
Effective injury prediction in professional soccer with GPS data and machine learning

Alessio Rossi · Luca Pappalardo · Paolo
Cintia · F. Marcello Iaia · Javier
Fernández · Daniel Medina

Abstract Injuries have a great impact on professional soccer, due to their large influence on team performance and the considerable costs of rehabilitation for players. Existing studies in the literature provide just a preliminary understanding of which factors mostly affect injury risk, while an evaluation of the potential of statistical models in forecasting injuries is still missing. In this paper, we propose a multidimensional approach to injury prediction in professional soccer which is based on GPS measurements and machine learning. By using GPS tracking technology, we collect data describing the training workload of players in a professional soccer club during a season. We show that our injury predictors are both accurate and interpretable by providing a set of case studies of interest to soccer practitioners. Our approach opens a novel perspective on injury prevention, providing a set of simple and practical rules for evaluating and interpreting the complex relations between injury risk and training performance in professional soccer.

Keywords sports analytics · data science · machine learning · sports science · predictive analytics

1 Introduction

Injuries of professional athletes have a great impact on sports industry, due to their large influence on both the mental state and the performance of a team

L. Pappalardo and P. Cintia
Department of Computer Science, University of Pisa, Italy
E-mail: lpappalardo@di.unipi.it, paolo.cintia@isti.cnr.it

J. Fernandez and D. Medina
Sports Science and Health Department, Football Club Barcelona, Spain
E-mail: javier.fernandez@pl.fcbarcelona.cat, daniel.medina@fcbarcelona.cat

A. Rossi and F.M. Iaia
Department of Biomedical Science for Health, University of Milan, Italy
E-mail: alessio.rossi2@gmail.com, marcello.iaia@unimi.it

[1,2]. Furthermore, the costs associated with the complex process of recovery and rehabilitation for the player is often considerable, both in terms of medical care and missed earnings deriving from the popularity of the player himself [3]. Recent research demonstrates that injuries in Spain cause an average of 16% of season absence by players, corresponding to a total cost estimation of 188 million euros just in one season [4]. It is not surprising hence that injury prediction is attracting a growing interest from researchers, managers, and coaches, who are interested in intervening with appropriate actions to reduce the likelihood of injuries of their players.

Historically, academic work on injury prediction has been deterred for decades by the limited availability of data describing the physical activity of players during the season. Nowadays, the data revolution and the Internet of Things have the potential to change rapidly this scenario thanks to Electronic Performance and Tracking Systems (EPTS) [5,6], new tracking technologies that provide high-fidelity data streams, based on video recordings by different cameras or observations by various kinds of fixed and mobile sensors [5,7,8,9,10]. Professional soccer clubs are starting to use these new technologies to collect data from official games and training sessions, to ensure they can remain in control of their players' performance as much as possible. These soccer data depict in detail the movements of players on the field [5,6] and have been used for many purposes, such as understanding game performance [11], identifying training patterns [12], or performing automatic tactical analysis [5].

Despite this wealth of data, a little effort has been put on investigating injury prediction in professional soccer so far [13,14,15]. Existing studies in the literature provide just a preliminary understanding of which variables mostly affect injury risk [13,14,15], while an evaluation of the potential of statistical models in forecasting injuries is still missing. A major limit of existing studies is that they follow a monodimensional approach: since they use just one variable at a time to estimate injury risk, they do not fully exploit the complex patterns underlying the available data. Professional soccer clubs are interested in practical, usable and interpretable models as support in decision making to coaches and athletic trainers [16]. The construction of such injury prediction models poses many challenges. On one hand, an injury prediction model must be highly accurate. Predictors which rarely forecast injuries are useless for coaches, as well as predictors which frequently produce "false alarms", i.e., misclassify healthy players as risky ones. On the other hand a "black box" approach is less desirable for practical use since it does not provide insights about the reason behind injuries. It is fundamental to understand the complex relationships between the performance of players and their injury risk through simple, interpretable, and easy-to-use tools. An interpretable model reveals the influence of variables to injuries and the reasons behind them, allowing soccer practitioners to react on time by modifying properly training demands. Therefore, predictive models for injury prediction must achieve a good tradeoff between accuracy and interpretability.

In this paper, we propose a data-driven approach to predict the injuries in a professional soccer club and demonstrate that it is accurate and easy-to-

interpret by soccer practitioners. We consider injury prediction as the problem of forecasting that a player will get injured in the next training session or official game, given his recent training workload. According to UEFA model [17], a non-contact injury is defined as any tissue damage sustained by a player that causes absence in next sports activities for at least the day after the day of the onset. Our approach is based on automatic data collection through standard GPS sensing technologies, and it is intended as a supporting tool to the decision making of soccer managers and coaches. In the first stage of our study, we collect data about training workload of players through GPS devices, covering half of a season of a professional soccer club. After a pre-processing task, we extract from the data a set of features used in sports science to describe kinematic, metabolic and mechanical aspects of training workload, and we enrich them with information about all the injuries which happen during the half season. Based on these features, we implement benchmark predictors based on injury risk estimation methods commonly used in sports science, showing that they are not effective for injury forecasting due to their low precision. We hence move to a multidimensional approach and use machine learning to train a repertoire of decision tree classifiers by using the variables extracted. We find that injuries can be successfully predicted with a small set of three variables: the presence of recent previous injuries, high metabolic load distance and sudden decelerations. We then evaluate our decision tree and observe that it can correctly forecast 76% of the injuries recorded throughout the half season, achieving a 94% precision after around 16 weeks of training data. We also investigate a real-world scenario where the classifiers are updated while new training workload and injury data become available as the season goes by. The machine learning approach is still precise in this more challenging scenario, stabilizing its predictive performance at around half of the observed period. Finally, we investigate the structure of the best decision tree, extract an easy-to-use set of decision rules, highlight critical values of the considered workload features and discuss related case studies of interest to soccer practitioners. Our results are remarkable if we consider that we use data from training sessions only, since the usage of sensing technologies in games was not allowed by FIFA when our data collection was performed. Given that FIFA now authorizes the usage of tracking systems in games [18], our approach can be further improved by including game data, which are the highest load demand in a player's week.

Although our results refer to the situation of a specific professional club, it shows that automatic data collection and machine learning allow for the construction of injury prevention tools which are both accurate and interpretable, opening an interesting perspective for engineering supporting tools for analysis and prediction in professional soccer. Data-driven models like ours can be used to describe and “nowcast” the health of players throughout the season in a fast, cheap and reliable way.

The rest of the paper is organized as follows. Section 2 revises the scientific literature relevant to our topic, Section 3 describes the data collection and the feature extraction processes. In Section 4 we implement existing methodologies

for injury risk estimation and construct related predictors. In Section 5 we describe and evaluate our machine learning approach to injury prediction, comparing it with existing approaches. We interpret and discuss the results of our experiments in Section 6 and Section 7. Finally, Section 8 concludes the paper describing future improvements of our approach.

2 Related Work

The relationship between training workload, sports performance and injury risk has been recently studied in the sports science literature [19, 20, 21, 22, 23, 24, 25, 26, 27]. These studies observe that injuries related to training workload are to some extent preventable. For example Gabbett et al. [19, 20, 21, 22, 23, 24, 27] investigate the case of rugby and find that a player has a high risk of injury when his workloads are increased above certain thresholds. As a consequence, they conclude that balancing workload is a critical aspect of predicting injuries.

To assess injury risk in cricket, Hulin et al. [28] propose the ratio between acute workload (i.e., the average workload in the last 7 days) and chronic workload (i.e., the average workload in the last 28 days) using total week distance to measure training workload. They observe that acute chronic ratio values lower than 1 ($\text{acute} < \text{chronic}$) refer to cricket players in good physical fit and correspond to a low injury risk (4%). In contrast, acute chronic ratio values higher than 2 ($\text{acute} > \text{chronic}$) refer to players in a state of fatigue and correspond to injury risk from 2 to 4 times higher than the other group of players. Hence, the acute chronic ratio is useful to emphasize both positive and negative consequences of training. Murray et al. [29] show that in rugby acute chronic workload ratio is better related to injury risk when it is calculated using exponential smoothed moving average instead of the simple average. Hulin et al. [28] and Ehrmann et al. [14] find the same results for elite rugby players and soccer players respectively. They find that injured players, in both rugby and soccer, show significantly higher physical activity in the week preceding the injury with respect to their seasonal averages.

In skating, Foster et al. [30] measure training workload by the “session load”, i.e., the product of perceived exertion and duration of the training session. They find that when the session load outweighs a skater’s ability to fully recover before the next session, the skater suffers from the so-called “overtraining syndrome”, a condition that can cause injury [30]. In basketball, Anderson et al. [24] find a strong correlation between injury risk and the so-called “monotony”, i.e., the ratio between the mean and the standard deviation of the session load recorded in last 7 days. Moreover, Brink et al. [13] observe that injured young soccer players (age < 18) recorded higher values of monotony in the week preceding the injury than non-injured players.

Venturelli et al. [15] perform several periodic physical tests on young soccer players (age < 18) and measure body size, endurance, flexibility and jump height. They find that jump height, body size and the presence of previous injuries are significantly correlated with the probability of thigh strain injury.

Talukder et al. [31] apply a machine learning approach to predict injuries in basketball, based on performance features extracted from NBA games. In particular, they create a classifier able to predict 19% of injuries occurred in NBA. They also show that the most important features for predicting injuries are the average speed, the number of past competitions played, the average distance covered, the number of minutes played to date and the average field goals attempted. However, Talukder et al. [31] do not take into account the players' workload in training sessions.

Position of our work. In literature, the only studies addressing the problem of injury prediction in soccer are the ones by Brink et al. [13], Ehrmann et al. [14] and Venturelli et al. [15]. However, these studies suffer a major limitation: while they observe the existence of a correlation between specific aspects of training workload and the chance of injury, they do not construct any predictor as a practical tool for coaches and athletic trainers to prevent injuries. Therefore, to the best of our knowledge, there is no quantification of the potential of predictive analytics in preventing non-traumatic injuries in professional soccer. In this paper, we fill this gap by proposing a machine learning approach to injury prevention and show that outperforms existing injury risk estimation methods for professional soccer players.

3 Data collection and feature extraction

We set up a study on twenty-six professional players of a professional soccer club in Italian Serie B (age = 26 ± 4 years; height = 179 ± 5 cm; body mass = 78 ± 8 kg) during season 2013/2014. Six central backs, three fullbacks, seven midfielders, eight wingers and two forwards were recruited. Participants gave their written informed consent to participate in the study. The study was also approved by the Ethical Committee of the University of Milan.

We monitored the physical activity of players during 23 weeks of training sessions – from January 1st to May 31st – using portable 10 Hz GPS devices, integrated with 100 Hz 3-D accelerometer, a 3-D gyroscope, a 3-D digital compass (STATSports Viper). Despite some concerns over the reliability of GPS measurement of accelerations [32], especially at low sampling rates, it has been demonstrated that GPS traces are an useful tool for analyzing the activity profile in team sports, including soccer [33]. The devices were placed between the players' scapulae through a tight vest and were activated 15 minutes before the data collection, in accordance with the manufacturer's instructions to optimize the acquisition of satellite signals. We recorded a total of 954 individual training sessions corresponding to 80 collective training sessions performed during the 23 weeks. From the data collected by the devices, we extracted a set of training workload indicators through the software package Viper Version 2.1 provided by STATSports 2014. Figure 1a visualizes the GPS trace of a player during a training session in our dataset.

The club's medical staff recorded all the non-contact injuries occurred during 23 weeks. According to UEFA regulations [17], a non-contact injury is defined as any tissue damage sustained by a player that causes absence in next football activities for at least the day after the day of the onset. We observed 21 non-contact injuries during the period of observation, 19 of them associated with players who got injured at least once in the past. The distribution of injuries per player is provided in Figure 1. We observe that half of the players never get injured during the period of observation, while the others get injured once (seven players), twice (five players) or four times (one player). For every player, we also collected information about his age, weight, height and role on the field. Moreover, for every single training session of a player, we collected information about the play time in the official game before the training session and the number of official games played before the training session.

From the players' GPS data of every training session we select 12 features describing different aspects of workload [34]. Two features – Total Distance (d_{TOT}) and High Speed Running Distance (d_{HSR}) – are *kinematic*, i.e., they quantify the overall movement of a player during a training session. Three features – Metabolic Distance d_{MET} , High Metabolic Load Distance d_{HML} and High Metabolic Load Distance per minute d_{HML}/m – are *metabolic*, i.e., they quantify the energy expenditure of the overall movement of a player during a training session. The remaining seven features – Explosive Distance (d_{EXP}), number of accelerations above 2m/s^2 (Acc_2), number of accelerations above 3m/s^2 (Acc_3), number of decelerations above 2m/s^2 (Dec_2), number of decelerations above 3m/s^2 (Dec_3), Dynamic Stress Load (DSL) and Fatigue Index (FI) – are *mechanical* features describing the overall muscular-skeletal load of a player during a training session. Table 1 provides a description of the workload features, Appendix A provides descriptive statistics of the variables.

4 Classic approaches to injury risk estimation

In this section, we replicate the two most used state-of-the-art approaches to injury risk estimation that can be implemented with the workload features available in our study. In particular we consider the acute chronic workload ratio (ACWR) used in cricket, rugby and soccer [14, 28, 29] and the mean standard deviation workload ratio (MSWR) used in non-professional soccer [13, 24]. First, since these approaches have been initially developed for other sports or for non-professional players, we test their effectiveness on injury risk estimation for professional soccer players. Secondly, since these approaches aim just at observing the relations between training workload and injury risk without providing any predictive model, we transform them in predictors and assess their practical usability for soccer practitioners.

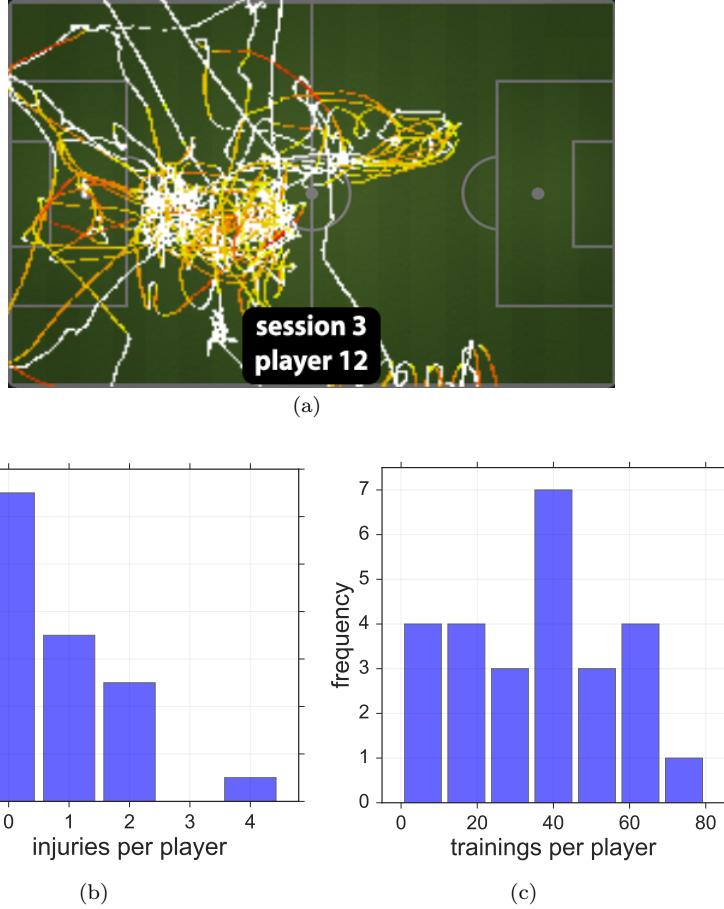


Fig. 1 (a) A visualization of the GPS trace of a player during a training session in our dataset. The color of the lines indicates, in a gradient from white to red, the speed of the player. (b) Distribution of the number of injuries per player. (c) Distribution of the number of training sessions per player.

4.1 Acute Chronic Workload Ratio (ACWR)

Many existing works on injury risk estimation rely on the so-called Acute Chronic Workload Ratio (ACWR) [14, 20, 22, 28, 29, 35], i.e., the ratio between the acute workload and the chronic workload of a player. Although its validity has been questioned by some authors [36, 37], ACWR is still one of the most used techniques in professional soccer clubs [38]. As proposed by Murray et al. [29], the acute workload of a player can be estimated by the exponential weighted moving average of the workload in the previous 7 days; the chronic workload of a player can be estimated as the exponential weighted moving average of the workload in the previous 28 days. ACWR can be calculated on any

Training workload features [34,39,40]

d_{TOT}	Distance in meters covered during the training session
d_{HSR}	Distance in meters covered above 5.5m/s
d_{MET}	Distance in meters covered at metabolic power
d_{HML}	Distance in meters covered by a player with a Metabolic Power is above 25.5W/Kg
d_{HML}/m	Average d_{HML} per minute
d_{EXP}	Distance in meters covered above 25.5W/Kg and below 19.8Km/h
Acc_2	Number of accelerations above 2m/s^2
Acc_3	Number of accelerations above 3m/s^2
Dec_2	Number of decelerations above 2m/s^2
Dec_3	Number of decelerations above 3m/s^2
DSL	Total of the weighted impacts of magnitude above 2g. Impacts are collisions and step impacts during running
FI	Ratio between DSL and speed intensity
Age	age of players
BMI	Body Mass Index: ratio between weight (in kg) and the square of height (in meters)
Role	Role of the player
PI	Number of injuries of the players before each training session
Play time	Minutes of play in previous games
Games	Number of games played before each training session

Table 1 Training workload features used in our study. Description of the training workload features extracted from GPS data and the players' personal features collected during the study. We defined four categories of features: kinematic features (blue), metabolic features (red), mechanical features (green) and personal features (white).

variable describing workload with a monodimensional approach: just one variable is considered at a time to estimate a player's injury risk [28,29]. ACWR values higher than 1 indicate that acute workload exceeds chronic workload, i.e., the average training workload in the last 7 days is much higher than the average workload in the previous 28 days, suggesting a training overload by the player. In contrast, ACWR values lower than 1 indicate an acute workload lower than the chronic workload.

Murray et al. [29] compute the ACWR for a set of workload features and categorize rugby players' training sessions in five groups: (1) $\text{ACWR} < 0.49$ (very low); (2) $\text{ACWR} \in [0.50, 0.99]$ (low); (3) $\text{ACWR} \in [1.00, 1.49]$ (moderate); (4) $\text{ACWR} \in [1.50, 1.99]$ (high); (5) $\text{ACWR} > 2.00$ (very high). Then injury likelihood (IL) is estimated in every ACWR group as the ratio between the number of players who get injured after the training session assigned to that ACWR group and the number of players who do not. Murray et al. [29] observe that players whose training sessions result in $\text{ACWR} > 2$ have a high injury risk (i.e., a high IL). We reproduce the ACWR methodology on our dataset for each of the 12 workload features described in Section 3, using the ACWR groups suggested by Murray et al. [29]. In contrast with literature,

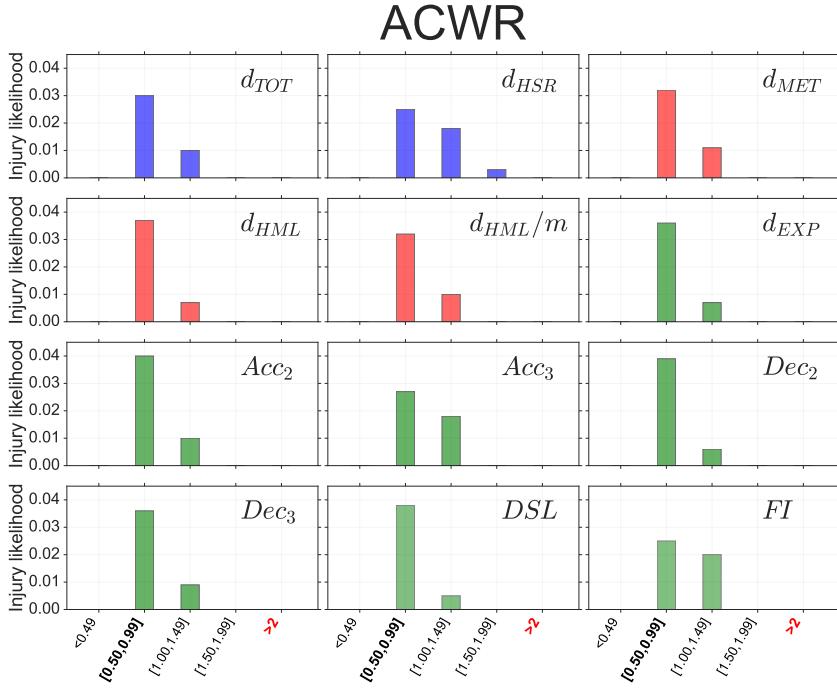


Fig. 2 Injury risk in ACWR groups. The plots show Injury Likelihood (IL) for pre-defined ACWR groups [29], for every of the 12 training workload features considered in our study. Bars are colored according to feature categorization defined in Table 3.

we do not find any individual training session resulting in $ACWR > 2$ (Figure 2), while we observe that players whose individual training sessions result in $ACWR < 1$ have the highest injury risk, i.e., the highest IL values (Figure 2). We replicate the experiment by using ACWR groups defined by the quintiles of the ACWR distribution instead of the pre-defined groups proposed by Murray et al. [29], finding similar results (see Appendix B).

ACWR predictor. The literature about the ACWR approach provides just a quantification of the relation between training workload and injury risk. We explore the usability in practice of this approach as a tool for injury prediction by constructing 12 predictive models, one for each of the 12 workload features introduced in Section 3 (Table 1). Given a player's training session, every predictive model C_h^{ACWR} predicts whether or not the player will get injured during next game or training session based on the value of workload feature h recorded in the current training session. If considering feature h the individual training session results in $ACWR < 1$, model C_h^{ACWR} predicts an injury (class 1) otherwise it predicts a non-injury (class 0). We validate the accuracy of predictive models against real data and observe that, in average, they have high recall (0.80 ± 0.08) and low precision (0.03 ± 0.003) on the

injury class (1): while most of the injuries (80%) are detected, in average the models wrongly predict an injury in 97% of the cases (see Table 7). We also explore the predictive power of the combination of the single predictors by constructing three combined predictive models. Predictive model $C_{\text{vote}}^{\text{ACWR}}$ predicts that a player will get injured if his training session results in $\text{ACWR} < 1$ for the majority of the workload features. Predictive model $C_{\text{all}}^{\text{ACWR}}$ predicts that a player will get injured if his training session results in $\text{ACWR} < 1$ for all the workload features. Model $C_{\text{one}}^{\text{ACWR}}$ predicts an injury if $\text{ACWR} < 1$ for at least one workload feature. Table 7 reports the accuracy of the three predictive models. Only model $C_{\text{vote}}^{\text{ACWR}}$ achieves a slightly better performance, in terms of precision on the injury class, w.r.t. models based on single features. We compare these predictive models with four baselines. Baseline B_1 randomly assigns a class to an example by respecting the distribution of classes. Baseline B_2 always assigns the majority class (i.e., class 0, a non-injury), while baseline B_3 always assigns the minority class (i.e., class 1, injury). Baseline B_4 is a classifier which assigns class 1 (injury) if the exponentially weighted average of variable $\text{PI} > 0$ (see Appendix D), and 0 (no injury) otherwise. Although the 15 predictive models constructed are significantly better than the baseline classifiers in terms of recall on the injury class, our results suggest that a predictor based on ACWR is not usable in practice due to its low precision. In a scenario where an athletic trainer bases his decisions on the suggestions of the predictors, in the vast majority of the cases he would generate “false alarms” by stopping a player with no risk of injury, which is not a practical solution to injury prevention in professional soccer.

4.2 Mean Standard deviation Workload Ratio (MSWR)

Another widely used approach to injury risk estimation is MSWR, defined in the literature as the ratio between the mean and the standard deviation of a player’s workload obtained in one week [24, 13, 30]. The higher the MSWR of a player, the lower is the variability of his workloads during the training week. In literature, it has been demonstrated that low MSWR values are associated with positive game performance while high MSWR values are associated with negative game performance and high injury risk [30]. Foster et al. [30] use this approach on a workload feature defined as the product of a player’s training duration and training workload as subjectively perceived by the players. In our study, we use objective workload as measured by the 12 features introduced in Section 3 and we compute MSWR for each of them. In detail, the MSWR value of a player at the end of a training session with respect to a workload feature h is the ratio between the mean and the standard deviation of the player’s distribution of feature h in the week preceding the individual training session. We investigate the relation between MSWR and injury risk by grouping the individual training sessions into quintiles according to the distribution of the workload features. For every quintile, we compute the corresponding injury likelihood (IL). We observe that high MSWR values are related to high injury

	ACWR				MSWR				
	class	prec	rec	F1	AUC	prec	rec	F1	AUC
$C_{d_{TOT}}$	0	0.99	0.44	0.61	0.65	0.98	0.80	0.88	0.57
	1	0.03	0.86	0.06		0.04	0.33	0.07	
$C_{d_{HSR}}$	0	0.99	0.37	0.54	0.57	0.98	1.00	0.99	0.50
	1	0.03	0.76	0.05		0.00	0.00	0.00	
$C_{d_{MET}}$	0	0.99	0.43	0.60	0.59	0.98	0.95	0.96	0.55
	1	0.03	0.76	0.06		0.06	0.14	0.09	
$C_{d_{HML}}$	0	0.99	0.43	0.60	0.60	0.98	0.96	0.97	0.53
	1	0.03	0.76	0.06		0.06	0.10	0.07	
$C_{d_{HML/m}}$	0	0.99	0.39	0.56	0.60	0.98	0.96	0.97	0.55
	1	0.03	0.81	0.06		0.08	0.14	0.10	
$C_{d_{EXP}}$	0	1.00	0.43	0.60	0.67	0.98	0.94	0.96	0.49
	1	0.04	0.91	0.07		0.02	0.05	0.03	
C_{Acc_2}	0	0.99	0.47	0.64	0.64	0.98	0.93	0.95	0.46
	1	0.03	0.80	0.06		0.00	0.00	0.00	
C_{Acc_3}	0	0.99	0.45	0.64	0.58	0.98	0.98	0.98	0.49
	1	0.03	0.71	0.06		0.00	0.00	0.00	
C_{Dec_2}	0	0.99	0.46	0.63	0.66	0.98	0.94	0.96	0.52
	1	0.04	0.86	0.07		0.04	0.10	0.05	
C_{Dec_3}	0	0.99	0.46	0.63	0.66	0.98	0.99	0.98	0.49
	1	0.04	0.86	0.07		0.00	0.00	0.00	
C_{DSL}	0	1.00	0.42	0.60	0.66	0.98	0.97	0.97	0.48
	1	0.03	0.90	0.07		0.00	0.00	0.00	
C_{FI}	0	0.98	0.47	0.64	0.55	0.98	0.72	0.83	0.50
	1	0.03	0.62	0.05		0.03	0.29	0.04	
C_{one}	0	1.00	0.09	0.17	0.54	0.98	0.56	0.71	0.54
	1	0.02	1.00	0.05		0.03	0.52	0.05	
C_{vote}	0	0.99	0.83	0.90	0.65	0.97	0.99	0.98	0.49
	1	0.06	0.48	0.11		0.00	0.00	0.00	
C_{all}	0	0.98	0.82	0.90	0.58	0.97	1.00	0.99	0.50
	1	0.04	0.33	0.07		0.00	0.00	0.00	
B_1	0	0.98	0.98	0.98	0.51	0.98	0.98	0.98	0.51
	1	0.06	0.05	0.05		0.06	0.05	0.05	
B_2	0	0.98	1.00	0.99	0.50	0.98	1.00	0.99	0.50
	1	0.00	0.00	0.00		0.00	0.00	0.00	
B_3	0	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50
	1	0.02	1.00	0.04		0.02	1.00	0.04	
B_4	0	0.98	0.77	0.86	0.60	0.98	0.77	0.86	0.60
	1	0.04	0.43	0.07		0.04	0.43	0.07	

Table 2 Performance of ACWR and MSWR predictors. We report precision (prec), recall (rec), F1-score (F1) and Area Under the Curve (AUC) for the injury class and the non-injury class for all the predictors based on ACWR and MSWR. We also provide predictive performance of four baseline predictors B_1 , B_2 , B_3 and B_4 .

risk for the majority of workload features, substantially confirming results observed in the literature by Foster et al. [30] (see Figure 3).

MSWR predictor. As done for ACWR, we explore the usability in practice of MSWR by constructing 12 predictive models based on the 12 training workload features introduced in Section 3. Given a player’s training session, every predictive model C_h^{MSWR} predicts whether or not the player will get injured

during next game or training session based on the value of workload feature h . If considering feature h the individual training session is associated with the MSWR group with the highest injury risk, model C_h^{MSWR} predicts an injury (class 1), otherwise it predicts a non-injury (class 0). We validate the accuracy of these predictive models against real data and observe that in average they have both low recall (mean is 0.10 ± 0.10) and low precision (mean is 0.03 ± 0.03) on the injury class: just 10% of the injuries are detected and in average the models wrongly predict an injury in 97% of the cases. As done for ACWR, we also construct three combined models – $C_{\text{vote}}^{\text{MSWR}}$, $C_{\text{all}}^{\text{MSWR}}$ and $C_{\text{one}}^{\text{MSWR}}$ – and observe that they have poor accuracy in detecting the injury class. In particular, the MSWR predictors have predictive power comparable to the ACWR predictors. Both methodologies, in conclusion, are not usable in practice due to their low precision.

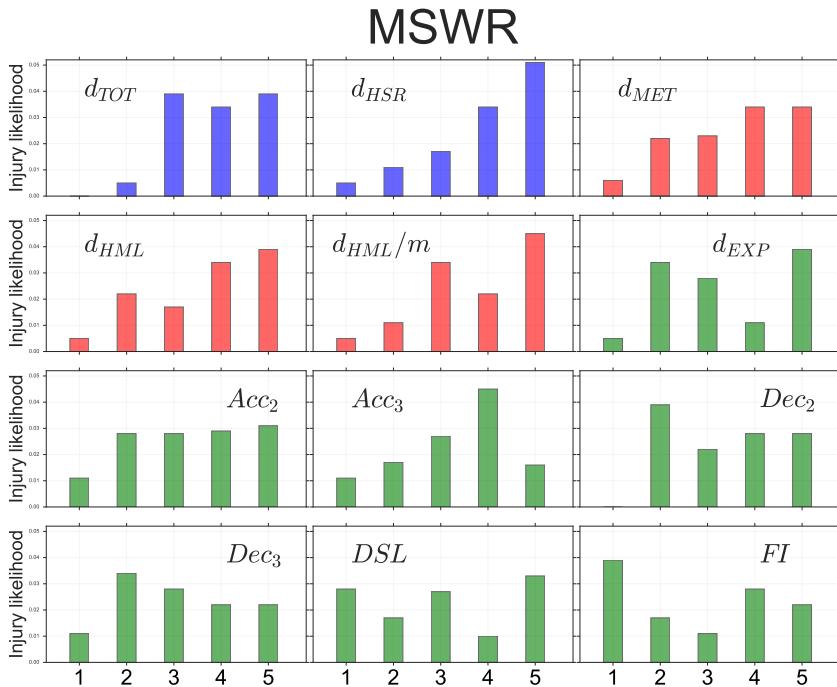


Fig. 3 Injury risk in MSWR groups. The plots show the Injury Likelihood (IL) for the MSWR groups for every of the 12 training workload features considered in our study. Bars are colored according to feature categorization defined in Table 1.

5 A machine learning approach to injury prediction

Given the inadequacy of the ACWR and MSWR predictors, we construct a repertoire of classifiers to predict whether or not a player will get injured based on his recent training workload. The construction of each classifier is made by using a training dataset where each example refers to a single player's training session and consists of: (i) a vector of the workload features describing the player's recent workload including the current training session; and (ii) the injury label, indicating whether or not the player got injured in next game or training session. We use a decision tree as predictive algorithm because it is easier to translate into practice and easy to understand even for people with a non-analytical background.

5.1 Construction of training dataset

Given a feature set S , the training dataset T_S for the learning task is constructed by a two-step procedure:

- (1) For every individual training session i we construct a feature vector $\mathbf{m}_i = (h_1, \dots, h_k)$ where $h_j \in S$, ($j=1, \dots, k$), is a training workload feature and $k = |S|$ is the number of features considered. All the feature vectors compose matrix $F_S = (\mathbf{m}_1, \dots, \mathbf{m}_n)$, where n is the number of individual training sessions in our dataset ($n = 954$);
- (2) Every feature vector \mathbf{m}_i is associated to a label $c_i \in \{0, 1\}$ which is 1 if the player gets injured in the next game or training session and 0 otherwise. Matrix F_S is hence associated to a vector of labels $\mathbf{c} = (c_1, \dots, c_n)$ (one for each training session). The training dataset for the learning task is finally $T_S = (F_S, \mathbf{c})$. Table 3 shows an example of training dataset.

	d_{TOT}	d_{EXP}	...	ACC_3	label
s_1	4,018.19	426.42	...	16.99	0
s_2	3,465.81	326.41	...	16.91	0
s_3	3,227.15	256.85	...	18.25	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
s_n	3,199.58	273.69	...	19.64	1

Table 3 Structure of the training dataset used for injury prediction. Every example describes a player's training session and consists of k features and a label $c \in \{0, 1\}$.

We construct four different training datasets based on four different feature sets built starting from the workload features in Table 1. Feature set WF consists of the exponential weighted moving average (EWMA) of the 12 workload

features in Table 1. We use EWMA since, as suggested in literature [29,31], a player’s injury risk depends on both the workload in the current training session and the workload in the recent training sessions. In detail, we consider a span equal to six in the EWMA and adjust every workload feature using EWMA. This choice is data-driven and justified by our experiments where we find that the best classification results are achieved when using a span equal to six (see Appendix C). Feature set ACWR uses the ACWR version of the 12 workload features in Table 1: given a player’s training session, for every feature we compute its acute chronic workload ratio as described in Section 4.1. Similarly, feature set MSWR uses the MSWR version of the 12 workload features in Table 1: given a player’s training session, for every feature we compute the exponential weighted moving standard deviation of the six most recent training sessions of the player. To take into account both the number of a player’s previous injuries and their distance to the current training sessions we compute $\text{PI}^{(\text{WF})}$, the EWMA of feature PI with a span equal to 6. $\text{PI}^{(\text{WF})}$ reflects the distance between the current training session and the training session when the player returned to regular training after an injury. $\text{PI}^{(\text{WF})} = 0$ indicates that the player never got injured in the past; $\text{PI}^{(\text{WF})} > 0$ indicates that the player got injured at least once in the past; $\text{PI}^{(\text{WF})} > 1$ indicate that the player got injured more than once in the past (see Appendix D). Finally, feature set ALL contains 42 features and it is the union of the three feature sets described above, plus the personal variables in Table 1 and $\text{PI}^{(\text{WF})}$. Every training set consists of 954 examples (i.e., individual training sessions) corresponding to 80 collective training sessions.

Example 1 (Construction of training dataset) Let us consider a toy dataset of a portion of the training sessions of a player $D = \{s_6, s_7, s_8, s_9\}$, where the last one is associated with an injury, i.e., the player will get injured during training session s_{10} . We construct the training dataset T_{ALL} based on feature set ALL as follows:

- (1) We create a new example in dataset T_{ALL} for every of the four training session in D , by computing the EWMA of the 42 player’s workload features in feature set ALL. Every example is described by a vector of length 42, $\mathbf{m}_i = (h_1, \dots, h_{42})$. All the four vectors compose matrix $F_{\text{ALL}} = (\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4)$;
- (2) Since the first three training sessions are not associated with injuries, the first three examples have label 0. In contrast, the last example has label 1 since it is associated with an injury. Therefore, the labels vector is hence $\mathbf{c} = (0, 0, 0, 1)$, indicating that the first three examples are not associated with an injury while the last training session produces an injury. The training dataset based on D and feature set ALL is finally $T_{\text{ALL}} = (F_{\text{ALL}}, \mathbf{c})$.

5.2 Experiments

For every feature set – WF, ACWR, MSWR, ALL – we construct the corresponding training datasets as shown in Section 5.1 and we perform two tasks. First, we

perform a feature selection process to determine the most relevant features for classification. The feature selection process has two motivations: *(i)* it aims at reducing the dimensionality of the feature space and the risk of overfitting; *(ii)* it allows for an easier interpretation of the machine learning models, due to the lower number of features [41]. The feature selection task is especially important for feature set ALL because it has 42 variables, a large number considering the small size of the dataset. We use recursive feature elimination with cross-validation (RFECV) [42] to select the best set of features.¹ RFECV is a wrapper method for feature selection [43] which initially starts by training a predictive model (a decision tree in our experiments) with all the features in the feature set. Then at every step, RFECV eliminates one feature, trains the decision tree on the reduced feature set and calculates the score on the validation data. The subset of features producing the maximum score on the validation data is considered to be the best feature subset [42]. On the new training dataset derived from the feature selection, we train a Decision Tree classifier (DT).² We choose DT because it provides a set of decision rules that are easy to understand. As already stated we are not interested in a “black box” approach since soccer practitioners are also interested in understanding why injury happened and what predictors are associated with it.

We hence construct four classification models C_{WF} , C_{ACWR} , C_{MSWR} and C_{ALL} . We validate the classifiers with a 3-fold stratified cross-validation strategy [44]: the real dataset is divided into 3 parts or folds and, for each fold, we use the 10% of the target values as test set, and the remaining 90% as training set. Each fold is made by preserving the percentage of samples for each class. Thus, each sample in the dataset was tested once, using a model that was not fitted with that sample. We also try cross-validation with 5 and 10 folds without finding significant differences in the results presented in this paper. We measure the goodness of the classifiers by four metrics:

- (1) **precision:** $prec = TP / (TP + FP)$, where TP and FP are true positives and false positives resulting from classification, where we consider the injury class (1) as the positive class. Given a class, precision indicates the fraction of examples that the classifier correctly classifies over the number of all examples the classifier assigns to that class;
- (2) **recall:** $rec = TP / (TP + FN)$, where FN are false negatives resulting from classification. Given a class, the recall indicates the ratio of examples of a given class correctly classified by the classifier;
- (3) **F-measure:** $F = 2(prec \times rec) / (prec + rec)$. This measure is the harmonic mean of precision and recall, which in our case coincides with the square of the geometric mean divided by the arithmetic mean;
- (4) **Area Under the Curve (AUC):** the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming “positive” ranks higher than “negative”). An AUC

¹ We use the RFECV function provided by the publicly available Python package scikit-learn (<http://scikit-learn.org/>).

² We use the Python package scikit-learn to train and validate the decision tree.

close to 1 represents an accurate classification, while an AUC close to 0.5 represents a classification close to randomness.

For injury prevention purposes, we are interested in achieving high values of precision and recall on class 1 (injury). Let us assume that a coach or an athletic trainer makes a decision about whether or not to “stop” a player based on the suggestion of the classifier, i.e., to prevent an injury, the players will skip next training session or game every time the classifier prediction associated with the player’s current training session is 1 (injury). In this scenario, precision indicates how much we can trust the classifier’s predictions: the higher the precision, the more a classifier’s predictions are reliable, i.e., the probability that the player will actually get injured is high. Trusting a classifier with low precision means producing many false positives (i.e., “false alarms”) and frequently stop players unnecessarily, a condition clubs want to avoid especially for the team’s key players.

Recall indicates the fraction of injuries the classifier detects over the total number of injuries: the higher the recall the more injuries the classifier can detect. A classifier with low recall detects just a small fraction of all the injuries, meaning that many players will attend next training session or game and actually get injured. In other words, a classifier with low recall would misclassify many actual injuries as non-injuries, i.e., in many cases, the classifier would classify a risky player as a healthy player.

5.3 Results

The feature selection task selects a small subset of the available features (see Figure 4a). Although just a small subset of features is selected, we observe that they cover all the four categories introduced in Table 1 – kinematic, metabolic, mechanical and personal – indicating that all these aspects are related to injury risk. In particular, two features are selected in feature sets WF, ACWR and MSWR, while three features are selected in feature set ALL (see Table 9). Figure 4a also shows the importance of the features in the classifiers, computed as the mean decrease in Gini coefficient, a measure of how each variable contributes to the homogeneity of nodes and leaves in the resulting decision tree [45]. Feature PI is the only one that is selected in every feature set, accounting for a relative importance ≥ 0.75 . Such a high importance of PI suggests that in our dataset a player’s injury risk increases if he already experienced injuries in the past. All the other personal features (i.e., age, BMI, role, play time and games) are discarded by the feature selection task.³

Table 9 shows the performance of classification resulting from our experiments: C_{ALL} is the best classifier in terms of both precision and recall, being

³ We observe that the feature selection task improves our predictive performance, since the decision trees trained on the entire sets of features (AUC=0.76) produces worse results than the decision trees trained on the reduced sets of features (AUC=0.88). In this paper we hence present results with feature selection only for the sake of brevity.

able to detect 76% of the injuries and achieving a 94% precision on the injury class (Figure 4b).⁴ We observe that all the decision trees constructed on the four feature sets provide a significant improvement in the prediction with respect to both the baselines B_1, \dots, B_4 and the ACWR and MSWR predictors, which have a precision close to zero (Table 7).⁵ The strikingly good performance of C_{ALL} indicates that: (i) the classifier is reliable due to its high precision, reducing false alarms and hence scenarios where players are “stopped” unnecessarily before games or training sessions; (ii) the classifier can detect more than 2/3 of the injuries occurred in the season, which can lead to a significant saving of economic resources by the club. Decision tree C_{WF} interestingly achieves the best precision ($\text{prec} = 1.0$), paying the cost for a low recall $\text{rec} = 0.33$. This classifier could be preferred to C_{ALL} if the club aims at having the absolute certainty that player stoppings are actually necessary to avoid their future injuries.

We investigate how long it takes for the classifiers – C_{ALL} , C_{WF} , C_{MSWR} and C_{ACWR} – to stabilize their predictive performance by training and evaluating them after every training week w_i . In other words, at week w_i we use all the training sessions up to week w_i , validating every classifier on the same data through 3-fold stratified cross-validation. Figure 4c shows the evolution of the cross-validated F1-score of the five classifiers compared to baseline B_4 . We observe that the classifiers’ performance significantly improve as new data become available during the season, stabilizing after 16 weeks (Figure 4c). In contrast, the performance of baseline B_4 is constant over time (F1-score ≈ 0.1).

5.4 Evolutive scenario of injury prevention

Here we investigate a scenario where we assume that, at the beginning of a season, no data are available to a soccer club who aims to use our approach. The soccer club equips with appropriate sensor technologies and starts recording training workload data since the first training session of the season. Assuming that the club train the classifier with new data after every training week, how many injuries the club can actually forecast? To answer this question we investigate the evolution of the classifier’s performance during the season, and how many of the injuries the classifier can actually forecast.

In the evolutive scenario, we proceed from the least recent training week, w_1 , to the most recent one, w_{n-1} . At training week w_i we train the five classifiers – C_{ALL} , C_{WF} , C_{MSWR} , C_{ACWR} , and B_4 – on data extracted from sessions w_1, \dots, w_i and evaluate their accuracy as the ability in predicting injuries in week w_{i+1} . Figure 5 shows the evolution of the F1-score of the five classifiers as the season goes by, where sessions are grouped by week for visualization

⁴ the standard deviations in the 3 folds of precision and recall on the injury class are 0.14 and 0.06, respectively.

⁵ We also implement a Random Forest Classifier and observe that it achieves slightly worse performance than DT (see Appendix section E).

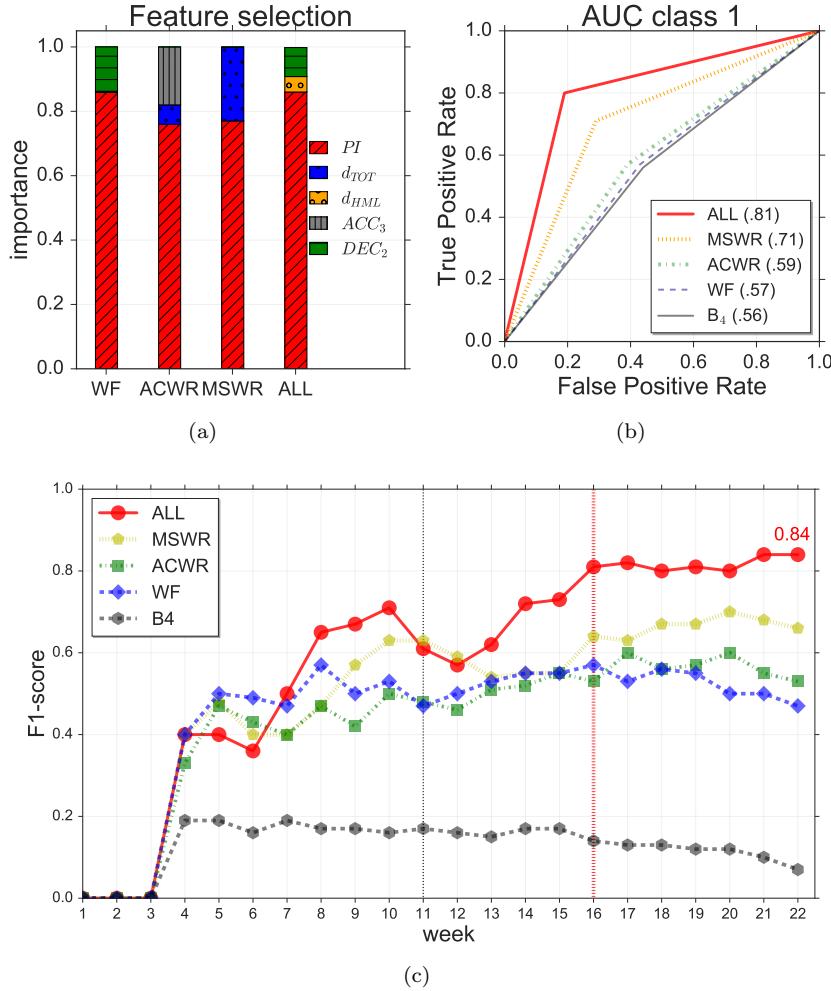


Fig. 4 (a) Relative importance of the features selected by the feature selection task, for every decision tree. (b) AUC curves of decision trees and baselines for class 1 (injury). (c) F1-score of classifiers as more data become available during the half season.

purposes. Due to the scarcity of data, the classifiers have a poor predictive performance at the beginning of the season and miss many injuries (the black crosses in Figure 5). However, predictive ability improves significantly by time and classifiers predict most of the injuries in the second half of the season (the red crosses in Figure 5). The best classifier, C_{ALL} , detects in this scenario 11 injuries out of 21, resulting in a cumulative F-score= 0.45 (Figure 5). The other classifiers – C_{WF} , C_{MSWR} and C_{ACWR} – follow a similar pattern as C_{ALL} ,

	model	class	prec	rec	F1	AUC	selected features
decision trees	C_{ALL}	0	0.98	1.00	0.99	0.88	$\text{PI}^{(\text{WF})}, d_{\text{HML}}^{(\text{MSWR})}, \text{Dec}_2^{(\text{WF})}$
		1	0.94	0.76	0.84		
	C_{MSWR}	0	0.99	0.99	0.99	0.78	$\text{PI}, d_{\text{TOT}}$
		1	0.80	0.57	0.66		
	C_{CACWR}	0	0.99	1.00	0.99	0.71	$\text{PI}, d_{\text{TOT}}, \text{Acc}_3$
		1	0.69	0.43	0.53		
	C_{WF}	0	0.98	1.00	0.99	0.64	PI, Dec_2
		1	1.00	0.33	0.47		
ACWR	$C_{\text{vote}}^{\text{ACWR}}$	0	0.99	0.83	0.90	0.65	12 workload features in Table 1
baselines	B_4	0	0.98	0.77	0.86	0.60	
		1	0.04	0.43	0.07		
	B_1	0	0.98	0.98	0.98	0.51	
		1	0.06	0.05	0.05		
B_2	0	0.98	1.00	0.99	0.50		
	1	0.00	0.00	0.00			
B_3	0	0.00	0.00	0.00	0.50		
	1	0.02	1.00	0.04			

Table 4 Performance of classifiers compared to baselines. We report the performance of classifiers C_{ALL} , C_{WF} , C_{MSWR} , C_{CACWR} in terms of precision, recall, F1 and AUC. We also report for every classifier the features selected by the feature selection task. We compare the classifier with four baseline B_1, \dots, B_4 and predictor $C_{\text{vote}}^{\text{ACWR}}$.

in contrast with B_4 which does not improve predictive performance as the season goes by.

6 Interpretation of injury prediction model

Figure 6 is a schematic visualization of the best decision tree resulting from our experiments, C_{ALL} . In the tree there are two types of node: decision nodes (black boxes) and leaf nodes (green and red boxes). Every decision node has two branches each indicating the next node to select in the tree depending on the range of values of the feature associated with the decision node. The next node to select can be another decision node or directly a leaf node. A leaf node represents the final prediction of the classifier based on the feature vector describing a player’s individual training session. There are two possible final decisions: Injury (red boxes) indicates that the player will get injured in next game or training session; No-Injury (green boxes) indicates that the player has no risk to get injury in next game or training session. Given a feature vector s_i describing a player’s training session, the prediction associated with s_i can be obtained by following the path from the root of the tree down to a leaf node, through decision nodes. For example, given a player’s training session with $\text{PI}^{(\text{WF})} = 0.15$, $d_{\text{HML}}^{(\text{MSWR})} > 1.80$ and $\text{DEC}_2^{(\text{WF})} = 50.20$, the decision path

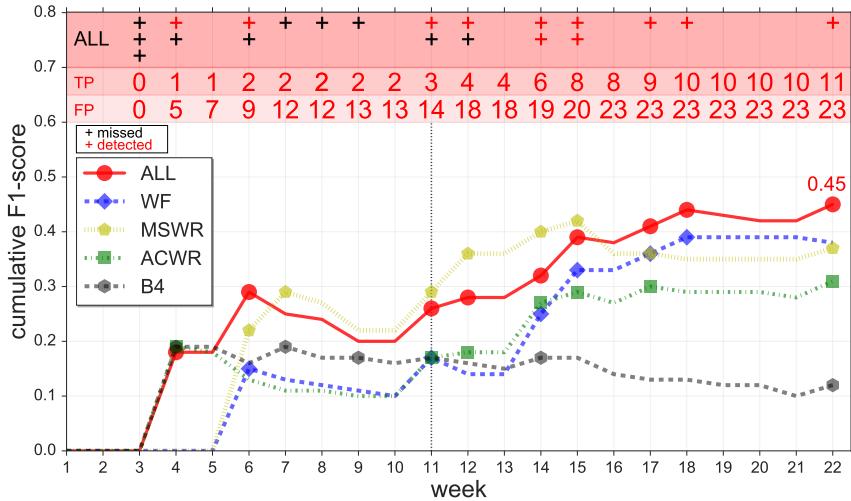


Fig. 5 Performance of classifiers in the evolutive scenario. As the season goes by, we plot week by week the F-score of the classifiers C_{ALL} , C_{WF} , C_{MSWR} , C_{ACWR} and B_4 trained on the data collected up to that week. Black crosses indicate injuries not detect by C_{ALL} , red crosses indicate injuries correctly predicted by C_{ALL} . For every week we highlight in red the number of injuries detected by C_{ALL} up to that week.

associated with this example is:

$$\boxed{\text{PI}^{(\text{WF})} > 0.14} \rightarrow \boxed{\text{PI}^{(\text{WF})} > 0.40} \rightarrow \boxed{d_{\text{HML}}^{(\text{MSWR})} \leq 1.36} \rightarrow \boxed{d_{\text{HML}}^{(\text{MSWR})} > 1.35} \rightarrow \boxed{\text{Injury}}$$

One of the most useful characteristics of decision trees is that they allow for the extraction of decision rules which summarize the reasoning of the classifier when taking specific decisions. In particular, C_{ALL} has 6 decision paths associated with injuries. We highlight in red these paths in Figure 6a and report them separately in Figure 6c. The structure of these rules provides us with useful insights about the reason behind injuries, such as the physical state of players and their training patterns before the injury. We extract three main injury scenarios which can be helpful to soccer practitioners to understand the reasons behind injuries in the considered club:

Previous injuries can lead to new ones.

The six rules derived from C_{ALL} highlight that the presence of previous injuries is crucial to determine whether or not a player will get injured in the near future. Indeed, feature $\text{PI}^{(\text{WF})}$ is present in all the six rules of Figure 6c, depicting three different scenarios:

1. a player can get injured immediately after its return to regular training after an injury (Rule 1: $\text{PI}^{(\text{WF})} \in (0.14, 0.40]$);
2. a player can get injured three days after its return to regular training after an injury (Rules 2, 4, 5 and 6: $\text{PI}^{(\text{WF})} > 0.40$);

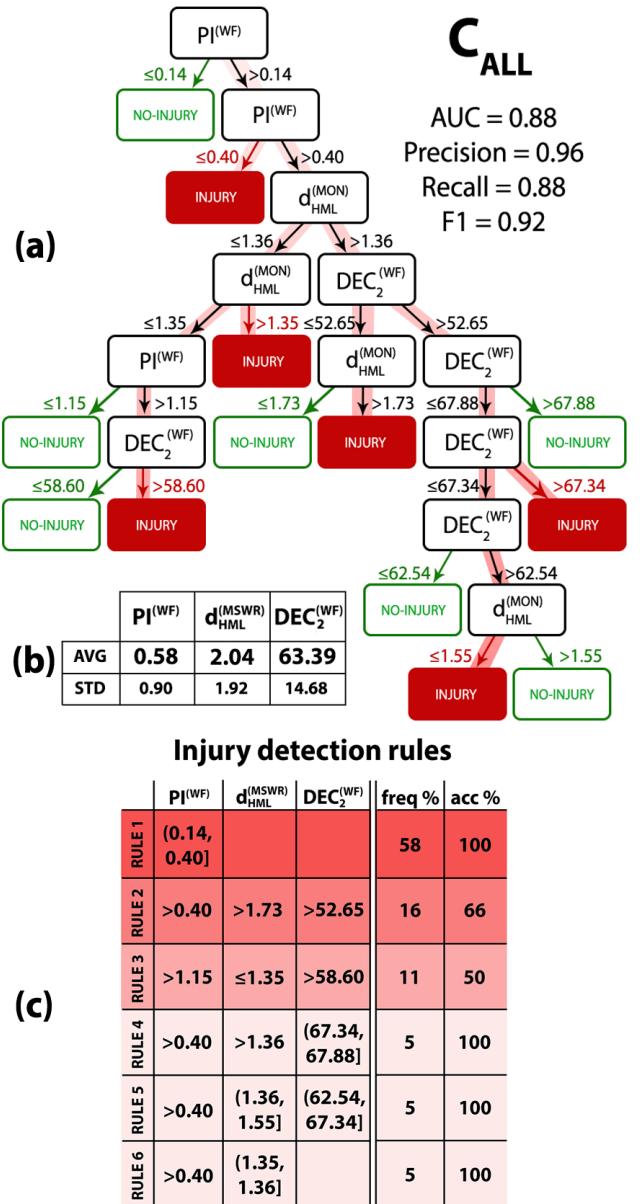


Fig. 6 Interpretation of injury prediction model. (a) A schematic visualization of decision tree C_{ALL}. Black boxes are decision nodes, green boxes are leaf nodes for class No-Injury, red boxes are leaf nodes for class Injury. (b) Mean (AVG) and standard deviation (STD) of the tree features used in the DT. (c) The six injury rules extracted from C_{ALL}. For every rule we show the range of values of every feature, the frequency and the accuracy of the rule.

3. a player can get injured the day after its return to regular training after two previous injuries (Rule 3: $\text{PI}^{(\text{WF})} > 1.15$).

From this scenario, it clearly emerges that, in order to prevent future injuries, players who return from a recent injury must be observed with particular attention by the club's athletic trainers. In particular, the first three days of training after an injury are the most critical, since 58% of injuries occur immediately after a player's return to regular training (Rule 1).

Special attention to high metabolic workload.

Among the players who got injured once in the past, *metabolic workload* $d_{\text{HML}}^{(\text{MSWR})}$ represents a critical value for subsequent injuries. When it is higher than the seasonal average ($d_{\text{HML}}^{(\text{MSWR})} > 1.73$, Rule 2), or close to a value in the 50th percentile range ($d_{\text{HML}}^{(\text{MSWR})} \in (1.35, 1.55+)$, Rules 4-6), metabolic workload is predictive of 31% of the injuries (Rules 2, 4, 5 and 6). A high metabolic workload can be risky for players that got injured in the past. This is true for players who have incurred in at least two injuries during the season ($\text{PI}^{(\text{WF})} > 1.15$).

High mechanical workload is related to injury risk.

In some cases, *mechanical workload* ($DEC_2^{(\text{WF})}$) is also predictive of injuries. In particular, $DEC_2^{(\text{WF})}$ has impact on predicting the injuries of players with a previous injury ($DEC_2^{(\text{WF})} > 52.65$ or $DEC_2^{(\text{WF})} \in (62.54, 67.88)$, Rules 2-5) or two previous injuries ($DEC_2^{(\text{WF})} > 58.60$, Rule 3). A high mechanical workload higher or close to the seasonal average of $DEC_2^{(\text{WF})}$, in conjunction with a high metabolic workload (Rules 2-5), identify 37% of the injuries. Such a result highlights the impact that high stress has, in terms of mechanical workload, on players who got injured in the past.

7 Discussion of results

The experiments on our dataset produce three interesting results. First, the best decision tree, C_{ALL} , can detect 76% of the injuries with 94% precision, far better than ACWR, MSWR and the baselines (Table 9). The decision tree's false positive rate is indeed small ($\text{FPR}=0.01$, Table 5), indicating that it infrequently produces "false alarms", i.e., situations where the classifier predicts that an injury will happen when it will not. In professional soccer, false alarms are despicable because the scarcity of players can negatively affect the performance of a team, both in single games and during the season [2]. C_{ALL} produces false alarms in less than 1% of the cases, while ACWR, MSWR, and the baselines produce false alarms in $>20\%$ of the cases. C_{ALL} also produces a moderate false negative rate ($\text{FNR}=0.24$, Table 5), meaning that situations where a player that will get injured is classified as out of risk are infrequent (Table 5). Although in general ACWR, MSWR and the baselines have a false negative rate lower or comparable to the decision trees (Table 5), they pay the price in precision producing a high number of false alarms. C_{ALL} achieves the

best trade-off between precision and recall indicating that, in contrast with existing approaches in sports science, a machine learning approach can forecast most of the injuries while still producing few false alarms.

		prediction				
		1	0			
C_{ALL}	true	1	16	5	TPR=0.76	FNR=0.24
		0	1	902	FPR=0.01	TNR=0.99
$C_{d_{EXP}}^{MSWR}$	true	1	19	2	TPR=0.90	FNR=0.10
		0	178	725	FPR=0.20	TNR=0.80
$C_{d_{TOT}}^{ACWR}$	true	1	19	2	TPR=0.90	FNR=0.10
		0	514	389	FPR=0.57	TNR=0.43
B_4	true	1	9	12	TPR=0.43	FNR=0.67
		0	217	686	FPR=0.24	TNR=0.76

Table 5 Confusion matrix of classification. Confusion matrix for C_{ALL} , the best predictors for both MSWR and ACWR approaches ($C_{d_{EXP}}$ and $C_{d_{TOT}}$, respectively) and baseline B_4 . TNR, TPR, FNR and FPR stays for True Negative Rate, True Positive Rate, False Negative Rate and False Positive Rate, respectively.

Second, in an evolutive scenario where a soccer club starts collecting the data for the first time and updates the classifiers after every session as the season goes by, at the end of the season C_{ALL} results in a cumulative F1-score=0.45 on the injury class (Figure 4):⁶ while it predicts more than half of the injuries throughout the season, it also produces many false alarms. The cumulative performance of the classifiers is highly affected by the initial period, where both training data and injury data are scarce. This suggests that trying to prevent injuries since the beginning could not be a good strategy since classification performance can be initially poor due to data scarcity. An initial period of data collection is needed in order to collect the adequate amount of data, and only then reliable classifiers can be trained on the collected data. The length of the data collection period clearly depends on the needs and the strategy of the club, including the frequency of training and games, the frequency of injuries, the number of available players and the tolerated level of false alarms. Regarding this aspect, in our dataset, we observe that the performance of the classifiers stabilizes after 16 weeks of data collection (Figure 4c). Moreover, in Figure 7 we investigate how many weeks of training we need to achieve a certain level of precision and recall on the injury class. We

⁶ cumulative precision and cumulative recall on the injury class at the end of the season are $prec = 0.40$ and $rec = 0.50$, respectively.

observe that at least 7 weeks of training are needed to achieve a precision on the injury class higher than 0.5, while at least 16 weeks of training are needed for a precision higher than 0.8. In our dataset a reasonable strategy could be to use the classifiers for injury prevention starting from the 16th week. This suggests that the considered club can effectively use the classifiers trained on data from a season to perform injury prediction since the first session of the next season.

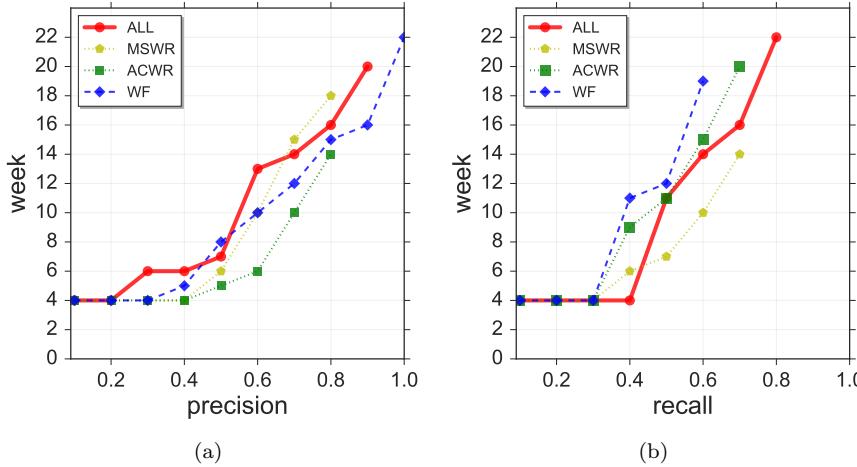


Fig. 7 Weeks of training vs precision and recall of the classifiers. (a) Number of weeks of training needed to achieve a given precision. (b) Number of weeks of training needed to achieve a given recall.

Third remarkable result is that, of 42 features, just 3 features are selected by the feature selection task for C_{ALL} : $\text{PI}^{(\text{WF})}$, $d_{\text{HML}}^{(\text{MSWR})}$ and $DEC_2^{(\text{WF})}$. Feature $\text{PI}^{(\text{WF})}$, the most important among the three, reflects the temporal distance between a player's current training session and the coming back to regular training of a player who got injured in the past. $\text{PI}^{(\text{WF})} > 0$ indicates that the player has incurred in at least one injury in the past, while $\text{PI}^{(\text{WF})} > 1$ indicates that the player injured at least twice in the past (see Appendix D). We remind that our dataset cover the second half of a season and just two injuries refer to players who never got injured before. Presumably, this is the reason why C_{ALL} can detect only injuries happened after a previous one ($\text{PI}^{(\text{WF})} > 0$). More than half of the injuries detected by the classifier (58%) happened immediately after the coming back to regular training of players who got injured (Rule 1, Figure 6c). Furthermore, 42% of the injuries detected by the classifier happened long after a previous one and are characterized by specific values of metabolic and mechanical features, $d_{\text{HML}}^{(\text{MSWR})}$ and $DEC_2^{(\text{WF})}$, which indicate the metabolic workload variability and the weighted mean of DEC_2 in the previous 6 days (Rules 2-6, Figure 6c). Rule 2, 4 and 5 describe injuries when

players record low levels of workload variability (i.e., high value of $d_{\text{HML}}^{(\text{MSWR})}$) and values higher or equal than the seasonal average of $DEC_2^{(\text{WF})}$. Rule 3 describes injuries when a player has a high value of $DEC_2^{(\text{WF})}$ in the last week and high variability in metabolic features (i.e., low value of $d_{\text{HML}}^{(\text{MSWR})}$), more than two day after the second injury. Finally, Rules 6 detects injuries in day long after previous injuries when the variability is on a specific values (i.e., $d_{\text{HML}}^{(\text{MSWR})} \in (1.35, 1.36]$).

Our results suggest that the professional soccer club should take care of the first training sessions of players which come back to regular training after an injury since in this conditions they are more likely to get injured again. In days long after the player's coming back to regular physical activities after an injury (i.e., more than 3 days), the club should control metabolic workload (i.e., $d_{\text{HML}}^{(\text{MSWR})}$) and mechanical workload (i.e., $DEC_2^{(\text{WF})}$), which can lead to injuries at specific values.

8 Conclusion

In this paper, we propose a machine learning approach to injury prediction for a professional soccer club. To the best of our knowledge, this is the first time machine learning is successfully applied to this challenging problem in professional soccer. Due to its high precision and interpretability, the proposed approach can be a valid support to coaches and athletic trainers, which can use it to prevent injuries and at the same time to understand the reasons behind the occurrences of these events. Our approach can be easily implemented for every club that performs automatic data collection with the standard sensing technologies available nowadays.

Our work can be extended in many interesting directions. First, in our experiments we extract a limited set of workload features widely accepted in the sports science literature. A way to enlarge our description of a player's health is to include performance features extracted from official games, where the player is exposed to high physical and psychological stress. When information about match workload is available, we expect physical effort features to be more relevant in the classification since games are the highest intensity demand in a week and considerably impact the general fitness of a player.

Second, we plan to investigate the “transferability” of our approach from a club to another, i.e., if a classifier trained on a set of players can be successfully applied to a distinct set of players, not used during the training process. If so, it would be possible to exploit collective information to train a more powerful classifier which includes training examples from different players, clubs, and leagues. Finally, if data covering several seasons of a player's activity are available, a specific classifier can be trained for every player to monitor his physical behavior. The quality of our approach could greatly benefit from the availability of other types of data measuring different aspects of health, such as heart rate, ventilation, and lactate, as new sensor technologies are available today to measure these specific quantities. When these data will become available,

they will allow us to refine our study on the relation between performance and injury risk.

In the meanwhile, experiences like ours may contribute to shaping the discussion on how to forecast injuries by a multidimensional approach based on automatically collected features of training workload and performance, that are starting to be massively available everywhere in the soccer industry. As we show in this paper, we have the potential of creating easy-to-use tools in support of soccer practitioners. This is crucial since the decisions of managers and coaches, and hence the success of clubs, also depend on what they measure, how good their measurements are, the quality of predictions and how well predictions are understood.

A Descriptive statistics of workload features

Table A shows the average (AVG) and the standard deviation (SD) of the distributions of the 12 training workload features considered in our study. We assess the normality of the distributions by using the Shapiro-Wilks' Normality test (SW) and observe that none of them is normally distributed (see Table A). Indeed, by a visual inspection of the distributions, we observe that they tend to be bimodal and right skewed (Figure 8).

	AVG	SD	SW
d_{TOT}	3882.94	1633.21	<0.01
d_{HSR}	133.22	66.41	<0.01
d_{MET}	1151.99	694.25	<0.01
d_{HML}	543.89	339.64	<0.01
$d_{HML/min}$	8.70	6.09	<0.01
d_{EXP}	410.67	221.29	<0.01
Acc_2	64.26	31.72	<0.01
Acc_3	16.16	10.97	<0.01
Dec_2	62.44	33.09	<0.01
Dec_3	19.14	12.78	<0.01
DSL	117.98	78.52	<0.01
FI	0.63	0.31	<0.01

SD = Standard Deviation; SW = Shapiro-Wilks' Normality test.

Table 6 Descriptive statistics of the 12 training workload features. We provide three categories of training workload features: kinematic features (blue), metabolic features (red) and mechanical features (green).

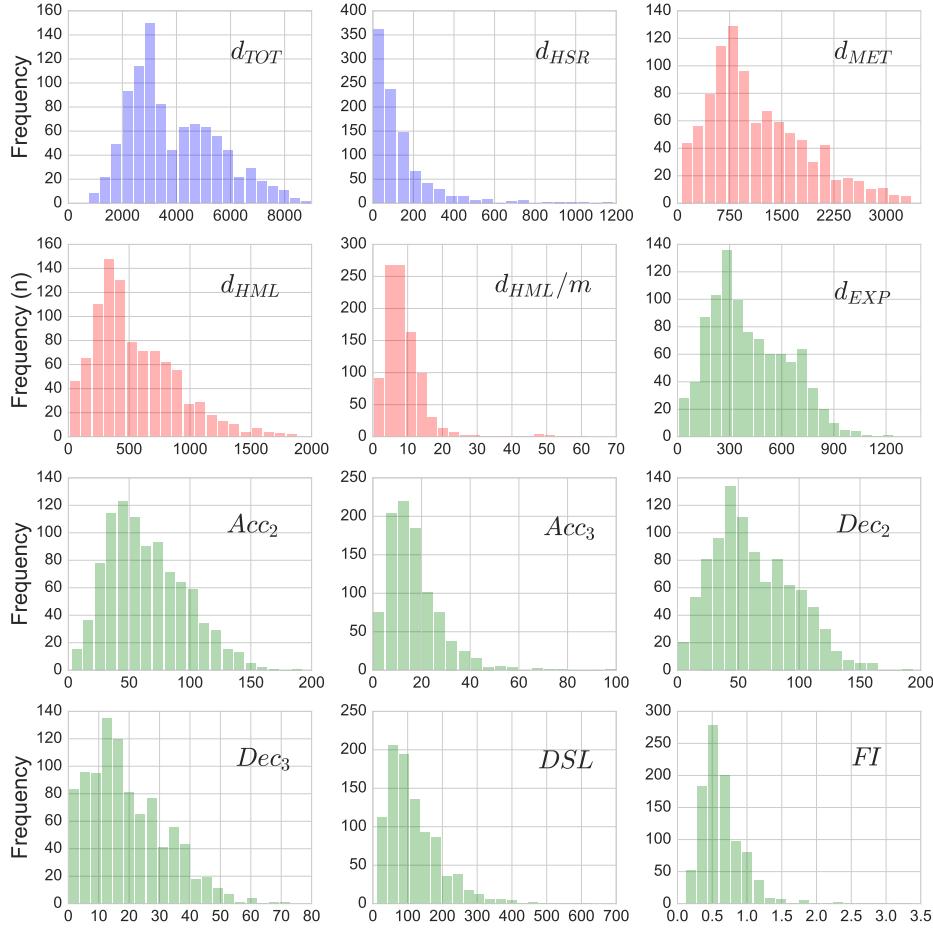


Fig. 8 Distribution of workload features. We provide three categories of training workload features: kinematic features (blue), metabolic features (red) and mechanical features (green).

B ACWR with data-driven groups

We repeat the ACWR approach by using the quintiles of the ACWR distribution (ACWR_q) instead of predefined ACWR groups suggested by Murray et al. [29]. Figure 9 shows the injury likelihood (IL) for every ACWR group in all the 12 workload features. We observe that groups with low ACWR_q are associated to the highest injury risk, substantially confirming the experiments made using predefined ACWR groups. We construct ACWR_q predictors following the same strategy described in the manuscript but using quantiles instead of predefined groups. Table 7 visualizes the results of classification. We observe results similar to those presented in the manuscript for the pre-defined ACWR groups: the predictors are a little usable in practice due to their too low precision.

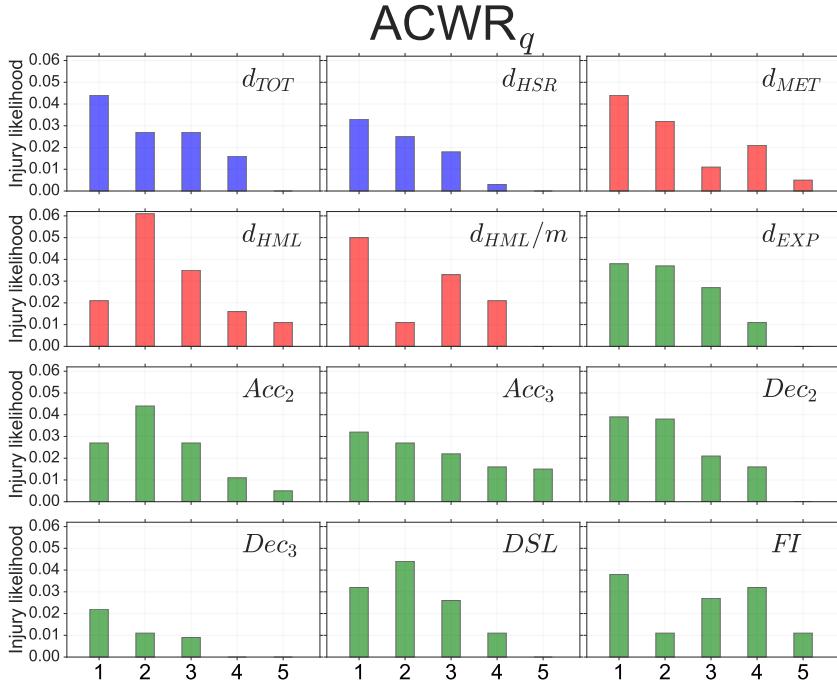


Fig. 9 Injury likelihood in $ACWR_q$ groups. The plots show IL for the ACWR groups defined the quantiles of the distribution, for every of the 12 training workload features considered in our study. We provide three categories of training workload features: kinematic features (blue), metabolic features (red) and mechanical features (green).

C Exponential Weighted Moving Average (EWMA)

To consider the recent training workload of a player, we compute the exponential weighted moving average (EWMA) of his most recent training sessions. The EWMA decreases exponentially the weights of the values according to their recency [46, 47], i.e., the more recent a value is the more it is weighted in an exponential function according to a decay $\alpha = 2/(span - 1)$. In accordance with the exponential function the moving average is computed as:

$$EWMA_t = \alpha[x_t - (x_{t-1} + (1 - \alpha)^2 x_{t-2} + \dots + (1 - \alpha)^{n-1} x_{t-n})] + x_t.$$

We vary $span = 1, \dots, 10$ to detect the value leading to the best classification performance. We hence train a decision tree on the feature set ALL by using every of the ten $span$ values. Figure 10 shows the cross-validated AUC and F1-score of the decision tree C_{ALL} varying the value of $span$. We observe that a 6 training span is the best predictive window to injury prediction in our dataset (Figure 10).

D Computation of PI

To take into account the injury history of a player, we compute the EWMA of the number of injuries in previous weeks. $PI^{(WF)}$ reflects the temporal distance between a player's training session and the return of the player to regular trianing after an injury. $PI^{(WF)} = 0$ represents

ACWR_q	class	prec	rec	F1	AUC
$C_{d_{TOT}}$	0	0.98	0.80	0.88	0.59
	1	0.04	0.38	0.08	
$C_{d_{HSR}}$	0	0.99	0.37	0.54	0.57
	1	0.03	0.76	0.05	
$C_{d_{MET}}$	0	0.98	0.80	0.88	0.59
	1	0.04	0.38	0.08	
$C_{d_{HML}}$	0	0.99	0.81	0.89	0.67
	1	0.06	0.52	0.10	
$C_{d_{HML/m}}$	0	0.98	0.81	0.89	0.62
	1	0.05	0.43	0.09	
$C_{d_{EXP}}$	0	0.98	0.80	0.88	0.57
	1	0.04	0.33	0.07	
C_{Acc_2}	0	0.98	0.80	0.88	0.59
	1	0.04	0.38	0.08	
C_{Acc_3}	0	0.98	0.80	0.64	0.54
	1	0.03	0.29	0.06	
C_{Dec_2}	0	0.98	0.80	0.88	0.57
	1	0.04	0.33	0.07	
C_{Dec_3}	0	0.99	0.46	0.63	0.66
	1	0.04	0.86	0.07	
C_{DSL}	0	0.98	0.80	0.88	0.59
	1	0.04	0.38	0.08	
C_{FI}	0	0.98	0.80	0.88	0.57
	1	0.04	0.33	0.07	
C_{ACWRq}^{one}	0	0.99	0.21	0.34	0.56
	1	0.02	0.91	0.05	
C_{ACWRq}^{vote}	0	0.99	0.83	0.90	0.64
	1	0.06	0.46	0.10	
C_{ACWRq}^{all}	0	0.98	1.00	0.99	0.50
	1	0.00	0.00	0.00	
B_1	0	0.98	0.98	0.98	0.51
	1	0.06	0.05	0.05	
B_2	0	0.98	1.00	0.99	0.50
	1	0.00	0.00	0.00	
B_3	0	0.00	0.00	0.00	0.50
	1	0.02	1.00	0.04	
B_4	0	0.98	0.77	0.86	0.60
	1	0.04	0.43	0.07	

Table 7 Injury prediction report of ACWR_q. We report precision (prec), recall (rec), F1-score (F1) and Area Under the Curve (AUC) for the injury class and the non-injury class for all the predictors defined on ACWR and monotony methodologies. We also provide predictive performance of four baseline predictors B_1 , B_2 , B_3 and B_4 .

players who never got injured in the past. $PI^{(WF)} > 0$ represents players who get injured at least once in the past. Table 8 provides specific $PI^{(WF)}$ thresholds in players incurred from 1 to 4 previous injuries. For example, $PI^{(WF)} = 0.50$ reflects a training performed by a player 3 days since his return to regular training after an injury.

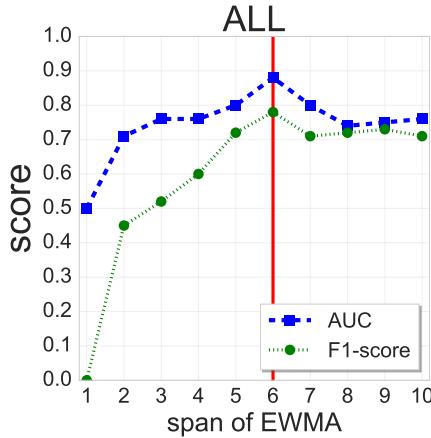


Fig. 10 We plot the AUC and F1-score of EWMA with $span = 1, \dots, 10$ in C_{ALL} . The red line reflects the best span to injury prediction.

injuries	PI_i					
	1	2	3	4	5	6+
1	0.29	0.49	0.64	0.74	0.81	> 0.86
2	1.27	1.48	1.63	1.74	1.81	> 1.86
3	2.27	2.46	2.62	2.72	2.80	> 2.85
4	3.25	3.46	3.53	3.66	3.76	> 3.83

Table 8 $PI^{(WF)}$ values after n training days (i.e., $n = 1 \dots 6$) since the return of a player to regular training. We report the values for different n of previous injuries (i.e., $n = 1 \dots 4$). PI_i is the number of training days long after players return to regular physical activity. 6+ indicates values for 6 and more than 6 days.

E Predictions using Random Forest

In addition to the Decision Tree (DT) provided in the manuscript, we also train an Extra Tree Random Forest Classifier (ETRFC) to assess the predictive power of ensemble methods.⁷ ETRFC is a robust model that fits a number of randomized decision trees on various sub-samples of the dataset, controlling the possible over-fitting. However, a ETRFC is more complex to interpret due to its complexity. Table 9 shows the performance of ETRFC classifier trained on WF, ACWR, MON, ALL. We use the same 3-fold stratified cross-validation used for DT. We observe that the performance of ETRFC is slightly worse than DT's performance. Again, C_{ALL} is the best classifier in terms of both precision and recall, being able to detect 78% of the injuries and achieving 88% precision. Moreover, the classifiers outperform both the baselines and predictors based on ACWR and MSWR.

Acknowledgements This work has been partially funded by the EU project SoBigData grant n. 654024.

⁷ We use the Python package `scikit-learn` to train and validate the decision tree.

	model	class	prec	rec	F1	AUC	selected features
random forest	C _{ALL}	0	0.99	0.99	0.99	0.86	PI ^(WF) , $d_{HML}^{(MON)}$, Dec ₃ ^(WF)
		1	0.88	0.78	0.82		
	C _{MON}	0	0.99	1.00	0.99	0.76	PI, d_{TOT}
		1	1.00	0.52	0.67		
CACWR	C _{CACWR}	0	0.99	1.00	0.99	0.73	PI, d_{TOT} , Acc ₃
		1	0.90	0.43	0.58		
	C _{WF}	0	0.98	1.00	0.99	0.63	PI, Dec ₃
		1	1.00	0.29	0.44		
ACWR	C _{CACWR} _{vote}	0	0.99	0.83	0.90	0.65	12 workload features in Table 1
baselines	B ₄	0	0.98	0.77	0.86	0.60	
		1	0.04	0.43	0.07		
	B ₁	0	0.98	0.98	0.98	0.51	
		1	0.06	0.05	0.05		
B ₂	B ₂	0	0.98	1.00	0.99	0.50	
		1	0.00	0.00	0.00		
	B ₃	0	0.00	0.00	0.00	0.50	
		1	0.02	1.00	0.04		

Table 9 Performance of ETRFC and baseline classifiers. The goodness of classifiers and baselines in terms of recall, precision, F1 and AUC. For every classifier we also show the features selected by the feature selection task.

Authors contribution A.R. and L.P. designed and performed all the experiments, A.R. made the plots and the figures. M.I. collected the GPS data and the injury data. A.R., L.P., J.F., D.M. and P.C. contributed to writing the manuscript.

References

1. Hägglund M, Waldén M, Magnusson H, Kristenson H, Bengtsson H and Exstrand J. Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. British Journal of Sports Medicine, doi: 10.1136/bjsports-2013-092215, 2013.
2. Hurley OA. Impact of Player Injuries on Teams' Mental States, and Subsequent Performances, at the Rugby World Cup 2015. Frontiers in Psychology 7:807, doi: 10.3389/fpsyg.2016.00807, 2016.
3. Lehmann EE, Schulze GG. What Does it Take to be a Star? – The Role of Performance and the Media for German Soccer Players. Applied Economics Quarterly 54:1, pp. 59-70, doi: 10.3790/aeq.54.1.59, 2008.
4. Fernández-Cuevas I., Gomez-Carmona P, Sillero-Quintana M, Noya-Salces J, Arnaiz-Lastras J, Pastor-Barrón A. Economic costs estimation of soccer injuries in first and second Spanish division professional teams. 15th Annual Congress of the European College of Sport Sciences ECSS, 23th 26th june. 2010.
5. Gudmundsoon H, Horton M. Spatio-Temporal Analysis of Team Sports - A Survey. CoRR: abs/1602.06994, 2016.
6. Stein M., Janetzko H., Seebacher D., Jäger A., Nagel M., Hölsch, J., Grossniklaus M. How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects. Data, 2:1, 2, doi:10.3390/data2010002, 2017.

7. Cintia P., Pappalardo L., Pedreschi D. Engine Matters: A First Large Scale Data Driven Study on Cyclists' Performance. 13th IEEE International Conference on Data Mining Workshops, pp. 147–153, doi:10.1109/ICDMW.2013.41, 2013.
8. Cintia P., Coscia M., Pappalardo L. The Haka network: Evaluating rugby team performance with dynamic graph analysis. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1095–1102, doi: 10.1109/ASONAM.2016.7752377, 2016
9. Cintia P., Pappalardo L., Pedreschi D., Giannotti F., Malvaldi M. The harsh rule of the goals: Data-driven performance indicators for football teams. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10, doi:10.1109/DSAA.2015.7344823, 2015
10. Pappalardo L., Cintia P. Quantifying the relation between performance and success in soccer. CoRR: abs/1705.00885, 2017.
11. Fernández J., Medina D., Gómez A., Arias M., Gavaldá R. From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data. 16th IEEE International Conference on Data Mining Workshops, doi:10.1109/ICDMW.2016.0027, 2016
12. Rossi A., Savino M., Perri E., Aliberti G., Trecroci A., Iaia M. Characterization of in-season elite football trainings by GPS features: The Identity Card of a Short-Term Football Training Cycle. 16th IEEE International Conference on Data Mining Workshops, pp. 160–166, doi: 10.1109/ICDMW.2016.0030, 2016.
13. Brink MS1, Visscher C., Arends S., Zwerver J., Post WJ., Lemmink KA. Monitoring stress and recovery: new insights for the prevention of injuries and illnesses in elite youth soccer players. *Br J Sports Med.* 2010;44: 809–15.
14. Ehrmann FE, Duncan CS, Sindhusake D, Franzsen WN, Greene DA. GPS and injury prevention in professional soccer. *J Strength Cond Res.* 2015;30:306–307.
15. Venturelli M, Schena F, Zanolla L, Bishop D. Injury risk factors in young soccer players detected by a multivariate survival model. *Journal of Science and Medicine in Sport.* 2011;14:293298.
16. Kirkendall D.T., Dvorak J. Effective Injury Prevention in Soccer. *The physician and sportsmedicine*, 38:1, doi: <http://dx.doi.org/10.3810/psm.2010.04.1772>, 2010.
17. Hagglund M, Walden M, Bahr R, Ekstrand J. Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *British Journal of Sports Medicine*, 39:6, pp. 340–346, doi:10.1136/bjsm.2005.018267, 2005.
18. FIFA circular no. 1494 of 8 July 2015, <http://bit.ly/2pBmhK7>.
19. Gabbett TJ. Reductions in pre-season training loads reduce training injury rates in rugby league players. *British Journal of Sports Medicine.* 2004;38: 743749.
20. Gabbett TJ, Jenkins DG. Relationship between training load and injury in professional rugby league players. *Journal of Science and Medicine in Sport.* 2011;14: 204209.
21. Gabbett TJ. Influence of training and match intensity on injuries in rugby league. *Journal of Sports Sciences.* 2004;22(5):409–417.
22. Gabbett TJ. The development and application of an injury prediction model for non-contact, soft-tissue injuries in elite collision sport athletes. *The Journal of Strength & Conditioning Research.* 2010;24(10):2593–2603.
23. Gabbett TJ, Domrow N. Relationships between training load, injury, and fitness in sub-elite collision sport athletes. *Journal of Sports Sciences.* 2007;25(13):1507–1519.
24. Anderson L, Triplett-McBride T, Foster C, Doberstein S, Brice G. Impact of training patterns on incidence of illness and injury during a women's collegiate basketball season. *The Journal of Strength & Conditioning Research.* 2003; 17: 734738.
25. Gabbett TJ, Ullah S. Relationship between running loads and soft-tissue injury in elite team sport athletes. *J Strength Cond Res.* 2012;26: 953960.
26. Rogalski B, Dawson B, Heasman J, Gabbett TJ. Training and game loads and injury risk in elite Australian footballers. *J Sci Med Sport.* 2013;16: 499503.
27. Gabbett TJ. The training-injury prevention paradox: should athletes be training smarter and harder? *Br J Sports Med.* 2016; bjsports-2015-095788.
28. Hulin BT, Gabbett TJ, Blanch P, Chapman P, Bailey D, Orchard JV. Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. *Br J Sports Med.* 2014;48:708–712.

29. Murray NB, Gabbett TJ, Townshend AD, Blanch P. Calculation acute:chronic workload ratios using exponential weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages. *Br J Sports Med.* 2016; bjsports-2016-097152.
30. Foster C. Monitoring training in athletes with reference to overtraining syndrome. *Med Sci Sports Exerc.* 1998;30:11641168.
31. Talukder H, Vincent T, Foster G, Hu C, Huerta J, Kumar A, et al. Preventing in-game injuries for NBA players. MIT Sloan Analytics Conference. Boston; 2016.
32. Buchheit M, Al Haddad H, Simpson BM, Palazzi D, Bourdon PC, Di Salvo V, Mendez-Villanueva A. Monitoring accelerations with GPS in football: time to slow down? *International Journal of sports physiology and performance,* 9 (3), pp. 442-445, doi: 10.1123/ijsspp.2013-0187, 2014.
33. Medina D, Pons E, Gomez A, Guitart M, Martin A, Vazquez-Guerrero J, Camenforte I, Carles B, Font R. Are There Potential Safety Issues Concerning the Safe Usage of Electronic Personal Tracking Devices? The Experience of a Multi-sport Elite Club. *International Journal of sports physiology and performance,* 4, pp. 1-12, doi: 10.1123/ijsspp.2016-0368, 2017.
34. Duncan MJ, Badland HM, Mummerly WK. Applying GPS to enhance understanding of transport-related physical activity. *Journal of Science and Medicine in Sport.* 2009;12: 549556.
35. Hulin BT, Gabbett TJ, Lawson DW, Caputi P, Sampson JA. The acute:chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players *Br J Sports Med.* 2016;50:231-236.
36. Bowen L., Gross AS, Gimpel M, Li FX. Accumulated workloads and the acute:chronic workload ratio relate to injury risk in elite youth football players. *British Journal of Sports Medicine,* 51, 452-459, doi:10.1136/bjsports-2015-095820, 2017.
37. Menaspá P. Are rolling averages a good way to assess training load for injury prevention? *British Journal of Sports Medicine,* 51, doi:<http://dx.doi.org/10.1136/bjsports-2016-096131>, 618-619, 2017.
38. Fuller CW, Ekstrand J, Junge A, Andersen TE, Bahr R, Dvorak J, Hägglund M, McCrory P, Meeuwisse WH. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *British Journal of Sports Medicine,* 40, pp. 193-201, doi:<http://dx.doi.org/10.1136/bjsm.2005.025270>, 2006.
39. Di Prampero PE, Fusi S, Sepulcri L, Morin JB, Belli A, Antonutto G. Sprint running: a new energetic approach. *J Exp Biol.* 2005;208: 28092816.
40. Gaudino P, Iaia FM, Alberti G, Strudwick AJ, Atkinson G, Gregson W. Monitoring training in elite soccer players: systematic bias between running speed and metabolic power data. *Int J Sports Med.* 2013;34: 963968.
41. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.* New York, NY: Springer New York; 2013.
42. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning* 46, 2002, 10.1023/A:1012487302797.
43. Li J., Cheng K., Wang S., Morstatter F., Trevino R., Jiliang T., Huan L. Feature Selection: A Data Perspective. *arXiv:1601.07996,* 2016
44. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning.* Springer series in Statistics, volume 1, Springer, Berlin, 2001.
45. Hastie TJ and Tibshirani RJ and Friedman JH *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics, isbn: 978-0-387-84857-0, 2009
46. Lowry CA, Woodall WH, Champ CW, Rigdon SE. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics.* *Technometrics.* 1992;34: 4653.
47. Lucas JM, Saccucci MS. Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. *Technometrics.* 1990;32: 112.