact, it was found that the DT classifier is the best machine learning process to this aim compared to ETRFC. As a matter of fact, the higher accuracy assessed by F1-scores was detected when the injury was predicted using this machine learning process (Table 3.2). Moreover, due to the fact that a lower prediction accuracy was detected by both stratified and most frequent dummy classifiers compared to DT one (Table 3.2), the latter can be considered a valid method to the injury prediction aim. In addition, it was found that the prediction accuracy of the DT increases when the predictive window refers to a 6 days' span (Figure 3.2). Thus, athletic trainers and coaches should take into account this period to detect possible players' illness that could induce injury. As a matter of fact, in professional Australian football it was found that the ratio between 6-days acute time window and 21-days chronic one is able to explain non-contact injury risk in the match day and at least in the next 5 days assessing discrepancy in the distance covered at a velocity of between 18 and 24 km/h (Carey et al. 2016). Differently, Talukder at. al (Talukder et al. 2016) showed that is possible predict injuries within 7-days after a match using players' technico-tactical match profiles (AUC=0.92) recorded in 14-days predictive windows. The difference in predictive windows span detected in previous studies could be explained by the different nature of the features used to this aim (e.g., trainings workloads and technico-tactical matches features). Moreover, different sports, coaching process and players' physical characteristics affected the prediction process as well.

The choice of the most important features able to discriminate injuries improves the ability of the DT classifier (Table 3.2 and Figure 3.3). Indeed, in a 6-days' span period, the field experts could monitor only a few features (i.e. Previous Injuries, Acceleration above 2 $m \cdot s^{-2}$, Acceleration above 3 $m \cdot s^{-2}$ and High Metabolic Load Distance per Minute) to discriminate the players more likely to incur injuries in the next training or match. As a matter of fact, significant discrepancies of these features were found between the injury's and no-injury's 6-days records (Table 3.3 and Figure 3.5). In contrast to previous studies, which found an influence of high loads on injury events (Tim J. Gabbett 2016; Tim J Gabbett 2004; Tim J. Gabbett and Ullah 2012; Tim J. Gabbett and Jenkins 2011; Rogalski et al. 2013), this study shows that the players who performed a significantly lower number of accelerations above $2m \cdot s^{-2}$ and High Metabolic Load distance per minute are those who are more likely to get injured (Table 3.3 and Figure 3.5). This is probably due to the fact that the players who got injured are also those who were not able to perform the training loads scheduled within six days before the injury events because of latent physical problems which had induced non-traumatic injuries. Moreover, due to the fact that a high number of relapses were observed in the dataset (i.e. about 62% of injuries happened few days after the injured player returned to regular training activities), it seems that the players who incurred an injury had not completely recover from their previous one, thus having an inability to perform regular training loads.

An accurate analysis of the decision tree provided in Figure 3.7 is needed in order to better understand how the machine learning process works to solve this unbalanced binary problem. 7 different scenarios were detected in the decision tree and were resumed in Table 3.4. When players perform 6-days trainings in accordance with the specific characteristics described in the scenarios, they are more likely to get injured in the next training or match (F1-score=0.78). The odds provided in each leaf of the decision tree (i.e. values score) represent the probability to get injured or not when players show a specific workloads scenario. Moreover, carefully analysing the decision tree provided in Figure 3.6 and the scenarios in Table 3.4, it is noticeable that the injuries predicted by our machine learning process are those derived from events happened within 6-days after the players have returned

to train (see supplementary materials section 4). Hence, the high importance of the Previous Injuries feature is due to the fact that 13 of the 21 injuries observed in the dataset happened within a few days after a player had returned to regular training activities. In particular, Scenario 1 (Table 3.4) shows that the 11 of the 21 injuries could be described using only Previous Injuries features. However, despite the 76.64% of the prediction variance was explained by Previous Injuries the other variable could help the machine learning to accurately predict injuries. As a matter of fact, Table 3.2 shows the lower ability of the machine learning processes to predict injuries using Injuries Previous as unique predictive features (Previous Injuries Dataset) than the ones that use the feature selected by LSVC feature selection (Feature Selection Dataset). In addition, the lower accuracy was detected using the Complete dataset as well. Thus, only a few GPS features should be taken into account to accurately predict injuries.

Athletic trainers and coaches could be interested in the prediction of injuries as the season goes by. To this aim, machine learning process performed in a Real Scenario showed that its accuracy increases in subsequent training weeks (Figure 3.8). As a matter of fact, 42.86% of the injuries could have been identified and avoided through the football season using this technique. In this way, the total number of injuries could have been drastically reduced thus keeping all the players available to all matches.

2.3.6. LIMITATIONS OF THIS STUDY

Although this study has reached its aims, there were some limitations. The Dataset used to train the machine learning process was not complete. Not all the players' trainings were documented because of technical problem on GPS tools during the recording. Moreover, the GPS data of the matches were not documented because it was not allowed in the football season 2014-2015 when the data were recorded. Thus, future investigations are required in order to create a most accurate machine learning process using a dataset, which take into consideration the GPS data of the matches in addition

to the training ones'. This dataset implementation is needed because the higher loads performed during the matches probably induce the higher risk of injuries, as well.

2.3.7. CONCLUSIONS

This study presents a predictive model that can be used to detect football players who are more likely to incur injuries. Thus, athletic trainers, coaches and physiotherapists could strategically put at rest the players in order to avoid possible injuries by assessing Previous Injuries, Acceleration above 2 $m \cdot s^{-2}$, Acceleration above 3 $m \cdot s^{-2}$ and High Metabolic Load Distance per Minute in a period of 6 days. In particular, the football team analysed in this study should keep under control the discrepancy of these features when players return to the regular training, because of the numerous times they fell back into injuries.

In order to avoid misunderstandings, it is important to highlight that the results provided in this investigation refer to the football team recruited for this study. The diversities of coaching processes and physical characteristics of the football players in each team do not permit to make inferences about the football players' population. Hence, this study provides only an example of how a machine learning process could be used to predict injuries. However, an enlargement of the dataset including different football teams might allow to build a generalized algorithm able to predict injuries. In addition, with more injury case will be possible to relax the binary label (injury/no-injury) used in this study according to the severity of injuries in order to produce a more accurate injury prediction. Thus, future study are required to better investigate the injury prediction problem using a huge and more detailed dataset.

2.3.8. APPENDIX

Section 1 – A moving average is a series of averages of different ordered subsets derived from the full data set. Therefore, the average of the initial fixed subset of data (from the first value of the time series to the n one in accordance to the fixed subset size) is the first element of the moving average. Then, the subgroup of the dataset shifts forward (i.e., excluding the first value of the fixed series and including the next one) until the last value in order to create a moving average dataset. Moreover, in order to emphasize particular values in the fixed subset, a moving average may be unequal weighted in accordance to specific functions (e.g., span, centre of mass and half-life). In particular, in this study, Exponential Weighted Moving Average (EWMA) was used.

The EWMA decreases exponentially the weights of the value (Lowry et al. 1992; Lucas and Saccucci 1990). In particular, the more recent a value is, the more it is weighted in an exponential function according to a decay α:

$$\alpha = 2/(span+1)$$

an example of span=10 is provided in Figure 4.1. In accordance with the exponential function the moving average is computed as:

$$EWMA=(x^{a1}+x^{a2}+\ldots+x^{an})/n$$

In this study, we vary n=1…10 in order to detect the best n to predict injuries (Figure 4.2).

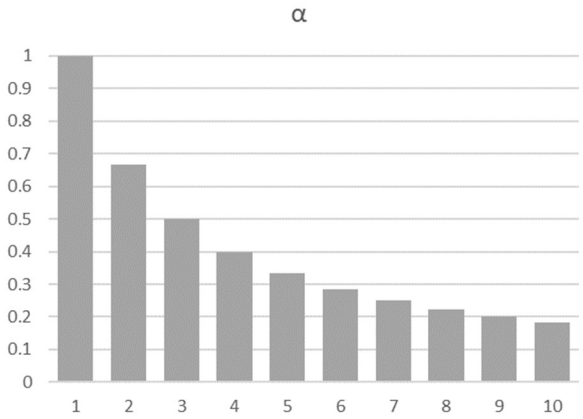

Figure 4.1. α of a span period of n=10.

Figure 4.2. Description of moving average approach used to predict injuries.

Section 2 **–** Feature Selection is a statistical technique able to select a subset of the most relevant features to solve a machine learning problem. This technique is widely used because it reduces the dimensionality of the dataset in order to simplify the interpretation of the machine learning process, enhancing the possibility to generalize the process and reducing overfitting (i.e. reduction of variance) (Flach 2012). In this study, a features selection based on Linear Support Vector Classification (LSVC) was used (James et al. 2013). This technique was employees because the LSVC tries to separate a binary labelled dataset with a hyper-plane that has a maximal distance between them (Maximal margin hyper-plane) working in combination with the kernels technique that automatically realizes a linear or non-linear mapping of the feature space. Thus, these LSVC characteristics permit us to select the best subset of features able to detect injuries in each sliding windows dataset (James et al. 2013).

Section 3 **–** The quality of the classification tasks was assessed by four measures: precision, recall, F1-score and area under the ROC curve (AUC) (Goutte and Gaussier 2005). In particular, the precision or positive predictive value is computed as:

$$Precision = \frac{True\ positive}{True\ positive + Fal \quad positive}$$

This value represents the fraction of retrieved events in a class by a classifier that are correctly classified in accordance to the true classification. Instead, the recall, true positive rate or sensitivity are computed as:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$

This denotes the fraction of true positive events that are successfully retrieved by a classifier. Moreover, the precision and the recall are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a model. The F1-score is the harmonic mean between these values computed as:

$$F1\ score \quad 2 \cdot \frac{Precision \cdot Recall}{Precision + Recal}$$

In addition, the accuracy of the classifier was assessed by using the area under the receiver operating characteristic (ROC) curve. An area of 1 represents a perfect classification and an area of 0.5 represents a worthless one. The ROC curve was created by plotting True Positive Rate (i.e. Recall) against the False Positive Rate (i.e. ratio between False Positive and the sum of False Positive and True Negative, which represents the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points) at many different thresholds. Therefore, the Area Under the ROC Curve (AUC) is the probability that a classifier will rank randomly chosen positive events higher than a randomly chosen negative one (Hajian-Tilaki 2013).

Section 4 – Previous Injuries permits to detect fall-backs into injury when players return to play. Table 4.1 provides specific values obtained by EWMA (n=6) to define the days when players incur injuries again after a previous one.

Table 4.1. Previous Injuries value obtained by Exponential Weighted Moving Average with n=6

	Days when player returns to play after an injury					
n of previous injuries	**1**	**2**	**3**	**4**	**5**	**6+**
1	0.17	0.33	0.50	0.67	0.83	1.00
2	1.17	1.27	1.33	1.38	1.43	1.47
3	1.63	1.72	1.77	1.81	1.84	1.86
4	2.03	2.10	2.15	2.17	2.20	2.22

Values in 6+ are referred to the 6 days when the players return to play and it is maintained since player get injured again.

CONCLUSIONS

In accordance with the results found in the two studies presented in this thesis, it is possible to assert that Machine Learning processes could be useful in order to help football coaches and athletic trainers on coaching process. As a matter of fact, as showed in the first work presented in this thesis, this technique could provide an objective evaluation of the weekly training loads that could be helpful for sport field experts to maintain the higher performance throughout the season balancing loads-recover ratio during the football training week. On the basis of the first investigation, future works are schedule with the purpose of comparing different kind of weekly training periodization throughout the football seasons or among different football clubs in order to define the best one able to positively affect the performance in the match day. In this way, it will possible to mathematically demonstrate the theory about sport team periodization by a high-dimension generalized model of the in-season short-term training cycle.

Moreover, as showed in the second work provided in this thesis, it could also be possible to detect players' risk of injuries by using a machine learning approach. This process could be useful to detect possible forced absence of essential players during the matches or training because of an injury. Using the algorithm provided in the second study, coaches and athletic trainers could detect when players have to rest from physical activity in order to reduce their risk of injuries and consequently the risk of force absence to essential matches. In this way, football clubs could schedule weekly training workloads program in accordance to players' demands to maximize the training effect on match performance with an acceptable risk of injury. Due to the fact that the injury prediction work is still very much in the beginning, future investigations are needed in order to improve the algorithm enlarging the datasets with data derived from matches and recovery time, relaxing the injury class (i.e., from binary to multiclass machine learning problem in accordance with the severity of injuries) and testing the algorithm on different football club in order to generalize the decision rules.

In this thesis, it was provided only two example that testify the potentiality of the machine learning process applied on sport sciences. As a matter of fact, this technique could be also used to better investigate several aspects of both the individuals' well-being and athletes' performances thanks to the big amount of information that we are now able to record due to the recent technological advent. Until now, coaches and athletic trainers make predictions on the game results based on their experience, instinct, and/or gut feeling. However, a quantitative prediction which uses some form of objective data to predict the outcome is useful to reduce the subjectivity of making predications based on instinct (Leung and Joseph 2014). In conclusion, thanks to this technique, sport scientists could solve old problems in a better way or new problems in the best one.

REFERENCE

Akenhead, Richard, Jamie Harley, and Simon Tweddle. 2016. "Examining the External Training Load of an English Premier League Football Team with Special Reference to Acceleration:" *Journal of Strength and Conditioning Research*, January, 1. doi:10.1519/JSC.0000000000001343.

Astorino, Todd A., Peter A. Tam, Jeremy C. Rietschel, Stephen M. Johnson, and Thomas P. Freedman. 2004. "Changes in Physical Fitness Parameters during a Competitive Field Hockey Season." *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 18 (4): 850–54. doi:10.1519/13723.1.

Aughey, Robert J. 2011. "Applications of GPS Technologies to Field Sports." *International Journal of Sports Physiology and Performance* 6 (3): 295–310.

Baker, D. 2001. "The Effects of an in-Season of Concurrent Training on the Maintenance of Maximal Strength and Power in Professional and College-Aged Rugby League Football Players." *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 15 (2): 172–77.

Baker, Daniel. 1998. "Applying the In-Season Periodization of Strength and Power Training to Football." *Strength & Conditioning Journal* 20 (2): 18–27.

Barnes, C., D. Archer, B. Hogg, M. Bush, and P. Bradley. 2014. "The Evolution of Physical and Technical Performance Parameters in the English Premier League." *International Journal of Sports Medicine* 35 (13): 1095–1100. doi:10.1055/s-0034-1375695.

Barris, Sian, and Chris Button. 2008. "A Review of Vision-Based Motion Analysis in Sport." *Sports Medicine* 38 (12): 1025–1043.

Bartlett, Jonathan D., Fergus O'Connor, Nathan Pitchford, Lorena Torres-Ronda, and Samuel J. Robertson. 2016. "Relationships Between Internal and External Training Load in Team Sport Athletes: Evidence for an Individualised Approach." *International Journal of Sports Physiology and Performance*, May. doi:10.1123/ijspp.2015-0791.

Ben Abdelkrim, Nidhal, Carlo Castagna, Saloua El Fazaa, and Jalila El Ati. 2010. "The Effect of Players' Standard and Tactical Strategy on Game Demands in Men's Basketball:" *Journal of Strength and Conditioning Research* 24 (10): 2652–62. doi:10.1519/JSC.0b013e3181e2e0a3.

Berg, Dr Mark de, Dr Marc van Kreveld, Prof Dr Mark Overmars, and Dr Otfried Cheong Schwarzkopf. 2000. "Computational Geometry." In *Computational Geometry*, 1–17. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-662-04245-8_1.

Berman, Jules J. 2013. "Introduction." In *Principles of Big Data*, xix–xxvi. Boston: Morgan Kaufmann. http://www.sciencedirect.com/science/article/pii/B9780124045767099809.

Bialkowski, A., P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. 2014. "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data." In *2014 IEEE International Conference on Data Mining*, 725–30. doi:10.1109/ICDM.2014.133.

Borresen, Jill, and Prof Michael Ian Lambert. 2012. "The Quantification of Training Load, the Training Response and the Effect on Performance." *Sports Medicine* 39 (9): 779–95. doi:10.2165/11317780-000000000-00000.

Borrie, Andrew, Gudberg K. Jonsson, and Magnus S. Magnusson. 2002. "Temporal Pattern Analysis and Its Applicability in Sport: An Explanation and Exemplar Data." *Journal of Sports Sciences* 20 (10): 845–52. doi:10.1080/026404102320675675.

Brink, Michel S., Esther Nederhof, Chris Visscher, Sandor L. Schmikli, and Koen A. P. M. Lemmink. 2010. "Monitoring Load, Recovery, and Performance in Young Elite Soccer Players." *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 24 (3): 597–603. doi:10.1519/JSC.0b013e3181c4d38b.

Brown, Lee E., and Mike Greenwood. 2005. "Periodization Essentials and Innovations in Resistance Training Protocols:" *Strength and Conditioning Journal* 27 (4): 80–85. doi:10.1519/00126548-200508000-00014.

Carey, David L, Peter Blanch, Kok-Leong Ong, Kay M Crossley, Justin Crow, and Meg E Morris. 2016. "Training Loads and Injury Risk in Australian Football—differing Acute: Chronic Workload Ratios Influence Match Injury Risk." *British Journal of Sports Medicine*, October, bjsports-2016-096309. doi:10.1136/bjsports-2016-096309.

Carli, G., C. L. Di Prisco, G. Martelli, and A. Viti. 1982. "Hormonal Changes in Soccer Players during an Agonistic Season." *The Journal of Sports Medicine and Physical Fitness* 22 (4): 489–94.

Cavagna, Giovanni A., L. Komarek, and Stefania Mazzoleni. 1971. "The Mechanics of Sprint Running." *The Journal of Physiology* 217 (3): 709–21.

Cavanillas, José María, Edward Curry, and Wolfgang Wahlster, eds. 2016. *New Horizons for a Data-Driven Economy*. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-21569-3.

Chapman, Brian E., and Jeannie Irwin. 2015. "Python as a First Programming Language for Biomedical Scientists." In *Proceedings of the 14th Python in Science Conference (SCIPY 2015): Published Online at Http://conference. Scipy. org/proceedings/scipy2015/(last Accessed 23-09-2015)*. http://conference.scipy.org/proceedings/scipy2015/pdfs/brian_chapman.pdf.

Charbonneau, Danielle, Julian Barling, and E. Kevin Kelloway. 2001. "Transformational Leadership and Sports Performance: The Mediating Role of Intrinsic Motivation." *Journal of Applied Social Psychology* 31 (7): 1521–1534.

Choi, Mona. 2014. "Book Review: Data Smart: Using Data Science to Transform Information into Insight." *Healthcare Informatics Research* 20 (3): 243–44. doi:10.4258/hir.2014.20.3.243.

Cintia, Paolo, Salvatore Rinzivillo, and Luca Pappalardo. 2015. "A Network-Based Approach to Evaluate the Performance of Football Teams." In *Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal*. https://www.researchgate.net/profile/Luca_Pappalardo/publication/280520469_A_network-based_approach_to_evaluate_the_performance_of_football_teams/links/55b7383408ae9289a08bd85d.pdf.

Clemente, Filipe Manuel, Micael Santos Couceiro, Fernando Manuel Lourenço Martins, and Rui Sousa Mendes. 2015. "Using Network Metrics in Soccer: A Macro-Analysis." *Journal of Human Kinetics* 45 (March): 123–34. doi:10.1515/hukin-2015-0013.

"Community Cleverness Required." 2008. *Nature* 455 (7209): 1–1. doi:10.1038/455001a.

Csataljay, Gabor, Peter O'Donoghue, Mike Hughes, and Henriette Dancs. 2009. "Performance Indicators That Distinguish Winning and Losing Teams in Basketball." *International Journal of Performance Analysis in Sport* 9 (1): 60–66.

Cummins, Cloe, Rhonda Orr, Helen O'Connor, and Cameron West. 2013. "Global Positioning Systems (GPS) and Microtechnology Sensors in Team Sports: A Systematic Review." *Sports Medicine (Auckland, N.Z.)* 43 (10): 1025–42. doi:10.1007/s40279-013-0069-2.

Dawson, B., R. Hopkinson, B. Appleby, G. Stewart, and C. Roberts. 2004. "Player Movement Patterns and Game Activities in the Australian Football League." *Journal of Science and Medicine in Sport / Sports Medicine Australia* 7 (3): 278–91.

Deutsch, M. U., G. A. Kearney, and N. J. Rehrer. 2007. "Time – Motion Analysis of Professional Rugby Union Players during Match-Play." *Journal of Sports Sciences* 25 (4): 461–72. doi:10.1080/02640410600631298.

Di Salvo, V., W. Gregson, G. Atkinson, P. Tordoff, and B. Drust. 2009. "Analysis of High Intensity Activity in Premier League Soccer." *International Journal of Sports Medicine* 30 (3): 205–12. doi:10.1055/s-0028-1105950.

Duncan, Mitch J., Hannah M. Badland, and W. Kerry Mummery. 2009. "Applying GPS to Enhance Understanding of Transport-Related Physical Activity." *Journal of Science and Medicine in Sport*, Physical Activity in Young People: Assessment and Methodological Issues, 12 (5): 549–56. doi:10.1016/j.jsams.2008.10.010.

Dunn, and Coffe. 2013. "The Four V's of Big Marketing." *Coffee + Dunn*. September 25. http://www.coffee-dunn.com/the-four-vs-of-big-marketing/.

Ehrmann, Fabian E., Craig S. Duncan, Doungkamol Sindhusake, William N. Franzsen, and David A. Greene. 2016. "GPS and Injury Prevention in Professional Soccer." *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 30 (2): 360–67. doi:10.1519/JSC.0000000000001093.

Eom, H. J., and R. W. Schutz. 1992. "Transition Play in Team Performance of Volleyball: A Log-Linear Analysis." *Research Quarterly for Exercise and Sport* 63 (3): 261–69. doi:10.1080/02701367.1992.10608741.

Flach, Peter A. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge ; New York: Cambridge University Press.

Fleck, Steven J. 1999. "Periodized Strength Training: A Critical Review." *The Journal of Strength & Conditioning Research* 13 (1): 82–89.

Fujimura, Akira, and Kokichi Sugihara. 2005. "Geometric Analysis and Quantitative Evaluation of Sport Teamwork." *Systems and Computers in Japan* 36 (6): 49–58. doi:10.1002/scj.20254.

Gabbett, Tim J. 2004. “Reductions in Pre-Season Training Loads Reduce Training Injury Rates in Rugby League Players.” *British Journal of Sports Medicine* 38 (6): 743–49. doi:10.1136/bjsm.2003.008391.
Gabbett, Tim J. 2016. “The Training-Injury Prevention Paradox: Should Athletes Be Training Smarter and Harder?” *British Journal of Sports Medicine*, January, bjsports-2015-095788. doi:10.1136/bjsports-2015-095788.
Gabbett, Tim J., and Nathan Domrow. 2007. “Relationships between Training Load, Injury, and Fitness in Sub-Elite Collision Sport Athletes.” *Journal of Sports Sciences* 25 (13): 1507–19. doi:10.1080/02640410701215066.
Gabbett, Tim J., and David G. Jenkins. 2011. “Relationship between Training Load and Injury in Professional Rugby League Players.” *Journal of Science and Medicine in Sport* 14 (3): 204–9. doi:10.1016/j.jsams.2010.12.002.
Gabbett, Tim J., and Shahid Ullah. 2012. “Relationship between Running Loads and Soft-Tissue Injury in Elite Team Sport Athletes.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 26 (4): 953–60. doi:10.1519/JSC.0b013e3182302023.
Gamble, Paul. 2006. “Periodization of Training for Team Sports Athletes.” *Strength and Conditioning Journal* 28 (5): 56. doi:10.1519/1533-4295(2006)28[56:POTFTS]2.0.CO;2.
Gantz, J., and D. Reinsel. 2011. “Extracting Value from Chaos.” *IDC IView*, no. 1142: 1–12.
Garganta, Júlio. 2009. “Trends of Tactical Performance Analysis in Team Sports: Bridging the Gap between Research, Training and Competition.” *Revista Portuguesa de Ciências Do Desporto* 9 (1): 81–89.
Gaudino, P., F. M. Iaia, G. Alberti, A. J. Strudwick, G. Atkinson, and W. Gregson. 2013. “Monitoring Training in Elite Soccer Players: Systematic Bias between Running Speed and Metabolic Power Data.” *International Journal of Sports Medicine* 34 (11): 963–68. doi:10.1055/s-0033-1337943.
Goutte, Cyril, and Eric Gaussier. 2005. “A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation.” In *Advances in Information Retrieval*, 345–59. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-31865-1_25.
Gréhaigne, J. F., D. Bouthier, and B. David. 1997. “Dynamic-System Analysis of Opponent Relationships in Collective Actions in Soccer.” *Journal of Sports Sciences* 15 (2): 137–49. doi:10.1080/026404197367416.
Gudmundsson, Joachim, and Thomas Wolle. 2010. “Towards Automated Football Analysis: Algorithms and Data Structures.” In *Proc. 10th Australasian Conf. on Mathematics and Computers in Sport*. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.6492&rep=rep1&type=pdf.
Gyarmati, Laszlo, Haewoon Kwak, and Pablo Rodriguez. 2014. “Searching for a Unique Style in Soccer.” *arXiv Preprint arXiv:1409.0308*. http://arxiv.org/abs/1409.0308.
Hajian-Tilaki, Karimollah. 2013. “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.” *Caspian Journal of Internal Medicine* 4 (2): 627–35.
Häkkinen, K. 1993. “Changes in Physical Fitness Profile in Female Volleyball Players during the Competitive Season.” *The Journal of Sports Medicine and Physical Fitness* 33 (3): 223–32.
Halson, Shona L., and Asker E. Jeukendrup. 2004. “Does Overtraining Exist? An Analysis of Overreaching and Overtraining Research.” *Sports Medicine (Auckland, N.Z.)* 34 (14): 967–81.
Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. 2015. “The Rise of ‘big Data’ on Cloud Computing: Review and Open Research Issues.” *Information Systems* 47: 98–115. doi:10.1016/j.is.2014.07.006.
Haughton, Dominique, Mark-David McLaughlin, Kevin Mentzer, and Changan Zhang. 2015. *Movie Analytics*. SpringerBriefs in Statistics. Cham: Springer International Publishing. http://link.springer.com/10.1007/978-3-319-09426-7.
Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. 1 edition. Redmond, Washington: Microsoft Research.
Hilbert, Martin, and Priscila López. 2011. “The World’s Technological Capacity to Store, Communicate, and Compute Information.” *Science* 332 (6025): 60–65. doi:10.1126/science.1200970.
Hollinger, John. 2005. *Pro Basketball Forecast: 2005-2006*. Washington, D.C.: Potomac Books Inc.
Hughes, Mike, and Richard Daniel. 2003. “Playing Patterns of Elite and Non-Elite Volleyball.” *International Journal of Performance Analysis in Sport* 3 (1): 50–56.
Iaia, F. Marcello, Ermanno Rampinini, and Jens Bangsbo. 2009. “High-Intensity Training in Football.” *International Journal of Sports Physiology and Performance* 4 (3): 291–306.
Issurin, V. 2008. “Block Periodization versus Traditional Training Theory: A Review.” *The Journal of Sports Medicine and Physical Fitness* 48 (1): 65–75.

Issurin, Vladimir B. 2010. “New Horizons for the Methodology and Physiology of Training Periodization.” *Sports Medicine* 40 (3): 189–206.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. http://link.springer.com/10.1007/978-1-4614-7138-7.
Kilpatrick, David. 2011. “Inverting the Pyramid: A History of Football Tactics (Review).” *Journal of Sport History* 38 (3): 530–31.
Korkmaz, Murat, Bülent Kılıç, Fatih Çatıkkaş, and Ali Serdar Yücel. 2014. “Financial Dimension of Sports Injuries.” https://openaccess.firat.edu.tr/xmlui/handle/11508/8153.
Kraemer, William J., Duncan N. French, Nigel J. Paxton, Keijo Häkkinen, Jeff S. Volek, Wayne J. Sebastianelli, Margot Putukian, et al. 2004. “Changes in Exercise Performance and Hormonal Concentrations over a Big Ten Soccer Season in Starters and Nonstarters.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 18 (1): 121–28.
Kuhlman, Dave. 2009. *A Python Book: Beginning Python, Advanced Python, and Python Exercises*. Dave Kuhlman. http://www.davekuhlman.org/python_book_01.pdf.
Lames, Martin, and Tim McGarry. 2007. “On the Search for Reliable Performance Indicators in Game Sports.” *International Journal of Performance Analysis in Sport* 7 (1): 62–79.
Leung, Carson K., and Kyle W. Joseph. 2014. “Sports Data Mining: Predicting Results for the College Football Games.” *Procedia Computer Science* 35: 710–19. doi:10.1016/j.procs.2014.08.153.
Lowry, Cynthia A., William H. Woodall, Charles W. Champ, and Steven E. Rigdon. 1992. “A Multivariate Exponentially Weighted Moving Average Control Chart.” *Technometrics* 34 (1): 46–53. doi:10.1080/00401706.1992.10485232.
Lucas, James M., and Michael S. Saccucci. 1990. “Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements.” *Technometrics* 32 (1): 1–12. doi:10.1080/00401706.1990.10484583.
Malone, James J., Rocco Di Michele, Ryland Morgans, Darren Burgess, James P. Morton, and Barry Drust. 2015. “Seasonal Training-Load Quantification in Elite English Premier League Soccer Players.” *International Journal of Sports Physiology and Performance* 10 (4): 489–97. doi:10.1123/ijspp.2014-0352.
Manzi, Vincenzo, Stefano D’Ottavio, Franco M. Impellizzeri, Anis Chaouachi, Karim Chamari, and Carlo Castagna. 2010. “Profile of Weekly Training Load in Elite Male Professional Basketball Players.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 24 (5): 1399–1406. doi:10.1519/JSC.0b013e3181d7552a.
McGarry, Tim, David I. Anderson, Stephen A. Wallace, Mike D. Hughes, and Ian M. Franks. 2002. “Sport Competition as a Dynamical Self-Organizing System.” *Journal of Sports Sciences* 20 (10): 771–81. doi:10.1080/026404102320675620.
Newton, Robert U., Ryan A. Rogers, Jeff S. Volek, Keijo Häkkinen, and William J. Kraemer. 2006. “Four Weeks of Optimal Load Ballistic Resistance Training at the End of Season Attenuates Declining Jump Performance of Women Volleyball Players.” *Journal of Strength and Conditioning Research / National Strength & Conditioning Association* 20 (4): 955–61. doi:10.1519/R-5050502x.1.
Novatchkov, Hristo, and Arnold Baca. 2013. “Artificial Intelligence in Sports on the Example of Weight Training.” *Journal of Sports Science & Medicine* 12 (1): 27–37.
O’Leary, D. E. 2013. “Artificial Intelligence and Big Data.” *IEEE Intelligent Systems* 28 (2): 96–99. doi:10.1109/MIS.2013.39.
Orchard, John. 2012. “Who Is to Blame for All the Football Injuries?” *British Journal of Sport Medicine*. June 20. http://blogs.bmj.com/bjsm/2012/06/20/who-is-to-blame-for-all-the-football-injuries/.
Osgnach, Cristian, Stefano Poser, Riccardo Bernardini, Roberto Rinaldo, and Pietro Enrico di Prampero. 2010. “Energy Cost and Metabolic Power in Elite Soccer: A New Match Analysis Approach.” *Medicine and Science in Sports and Exercise* 42 (1): 170–78. doi:10.1249/MSS.0b013e3181ae5cfd.
Piggott, Benjamin. 2008. “The Relationship between Training Load and Incidence of Injury and Illness over a Pre-Season at an Australian Football League Club.” *Theses: Doctorates and Masters*, January. http://ro.ecu.edu.au/theses/25.
Pinder, Ross A., Keith W. Davids, Ian Renshaw, and Duarte Araujo. 2011. “Representative Learning Design and Functionality of Research and Practice in Sport.” *Journal of Sport and Exercise Psychology* 33 (1): 146–55.

Plisk, Steven S., and Michael H. Stone. 2003. “Periodization Strategies.” *Strength & Conditioning Journal* 25 (6): 19–37.
Portas, Matthew D., Jamie A. Harley, Christopher A. Barnes, and Christopher J. Rush. 2010. “The Validity and Reliability of 1-Hz and 5-Hz Global Positioning Systems for Linear, Multidirectional, and Soccer-Specific Activities.” *International Journal of Sports Physiology and Performance* 5 (4): 448–58.
Prampero, P. E. di, S. Fusi, L. Sepulcri, J. B. Morin, A. Belli, and G. Antonutto. 2005. “Sprint Running: A New Energetic Approach.” *The Journal of Experimental Biology* 208 (Pt 14): 2809–16. doi:10.1242/jeb.01700.
Reep, C., and B. Benjamin. 1968. “Skill and Chance in Association Football.” *Journal of the Royal Statistical Society. Series A (General)* 131 (4): 581–85. doi:10.2307/2343726.
Reep, C., R. Pollard, and B. Benjamin. 1971. “Skill and Chance in Ball Games.” *Journal of the Royal Statistical Society. Series A (General)* 134 (4): 623–29. doi:10.2307/2343657.
Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. “Challenges and Opportunities of Open Data in Ecology.” *Science* 331 (6018): 703–5. doi:10.1126/science.1197962.
Rogalski, Brent, Brian Dawson, Jarryd Heasman, and Tim J. Gabbett. 2013. “Training and Game Loads and Injury Risk in Elite Australian Footballers.” *Journal of Science and Medicine in Sport / Sports Medicine Australia* 16 (6): 499–503. doi:10.1016/j.jsams.2012.12.004.
Rygula, Igor. 2005. “Artificial Neural Networks as a Tool of Modeling of Training Loads.” *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 3: 2985–88. doi:10.1109/IEMBS.2005.1617101.
Schöllhorn, Wl. 2003. “Coordination Dynamics and Its Consequences on Sports.” *International Journal of Computer Science in Sport* 2 (2): 40–46.
Scott, Brendan R., Robert G. Lockie, Timothy J. Knight, Andrew C. Clark, and Xanne A. K. Janse de Jonge. 2013. “A Comparison of Methods to Quantify the in-Season Training Load of Professional Soccer Players.” *International Journal of Sports Physiology and Performance* 8 (2): 195–202.
Selye, H. 1946. “The General Adaptation Syndrome and the Diseases of Adaptation.” *The Journal of Clinical Endocrinology and Metabolism* 6 (February): 117–230. doi:10.1210/jcem-6-2-117.
Smith, Lloyd, Bret Lipscomb, and Adam Simkins. 2007. “Data Mining in Sports: Predicting Cy Young Award Winners.” *J. Comput. Sci. Coll.* 22 (4): 115–121.
Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. “‘ Big Data’: Big Gaps of Knowledge in the Field of Internet Science.” *International Journal of Internet Science* 7 (1): 1–5.
Spencer, Matt, Steven Lawrence, Claire Rechichi, David Bishop, Brian Dawson, and Carmel Goodman. 2004. “Time–motion Analysis of Elite Field Hockey, with Special Reference to Repeated-Sprint Activity.” *Journal of Sports Sciences* 22 (9): 843–50. doi:10.1080/02640410410001716715.
Taki, T., and J. Hasegawa. 2000. “Visualization of Dominant Region in Team Games and Its Application to Teamwork Analysis.” In *Computer Graphics International, 2000. Proceedings*, 227–35. doi:10.1109/CGI.2000.852338.
Talukder, Hisham, Thomas Vincent, Geoff Foster, Camden Hu, Juan Huerta, Aparna Kumar, Mark Malazarte, Diego Saldana, and Shawn Simpson. 2016. “Preventing in-Game Injuries for NBA Players.” In *MIT Sloan Analytics Conference*. Boston. http://www.sloansportsconference.com/content/preventing-in-game-injuries-for-nba-players/.
Tamura, Kohei, and Naoki Masuda. 2015. “Win-Stay Lose-Shift Strategy in Formation Changes in Football.” *EPJ Data Science* 4 (1). doi:10.1140/epjds/s13688-015-0045-1.
Tenga, Albin, Ingar Holme, Lars Tore Ronglan, and Roald Bahr. 2010. “Effect of Playing Tactics on Goal Scoring in Norwegian Professional Soccer.” *Journal of Sports Sciences* 28 (3): 237–44. doi:10.1080/02640410903502774.
Urhausen, Axel, and Wilfried Kindermann. 2002. “Diagnosis of Overtraining: What Tools Do We Have?” *Sports Medicine (Auckland, N.Z.)* 32 (2): 95–102.
Varley, Matthew C., Ian H. Fairweather, and Robert J. Aughey. 2012. “Validity and Reliability of GPS for Measuring Instantaneous Velocity during Acceleration, Deceleration, and Constant Motion.” *Journal of Sports Sciences* 30 (2): 121–27. doi:10.1080/02640414.2011.627941.
Waldron, Mark, Paul Worsfold, Craig Twist, and Kevin Lamb. 2011. “Concurrent Validity and Test-Retest Reliability of a Global Positioning System (GPS) and Timing Gates to Assess Sprint Performance Variables.” *Journal of Sports Sciences* 29 (15): 1613–19. doi:10.1080/02640414.2011.608703.

Williams, J. G., and R. G. Eston. 1989. “Determination of the Intensity Dimension in Vigorous Exercise Programmes with Particular Reference to the Use of the Rating of Perceived Exertion.” *Sports Medicine (Auckland, N.Z.)* 8 (3): 177–89.

Y, Hong, Robinson Pd, Chan Wk, Clark Cr, and Choi T. 1996. “Notational Analysis on Game Strategy Used by the World’s Top Male Squash Players in International Competition.” *Australian Journal of Science and Medicine in Sport* 28 (1): 18–23.