Who is going to get hurt? Predicting injuries in professional soccer

Alessio Rossi[1], Luca Pappalardo[2,3], Paolo Cintia[2,3],
Javier Fernandez[4], F. Marcello Iaia[1], and Daniel Medina[4]

[1] Department of Biomedical Science for Health, University of Milan, Italy
alessio.rossi2@gmail.com
[2] Department of Computer Science, University of Pisa, Pisa, Italy
lpappalardo@di.unipi.it
[3] ISTI-CNR, Pisa, Italy
[4] Sports Science and Health Department, Football Club Barcelona, Spain

Abstract. Injury prevention has a fundamental role in professional soccer due to the high cost of recovery for players and the strong influence of injuries on a club's performance. In this paper we provide a predictive model to prevent injuries of soccer players using a multidimensional approach based on GPS measurements and machine learning. In an evolutive scenario, where a soccer club starts collecting the data for the first time and updates the predictive model as the season goes by, our approach can detect around half of the injuries, allowing the soccer club to save 70% of a season's economic costs related to injuries. The proposed approach can be a valuable support for coaches, helping the soccer club to reduce injury incidence, save money and increase team performance.

Keywords: sports analytics, data science, machine learning, sports science, predictive analytics.

1 Introduction

Injuries are an important issue in professional soccer, as they can negatively affect team performance and represent a remarkable expense for soccer clubs. The cost associated with the process of recovery and rehabilitation for a player is often considerable, especially in terms of medical care and missed earnings from merchandising [1]. It has been observed that injuries in Spain cause in average around 16% of season absence by players, corresponding to a total cost estimation of 188 million euros just in one season [2]. Hence, it is not surprising that injury prediction is attracting a growing interest from soccer managers, who are interested in intervening with appropriate actions to reduce the likelihood of injuries of their players. Due to its importance for a club's economy and success, a big effort has been put in the sports science literature on investigating injury prediction in professional soccer [10–12]. A major limitation of existing studies is that they follow a monodimensional approach, i.e., they use just one variable at a time to estimate injury risk thus not fully exploiting the complex patterns underlying measurable aspects of soccer performance. Moreover, in these works statistical modeling is used mainly to quantify the relation between the chosen variable and injury likelihood, while an evaluation of the predictive power of a player's performance is still missing [8, 6].

In this paper, we propose a data-driven, multidimensional approach to injury prediction, considered as the problem of forecasting whether or not a player will get injured in the next training session or official game, given his recent training workload. Our approach is based on automatic data collection through standard Electronic Performance and Tracking Systems (EPTS) [4, 7, 5, 6], 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 preprocessing task, we extract from the data a set of features used in sports science to describe aspects of training workload, and we enrich them with information about all the injuries which happen during the half season. We found 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 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 can detect more than half of the injuries during the season, indicating that by using our predictor the soccer club could have been saved 70% of injury-related costs.

2 Related Work

Several studies performed by Gabbett et al. [13–18, 21] show that muscular injuries are to some extent preventable. In rugby, they find that a player has a high injury risk when his workload is above a certain threshold. The same results are observed by Hulin et al. [22] and Ehrmann et al. [11] for cricket players and soccer players, respectively. In particular, all these studies assess 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), defining specific thresholds to detect players who could incur in a injury in the future training sessions.

The "monotony session load", i.e., the ratio between the mean and the standard deviation of the session load is widely used in literature. In skating, Foster et al. [23] 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 [23]. In basketball, Anderson et al. [18] find a correlation between injury risk and monotony session load. In soccer, Brink et al. [24] observe that injured players record higher values of monotony in the week preceding the injury than non-injured players.

Some studies also show that technical-tactical performance during official matches can affect the players' physical fit. Talukder et al. [25] propose a classifier able to predict 19% of the injuries occurred in NBA using the players' technical-tactical performance. They show that the most important features for injury prediction in basket 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.

From the literature, it is clear that all injury prediction studies for soccer suffer from a major limitation: they investigate the correlation between a single aspect of training workload and injury likelihood but they do not construct any predictor as a tool to make predictions and prevent injuries. Therefore, to the best of our knowledge, there is no quantification of the potential of predictive analytics in preventing injuries in professional soccer.

3 Dataset preparation

3.1 Data collection and feature extraction

During the season 2013/2014 we monitor the position of twenty-six professional football players competing in the Italian Serie B during 23 training sessions – from January 1st to May 31st – 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). Each player wore a tight vest where the receiver was placed between their scapulae, and every player wore his own GPS device for each training session. We recorded a total of 954 individual training sessions during the 23 weeks and extracted from the data a set of training workload indicators through the software package Viper Version 2.1 (STATSports 2014). From every training session we extracted 12 features describing kinematic, metabolic and mechanical aspects of the individuals' trainings. For each player, we also collected information about age, weight, height and role on the field. Moreover, for each player's training session 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. Table 1 provides a description of the considered features.

The club's medical staff recorded all the non-contact injuries occurred during 23 weeks. 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. In this dataset there are 21 non-contact injuries in total.

3.2 Feature engineering and dataset construction

We construct four training sets transforming the 12 workloads features described in Table 1 in the following way:

1. *Workload Features set (WF)* – we consider the training workloads in the 6 most recent training sessions by using an exponential weighted moving average (EWMA). We also compute the EWMA of feature PI with a span equal to 6 (PI^{WF}) in order to take into account both the number of a player's previous injuries and their temporal distance to the current training sessions. $\text{PI}^{\text{WF}} = 0$ indicates that the player never got injured in the past; $\text{PI}^{\text{WF}} > 0$

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. Description of the training workload features extracted from GPS data and the players' personal features collected during the study.

indicates that the player got injured at least once in the past; $\text{PI}^{\text{WF}} > 1$ indicates that the player got injured more than once in the past.

2. *Acute:Chronic Workload Ratio features set (ACWR)* – here we consider the standard *de facto* used in sports science to estimate injury likelihood [9] and compute the ratio between the 6 most recent training sessions by the EWMA and the EWMA of the previous 28 days.
3. *Mean over Standard deviation Workload Ratio (MSWR)* – we consider another way proposed in literature to estimate injury likelihood [10] and compute the ratio between the mean and the standard deviation of the training workloads in the 6 most recent days. The higher the MSWR of a player, the lower is the variability of his workloads during the training week.
4. we build a dataset based on the union of the three feature sets described above (WF, ACWR and MSWR) and the personal features in Table 1. This dataset consists of a vector of 42 features and the injury label indicating whether or not the player gets injured in next match or training session.

Every training set consists of 954 examples (i.e., individual training sessions) corresponding to 80 collective training sessions.

4 Experiments

First of all, we perform a feature selection process based on a Decision Tree Classifier in order to reduce the dimensionality of the feature space and consequently the risk of overfitting. We use recursive feature elimination with cross-validation (RFECV) to select the best set of features able to predict injuries in our dataset.

On the new training dataset derived from the feature selection, we train a Decision Tree classifier (DT) and a Random Forest Classifier (ETRFC).[5] In particular, we investigate a scenario where the club starts to record data at the beginning of a season and trains the classifier as the season goes by. Hence, we proceed from the first training week (w_1) to the most recent one ($w_{\text{i-1}}$). At training week w_{i} we train the classifiers on weeks $w_1 \dots w_{\text{i}}$ and evaluate their ability to predict injuries on week $w_{\text{i}+1}$.

Considering injury prediction as a binary classification problem where the injury class (1) is the positive class, we measure the goodness of the classifiers week by week in terms of precision, recall, F1-score and AUC [27]. Precision indicates the fraction of examples that the classifier correctly classifies over the number of all examples the classifier assigns to that class. Recall indicates the ratio of examples of a given class correctly classified by the classifier, while F1-score is the harmonic mean of precision and recall. AUC (Area Under the Curve) is 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 close to 1 represents an accurate classification, while an AUC close to 0.5 represents a random classification.

We compare the goodness of DT and ETRFC 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 PI > 0, and 0 (no injury) otherwise. Finally, we estimate the economic cost of the injuries for the considered soccer club by using the methodology suggested by Fernandez et al. [3], i.e., we multiply the number of days of "work" absence by the minimal legal salary per day in the Italian Serie B.

4.1 Results

Just 3 features out of 42 are selected by the feature selection task: $\text{PI}^{(\text{WF})}$, $d_{\text{HML}}^{(\text{MSWR})}$ and $DEC_2^{(\text{WF})}$. Feature $\text{PI}^{(\text{WF})}$ 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. Features $d_{\text{HML}}^{(\text{MSWR})}$ and $DEC_2^{(\text{WF})}$ are two training features indicating high metabolic load and sudden decelerations, respectively. We observe that 42% of the injuries detected by the classifier happened immediately after the coming back to regular training of players who got injured in the past, and are characterized by specific values of $d_{\text{HML}}^{(\text{MSWR})}$ and $DEC_2^{(\text{WF})}$, which indicate

[5] We use the Python package `scikit-learn` to train and test all the classifiers.

the metabolic workload variability and the average of sudden decelerations in the previous 6 days, respectively.

Figure 1 shows the evolution of the F1-score of DT, ETRFC and the four baselines (i.e., B_1, ..., B_4) as the season goes by. Due to the low number of injury examples, the classifiers have a poor predictive performance at the beginning of the season and miss many injuries (black crosses in Figure 1). However, the predictive ability improves by time and the classifiers predict most of the injuries in the second half of the season (red crosses in Figure 1). The cumulative performance of the classifiers is highly affected by the initial period, where injury examples 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, whose length depends on the needs and strategy of the club, is needed in order to collect the adequate amount of data, and only then reliable classifiers can be trained on the collected data. Regarding this aspect, in our dataset, we observe that the performance of the classifiers stabilizes after 16 weeks of data collection (Figure 1). In our case, a reasonable strategy could be to use the classifiers for injury prevention starting from the 16th week. This suggests that the considered club could effectively use the classifiers trained on data from a season to perform injury prediction since the first session of the second half of the current season.

We observe that DT is the best classifier in this scenario detecting more than half of the injuries (11 injuries out of 21), resulting in a cumulative F1-score = 0.45 (Figure 1).[6] Table 2 shows the classification reports of the two classifiers and the four baseline at the end of the season. We find that DT is significantly better than the baselines (Table 2). At the end of the season, DT detects 58% of the injuries (recall = 0.58) and it correctly predicts 38% of the cases classified as injuries (precision = 0.38). Although the machine learning approach significantly adds predictive power with respect to existing methods, there is still room for improvement. Soccer clubs are indeed interested in an algorithm with high precision to reduce “false alarms”, which could negatively affect a team’s performance due to the forced absence of crucial players.

We also train DT, ETRFC and the baselines using the entire feature set, i.e., without performing any feature selection process. These classifiers perform slightly worse than the classifiers build on the three selected features (precision, recall, F1-score and AUC are 0.36, 0.52, 0.43, and 0.74, respectively). To understand if the role of a player affects injury likelihood, we train distinct classifiers for every role (defender, midfielder, forwards) and find that they perform much worse that the classifiers trained without distinguishing between the roles (precision, recall, f1-score and AUC are 0.01, 0.04, 0.03 and 0.51, respectively).

Figure 2 shows the distribution of the number of days of work absence recorded during the season. The number of work days of absence due to injuries is 139, i.e., 6% of the working days. Generally, a player returns to regular

[6] DT has the following meta-parameters: max depth = 3, minimum samples for a leaf = 2, minimum sample split = 11. For all the other meta-parameters we use default values suggested by `sciki-learn` (see documentation: http://bit.ly/1T5sf92).

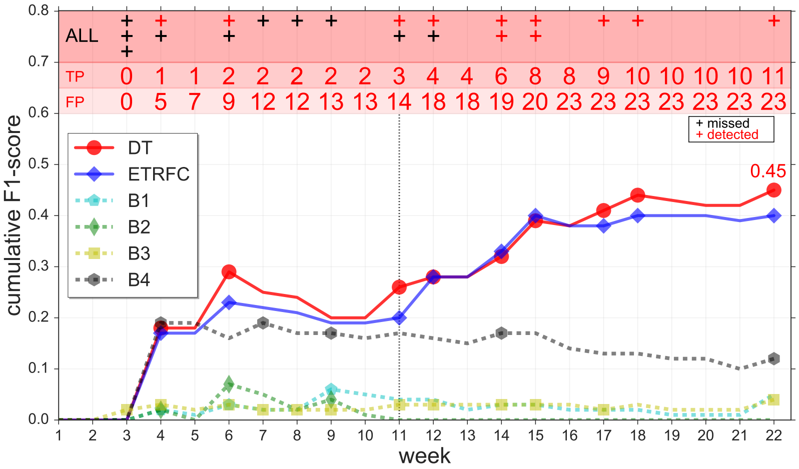

Fig. 1. Performance of classifiers in the evolutive scenario. We plot the cumulative F-score of the classifiers and the baselines, week by week. For every week we highlight in red the number of injuries detected by *DT* up to that week.

model	class	prec	rec	F1	AUC
DT	0 **1**	0.98 **0.38**	0.99 **0.58**	0.99 **0.45**	**0.76**
ETRFC	0 1	1.00 0.35	0.98 0.57	0.98 0.43	0.71
B_4	0 1	0.98 0.04	0.71 0.20	0.83 0.12	0.56
B_1	0 1	0.98 0.06	0.98 0.05	0.98 0.05	0.51
B_2	0 1	0.98 0.00	1.00 0.00	0.99 0.00	0.51
B_3	0 1	0.00 0.02	0.00 1.00	0.00 0.04	0.51

Table 2. Performance of classifiers compared to baselines. We report the performance of classifiers DT and ETRFC in terms of precision, recall, F1 and AUC at the end of the season. We compare the classifier with four baseline $B_1, \dots, B_4$.

physical activity within 5 days (i.e., 15 times out of 21 injuries), while only 6 times a player needed more than 5 days to recover. We estimate a (minimum) total cost related to injuries of 11,583 euros (139x83 euros = *days of absence* x *minimal legal salary per day*) corresponding to 3.81% of the salary cost of the soccer club (from January 1st to May 31st the club spent 303,750 euros for the players' salary). By using DT to predict injuries as the season goes by, the soccer club could had been able to prevent 11 injuries and save 8,300 euros, 70% of the economic costs related to injuries during the season (100x83 euros = *day of absence* x *minimal legal salary per day*).

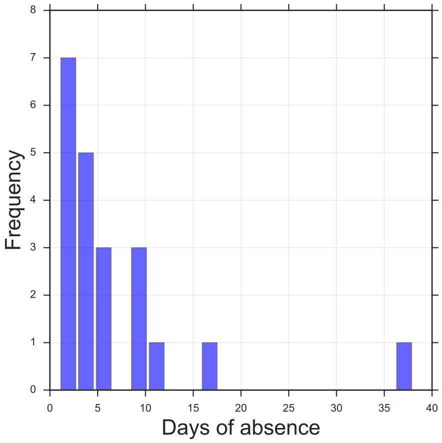

Fig. 2. Distribution of the number of days of work absence after an injury.

5 Conclusion

This study presents a method to predict injuries of soccer players. Athletic trainers, coaches and physiotherapists can use our method to make decisions about whether or not to stop a player in next official match, thus eventually preventing his injury, improving team performance and reducing the club's costs. The proposed study provides an example of how machine learning can be used to solve a difficult problem in sports analytics such as predicting injuries. An enlargement of the dataset to include different teams, which is planned by the authors of this paper, might allow to build a more general and robust algorithm for injury forecasting. With more injury cases we could transform the problem from a binary classification (injury/no-injury) to a multi-class classification or a regression problem, where information about the typology or the severity of the injuries can be exploited to produce more diverse predictions. Finally, due to its flexibility, our multidimensional approach can be easily extended to predict injuries in other professional sports, like rugby [13] and cycling [28].

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

References

1. 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.
2. 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.
3. Fernndez I, Gomez PM, Sillero M, Noya J, Arnaiz J, Pastor 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, At Antalya (Turkey), 2010.
4. Gudmundsoon H, Horton M. Spatio-Temporal Analysis of Team Sports - A Survey. CoRR: abs/1602.06994, 2016.
5. Cintia P., Rinzivillo S., Pappalardo L. A network-based approach to evaluate the performance of football teams. In Proceedings of the Machine Learning and Data Mining for Sports Analytics workshop (MLSA'15), ECML/PKDD 2015, Porto, Portugal.
6. Pappalardo L., Cintia P. Quantifying the relation between performance and success in soccer. eprint arXiv:1705.00885, 2017.
7. 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
8. Cintia P., Rinzivillo S., Pappalardo L. A network-based approach to evaluate the performance of football teams, Proceedings of the Machine Learning and Data Mining for Sports Analytics workshop (MLSA'15), ECML/PKDD 2015, 2015
9. 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.
10. 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.
11. 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.
12. 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.
13. Gabbett TJ. Reductions in pre-season training loads reduce training injury rates in rugby league players. British Journal of Sports Medicine. 2004;38: 743749.
14. 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.
15. Gabbett TJ. Influence of training and match intensity on injuries in rugby league. Journal of Sports Sciences. 2004;22(5):409-417.
16. Gabbett TJ. The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. The Journal of Strength & Conditioning Research. 2010;24(10):2593-2603.

17. 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.
18. 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.
19. Gabbett TJ, Ullah S. Relationship between running loads and soft-tissue injury in elite team sport athletes. J Strength Cond Res. 2012;26: 953960.
20. 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.
21. Gabbett TJ. The training-injury prevention paradox: should athletes be training smarter and harder? Br J Sports Med. 2016; bjsports-2015-095788.
22. 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.
23. Foster C. Monitoring training in athletes with reference to overtraining syndrome. Med Sci Sports Exerc. 1998;30:11641168.
24. 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.
25. 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.
26. 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.
27. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.
28. 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, ICDM Workshops, TX, USA, December 7-10, 2013, doi = 10.1109/ICDMW.2013.41.