Sport Sciences for Health (2023) 19:285–296
https://doi.org/10.1007/s11332-022-00932-1

ORIGINAL ARTICLE

Blood sample profile helps to injury forecasting in elite soccer players

Alessio Rossi[1] · **Luca Pappalardo**[2] · **Cristoforo Filetti**[3,4,5] · **Paolo Cintia**[1]

Received: 22 October 2021 / Accepted: 15 March 2022 / Published online: 11 May 2022

Abstract

Purpose By analyzing external workloads with machine learning models (ML), it is now possible to predict injuries, but with a moderate accuracy. The increment of the prediction ability is nowadays mandatory to reduce the high number of false positives. The aim of this study was to investigate if players' blood sample profiles could increase the predictive ability of the models trained only on external training workloads.

Method Eighteen elite soccer players competing in Italian league (Serie B) during the seasons 2017/2018 and 2018/2019 took part in this study. Players' blood samples parameters (i.e., Hematocrit, Hemoglobin, number of red blood cells, ferritin, and sideremia) were recorded through the two soccer seasons to group them into two main groups using a non-supervised ML algorithm (k-means). Additionally to external workloads data recorded every training or match day using a GPS device (K-GPS 10 Hz, K-Sport International, Italy), this grouping was used as a predictor for injury risk. The goodness of ML models trained were tested to assess the influence of blood sample profile to injury prediction.

Results Hematocrit, Hemoglobin, number of red blood cells, testosterone, and ferritin were the most important features that allowed to profile players and to analyze the response to external workloads for each type of player profile. Players' blood samples' characteristics permitted to personalize the decision-making rules of the ML models based on external workloads reaching an accuracy of 63%. This approach increased the injury prediction ability of about 15% compared to models that take into consideration only training workloads' features. The influence of each external workload varied in accordance with the players' blood sample characteristics and the physiological demands of a specific period of the season.

Conclusion Field experts should hence not only monitor the external workloads to assess the status of the players, but additional information derived from individuals' characteristics permits to have a more complete overview of the players well-being. In this way, coaches could better personalize the training program maximizing the training effect and minimizing the injury risk.

Keywords Injury risk · Training workload · GPS · Predictive model

Introduction

Preventing and predicting injuries are a hot topic for the sports industry due to their high impact on both economic and performance point of views [1, 2]. Hence, it is not surprising that injury prevention and in particular prediction is attracting a growing interest from researchers and field experts [3]. In fact, the scientific literature is growing fast providing machine learning algorithms that permit us to accurately predict when players will get injured or not [4]. Actually, the use of multidimensional models is fundamental to injury prediction, because sports injuries are a consequence of complex interactions of multiple risk factors [5, 6]. Moreover, in addition to prediction task, predictive modeling should provide injury risk factors to implement

✉ Alessio Rossi
alessio.rossi@di.unipi.it

1 Department of Computer Science, University of Pisa, Pisa, Italy

2 Institute of Information Science and Technologies, National Research Council, Pisa, Italy

3 Performance Department, Paris Saint-Germain FC, Paris, France

4 Faculty of Medicine and Surgery, School of Sport and Exercise Sciences, University of Rome Tor Vergata, Rome, Italy

5 School of Sports and Exercise Sciences, San Raffaele University, Rome, Italy

interventions to minimize the level of risk maximizing the training effect [5–7].

Actually, the recent scientific literature is focused on detecting multidimensional pattern in external workloads related to injury incidence [3, 4], but individuals' characteristics, such as sleep quality, muscular strength, and morphological characteristics combined with external and internal workloads data, could have an impact on players' wellness status and consequently on the risk of injuries. To the best of our knowledge, this is the first injury prediction study which combines training workloads and players' individual characteristics. In particular, profiling players in accordance with performance- and health-related status could help machine learning models to personalize the decision-making process in accordance with their physiological requests and characteristics, paving the way for transfer learning techniques: so far, it has been hard, for a squad, to use machine learning models trained on different squad: initial 4–8 weeks of training data are required every time the team wants to use an injury prediction model. The use of an approach based on baseline screening tests and a one focused on continuously monitoring the training workload as the season goes by were both used as indicators of the risk of injuries [3]. In the baseline screening test approach [8–10], the athletes were tested before the start of the competitive season (singular timepoint). Seow and colleagues [3] in their literature review demonstrated that a one-off baseline testing score may not be a true representation months later. Actually, the previous papers that use a baseline testing approach did not have a strong injury predictive performance [8–12], highlighting the fact that the use of a single time-point data recording may not be a true representation of the players' status. For example, Ruddy et al. further reported the inability to predict hamstring injuries using the baseline testing approach [9]. Differently, performing tests periodically (monitoring approach) showed a moderate–high injury prediction accuracy [3]. Monitoring training workloads along the season permits having an overview of the players' fitness status and consequently of their injury risk [7, 13–18]. As a matter of fact, it was demonstrated that the greater the training exposure is, the greater the injury risk is [13, 19, 20]. However, there may also be currently unknown factors, e.g., sleep, nutrition, and blood markers that could have a role in injury prediction [3].

This study is focused on profiling players in accordance with blood sample features due to the fact that blood analysis is a simple and powerful way to get data critical to anyone interested in assessing the individuals' biomedical status with the aim of improving athletic and personal performance [21, 22]. Actually, the decrements of hemoglobin, hematocrit, and red blood cell count are associated with the increment of training workloads [23]. As a matter of fact, the hematocrit "paradox" proposed by Brun et al. [24] shows that low values of hematocrit (< 40%) were found to be associated with a higher aerobic capacity and isometric adductor strength, while athletes with high hematocrits (> 44.6%) result in a status of over-training and/or iron-deficient, and with an increased blood viscosity and red cell disaggregability. Additionally, it was found a negative correlation between blood viscosity and fitness, while a positive one was detected against over-training score [21]. Moreover, it was also found that when hematocrit increases, there is a decrease in athletes' fitness status and ferritin, and an increment in overtraining risk [24]. Hence, based on the previous results, it is possible to suppose that insight derived from the blood sample features could provide important information on the athletes' status that could be related to injury risk.

Hence, the aim of this study was to assess if the combination of the periodical screening tests (player profile in accordance with blood sample analysis) and continuous monitoring of the training workloads could increase the accuracy of a machine learning model to injury forecasting. In particular, this study compares machine learning performance models trained using only workloads data, by creating independent models for each blood sample class, and by adding the blood sample classes as independent features in the model. In this way, it was possible to obtain information about how much the accuracy changed by providing blood sample information in addition to the external workload features.

Methods

Subjects

Eighteen elite soccer players—age = 24.7 (4.3) yrs; height = 183.73 (7.16) cm; weight = 78.81 (7.32) kg—competing in the Italian league (Serie B) during the seasons 2017/2018 and 2018/2019 took part in this study—128 (36) sessions per player. Before starting the data recording, the soccer player signed the informed consent with the soccer club giving their consent to the collection and use of their data for research purposes. The soccer club recorded the data during its daily routine and shared it with the researchers involved in this study through a Non-Disclosure Agreement only for research purposes. The owner of the data remains the soccer club that has the right to choose which information, results, and data can be made publicly available.

Data-driven clustering by blood sample features

Eighty-nine blood samples—3.30 (1.41) blood samples per player—were recorded through the two soccer seasons. All venous blood samples were taken in the early morning (around 8 AM) in an antecubital vein in a seated position.

The players were asked to fast from the previous evening. 10 ml of blood was collected in vacutainer tubes, using an anticoagulant. The freshly drawn blood was immediately centrifuged at 3000 repetitions per minute (825 g) for 10 min to remove the plasma. Plasma was separated into several aliquots and was rapidly frozen at − 80 °C for later biochemical analysis. Analyses were performed using a coulter blood counter (Model S-plus II, Coulter Electronics Inc., Hialeah, Florida, USA) and Hematocrit (%), Hemoglobin (g/dl), number of red blood cells (cells/microL), ferritin (ng/ml), and sideremia (ug/dl) were obtained from each blood sample. Immunoenzymatic plasma testosterone (ng/ml) and cortisol (nmol/L) measurements were taken with VIDAS testosterone (Ref. 30418) and VIDAS cortisol S (Ref. 30451) commercial test kits (bioMerieux, Carnaxide, Portugal). The T/C ratio was also calculated and expressed in percentage.

A silhouette analysis based on k-means algorithm was performed to detect the best number of clusters to group players in accordance with the blood sample. Silhouette analysis can be used to study the separation distance between the resulting clusters. In particular, it measures how close each point in one cluster is to points in the neighboring clusters (silhouette coefficient). This measure has a range between − 1 and 1. + 1 indicates that the sample is far away from the neighboring clusters, 0 refers to a sample that is very close to the decision boundary between two neighboring clusters, and negative values indicate that those samples might have been assigned to the wrong cluster. In particular, the Silhouette Coefficient was calculated using the mean intra-cluster distance (*a*) and the mean nearest-cluster distance (*b*) for each sample. The Silhouette Coefficient for a sample is *(b − a)/max(a, b)*. To clarify, *b* is the distance between a sample and the nearest cluster that the sample is not a part of. The best value is 1 and the worst value is − 1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar. The mean of the silhouette values for all the samples provides an index of cluster goodness.

After detecting the best number of clusters, an unpaired t test analysis was conducted in each blood sample feature to detect differences between blood sample groups.

Injury forecaster

Data

GPS data Players wore a Global Position System (K-GPS 10 Hz, K-Sport International, Italy) [25] during each training or match to obtain external workloads data of each session. Seventeen workload features were used (see Table 1 for more details about the GPS variables). All of these features were pre-processed to obtain metrics describing the Acute (exponential weighted moving average of the previous 7 days) and Chronic (exponential weighted moving average of the previous 28 days) workloads. Moreover, the ratio between Acute and Chronic features (ACWR) for each GPS variable was computed to obtain information about the intensity of the

Table 1 Workload features' description

GPS features	Description
T (s)	Duration of the training session in seconds
D (m)	Distance in meter covered during the training session
D_{MPHI} (m)	Distance in meter covered at high-metabolic intensity (> 20 W/Kg)
D_{SHI} (m)	Distance in meter covered at high intensity speed (> 16 km/h)
D_{AccHI} (m)	Distance in meter covered with high-intensity acceleration (> 2 m/s^2)
D_{DecHI} (m)	Distance in meter covered with high-intensity deceleration (> 2 m/s^2)
$\mathrm{ND}_{\mathrm{MPI}}$ (n)	Number of events with a metabolic power between 0 and 5 W/Kg
ED (%)	Percentage of equivalent distance, i.e., the distance that the athlete would have covered at a constant speed using the energy consumed when the metabolic power is greater than 25.5 W/Kg
AI (%)	Percentage of anaerobic index, i.e., the ratio between the energy expenditure higher than 21 km/h and the total energy expenditure
AMP (W/Kg)	Average metabolic power during the training session
EEE (Kj/Kg)	Estimating energy expenditure during the training session
$\mathrm{EEE}_{\mathrm{A4}}$ (Kj/Kg)	Estimating energy expenditure performed at moderate-acceleration intensity (2–3 m/s^2)
$\mathrm{EEE}_{\mathrm{A5}}$ (Kj/Kg)	Estimating energy expenditure performed at high-acceleration intensity (3–4 m/s2)
$\mathrm{EEE}_{\mathrm{S4}}$ (Kj/Kg)	Estimating energy expenditure performed at moderate-speed intensity (16–21 km/h)
$\mathrm{EEE}_{\mathrm{S5}}$ (Kj/Kg)	Estimating energy expenditure performed at high-speed intensity (21–24 km/h)
$\mathrm{EEE}_{\mathrm{MP4}}$ (Kj/Kg)	Estimating energy expenditure performed at moderate metabolic power (10–20 W/Kg)
$\mathrm{EEE}_{\mathrm{MP5}}$ (Kj/Kg)	Estimating energy expenditure performed at high metabolic power (20–40 W/Kg)

recent workloads (acute workload) in relation to the one that the player is used to perform (chronic workload). ACWR values higher than 1 indicate that a player performs a higher workload in the past week compared to the past month. In particular, values higher than 1.5 result in a high risk of injury (over-training status), while values between 1 and 1.5 indicate the optimal training zone [13, 19, 20]. Differently, values lower than 1 indicate under-training status [13, 19, 20]. Hence, in total, each training vector is composed of 68 workload features (17 features for each aggregation method, i.e., Daily, Acute, Chronic, and ACWR).

Blood sample groups Based on the machine learning approach described in "Data-driven clustering by blood sample features", all the players were grouped in accordance with the individual's blood sample features. However, in train and test scenarios (see "Machine learning approach"), if the players performed 2 or more blood sample tests during the soccer season, the players were grouped in accordance with the most frequent blood sample class (0 = high blood sample group; 1 = low blood sample group). Players with an equal number of classes during the season were set as High groups. Differently, in the real scenario (see "Machine learning approach"), the blood sample profile for each player was re-definite every time the players performed a new blood sample test.

Injury label The club's medical staff recorded 28 non-contact injuries during soccer seasons. A non-contact injury was defined as any tissue damage sustained by a player that causes an absence in physical activities for at least the day after the day of the onset [22, 26]. To predict future injuries, the training session examples of the previous 7 days were labeled as injury (a session with a high risk of injury). The days when the injuries occurred were deleted from the dataset due to the fact that the aim of this study is to predict players that will get injured in the next few days.

Moreover, to take into consideration the individuals' injury history (previous injury), the exponential weighted moving average of the past 28 days was computed on the injury label time series. The higher the previous injury index was, the closer the previous injury was. These values were used as input in the machine learning model. This metric was found to be an important index for injury prediction in the previous paper [7, 22, 27, 28].

Machine learning approach

In this section, the description of the two approaches that were used to evaluate the prediction ability of the machine learning algorithms in this study was provided. In the first approach, the dataset was randomly split into two parts. In the first one, the predictive models were trained in the first part, while they were tested in the second part (train and test approach). The second approach simulated the real scenario by training and testing the machine learning models as the season went by (real scenario approach). Moreover, a description of the predictive metrics to evaluate the predictive performance of the machine learning models was provided in the predictive performance metrics paragraph. Finally, the way used to interpret the decision-making process was described in the last paragraph of this section.

Train and test approach Multi-dimensional models were developed to predict the risk of injuries in the next 7 days. In particular, three different approaches was set: (i) "endorse groups" model, i.e., the models were trained for high and low blood sample groups separately; (ii) "blood group as variable", i.e., a variable indicating the blood sample groups of each player was used as predictive features; (iii) "No-split", i.e., no information about players' blood sample profile was provided. 70% of data were used to train the models, while in the remaining 30% of the dataset, they were tested. The dataset was split into train and test sets in accordance with the distribution of the injury and no-injury examples (stratified approach). Moreover, to solve the problem of data unbalancing, the injury class was oversampled in the train set using the adaptive synthetic sampling approach (ADASYN). The ADASYN algorithm generates examples of the minority class permitting to equalize the distribution of classes and reducing the learning bias. Moreover, to reduce the feature dimension space by selecting the most important features and consequently the risk of overfitting, a Recursive Feature Elimination with Cross-Validation (RFECV) approach was performed in the oversampled train set. Actually, RFECV selects a subset of features producing the maximum score on the validation data. Two machine learning models were trained and tested in this study: (i) Decision Tree classifier (DT) and (ii) Gradient Boosting classifier (XGB). Additionally, to validate the prediction ability of these classifiers, a baseline classifier (Dummy) that randomly assigns a class to an example by respecting the distribution of classes was built.

Real scenario approach This evaluation will permit us to assess the predictive performance of the machine learning algorithms in an evolutive scenario re-training the models by inserting new injury and no-injury examples as the season go by. Let assume that a soccer club has data until week n and it wants to detect the risk of injuries of the next week $n+1$. The machine learning models were trained over the train set (i.e., until week n) and then tested by predicting the injury class on new data recorded in week $n+1$. Every time that the algorithm is retrained, the data were pre-processed to oversample the minority class using ADASYN approach and the feature selection process (RFECV) was used to reduce the feature dimension space selecting the most important features for injury prediction in accordance with the season period. Two machine learning models (i.e., DT and XGB) and one baseline classifier (Dummy) were trained and tested

week-by-week. To assess the advantage that profiling players in accordance with blood samples could provide on injury prediction, the models were trained and tested on two different datasets: (i) with blood group as variable and (ii) with no blood sample group information (No-Split).

Predictive performance metrics Precision, recall, and f1-scores were computed to assess the model goodness. Actually, precision (specificity) is the ratio of correctly predicted positive observations to the total predicted positive observations, while recall (sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. Additionally, f1-score is the weighted mean of precision and recall. Finally, accuracy is the ratio of correctly predicted observations to the total observations.

Machine learning models' interpretation To globally and locally explain the decision-making process that the machine learning models make to predict injuries, SHapley Additive exPlanations (SHAP) values were computed (http://scikit-learn.org/imbalanced-learn). It allows exploring the relationships between predictive variables and injury risk assigning to each feature importance the permits to detect the influence of each feature to the final prediction. Moreover, SHAP permits evaluating how much each predictor contributes, either positively or negatively, to the target variable. Understanding why a model makes a certain prediction can help the team's staff to evaluate the reason underlying the model's decisions and consequently change the training program in accordance with players' demands.

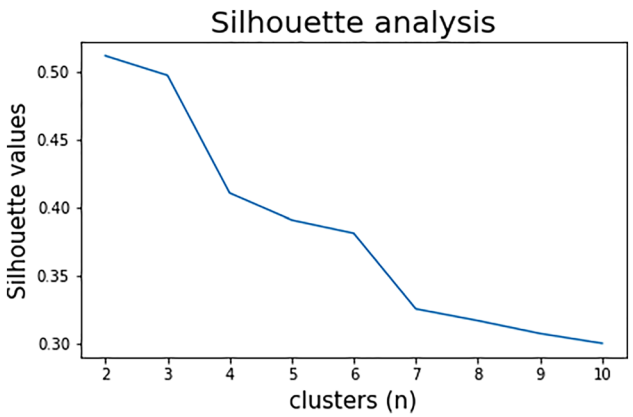

Fig. 1 Silhouette analysis

Results

Blood sample groups

K-means algorithm [29] was used to search for the optimal number of clusters according to silhouette score. Figure 1 shows the silhouette scores associated with each number of clusters selected (from 2 to 10 groups). This figure shows that the best value for silhouette was obtained by splitting players into two groups. The blood sample profiles were labeled as High and Low. Table 2 provides the descriptive statistics of the two groups and the statistical difference between the High and Low groups. The High group showed statistically higher Hematocrit ($p < 0.001$), Hemoglobin ($p < 0.001$), number of red blood cells ($p < 0.001$), and testosterone ($p = 0.02$), but lower ferritin ($p = 0.03$) compared to Low group. Moreover, similar level of cortisol ($p = 0.15$), sideremia ($p = 0.64$), and T/C ($p = 0.63$) were detected.

Injury forecaster

Train and test approach

As shown in Table 3, XGB is the best machine learning model to predict injuries in all the datasets. In particular, the higher prediction performance was detected in the model trained on the "blood group as variable" dataset (f1-score of the injury class = 63%). Similar results were detected when XGB is trained in players with different blood sample profiles (endorse group: f1-score of the injury class = 58%). Providing information about blood sample profiles at the machine learning model permitted to increase the injuries prediction ability of about 15% (i.e., precision and recall increase of about 18% and 4%, respectively) compared to the

Table 2 Descriptive statistics and statistical differences for and between the blood sample groups (i.e., High and Low)

Features	High	Low	*p* value
Hematocrit (%)	46.10 (1.99)	42.27 (2.06)	< 0.001***
Hemoglobin (g/dl)	15.64 (0.71)	14.44 (0.76)	< 0.001***
Red blood cells (cells/microL	5251.92 (163.90)	4684.60 (190.60)	< 0.001***
Ferritin (ng/ml)	110.67 (54.99)	139.91 (68.19)	0.03*
Testosterone (ng/ml)	6.45 (1.44)	5.62 (1.68)	0.02*
Cortisol (nmol/L)	502.69 (81.60)	471.75 (117.21)	0.15
Sideremia (ug/dL)	126.52 (78.75)	112.24 (40.93)	0.64
T/C	0.013 (0.004)	0.014 (0.011)	0.63

Statistical significance: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 3 Train and test approach. Performance of the machine learning models indifferent blood sample groups

Blood sample group	Algorithm	Classes	Precision	Recall	F1-score
High	DT	No-injury	0.97	0.81	0.88
		Injury	0.19	0.68	0.30
	XGB	No-injury	0.96	0.94	0.95
		Injury	0.33	0.42	0.37
	Dummy	No-injury	0.92	0.50	0.65
		Injury	0.05	0.37	0.08
Low	DT	No-injury	0.98	0.92	0.95
		Injury	0.44	0.73	0.55
	XGB	No-injury	0.98	0.95	0.96
		Injury	0.56	0.80	0.66
	Dummy	No-injury	0.95	0.52	0.67
		Injury	0.11	0.70	0.18
Endorse	DT	No-injury	0.98	0.87	0.92
		Injury	0.30	0.71	0.42
	XGB	No-injury	0.97	0.94	0.96
		Injury	0.48	0.65	0.58
	Dummy	No-injury	0.94	0.51	0.66
		Injury	0.08	0.57	0.14
Blood group as variable	DT	No-injury	0.98	0.84	0.90
		Injury	0.20	0.68	0.31
	XGB	No-injury	0.97	0.98	0.98
		Injury	0.58	0.65	0.63
	Dummy	No-injury	0.94	0.49	0.65
		Injury	0.05	0.45	0.09
No-split	DT	No-injury	0.97	0.78	0.86
		Injury	0.14	0.63	0.23
	XGB	No-injury	0.98	0.95	0.96
		Injury	0.40	0.61	0.48
	Dummy	No-injury	0.94	0.49	0.65
		Injury	0.05	0.45	0.09

results obtained from the "No-Split" dataset. All the models in all the datasets were valid to predict the injury due to the fact that higher predictive performance was detected with XGB and DTC algorithms compared to the Dummy one.

Figure 2 shows the feature importance for each model trained. Different features were selected in High and Low blood sample groups, indicating that the response to external stimuli was different between players with different blood sample profiles. Differently, the most important feature without splitting the players in accordance with blood sample profile ("No-Split" dataset) was the mix of the workloads variables extracted from the High and Low blood sample groups dataset. In the model trained on the "Blood group as variable" dataset, the most important feature was the blood sample group and the other features were a mix between High and Low groups features. This result corroborates the fact that the information derived by the blood sample profile (indices of health and fit status) is an important factor affecting the players' injury risk.

Real scenario approach

Table 4 and Fig. 3 show that profiling soccer players in accordance with their blood sample improve the prediction ability throughout the soccer season. In particular, higher precision, recall, and f1-score were detected in "Blood group as variable" compared to "No-split" at the end of the soccer season (i.e., increased prediction ability of about 5%, 3%, and 4%, respectively; Table 4). Additionally, the higher prediction performance detected in XGB compared to Dummy classifier validated the fact that this model is able to accurately distinguish between players with different risk of injury (Table 4). Figure 3 shows a higher predictive performance of XGB compared to DT and Dummy classifiers throughout the entire soccer season. Additionally, Fig. 4 provides the influence of each feature on injury prediction week-by-week. To be noticed, the influence of the predictive features changed in accordance with the season period. This aspect corroborates the fact that the training workloads differently affect the risk of injury in different parts of the soccer season based on teams' training schedule and players' physiological demand.

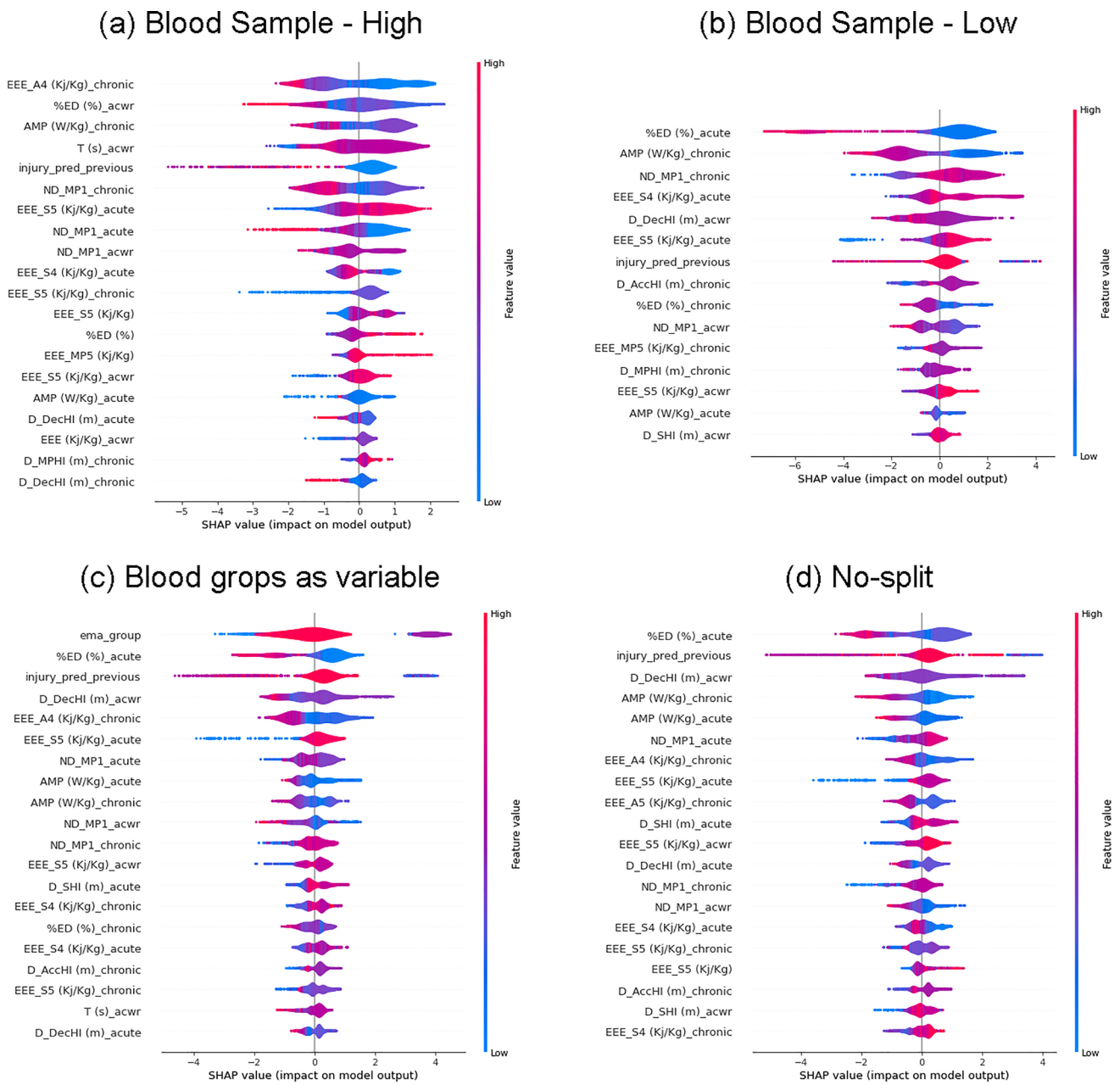

Fig. 2 Violin plots of the SHAP values computed for each feature in the XGB model trained on different datasets. The colors vary from blue (low feature value) and red (high feature value). The SHAP values indicate the influence of each point on injury risk. Negative SHAP values indicate that a specific feature value reduces the injury risk, while positive ones increase the risk. For example, a blue dot with a negative SHAP value indicates that the lower the feature value is the lower is the risk of injury. The sum of these influences indicates the risk of injury

Table 4 Real scenario approach. Performance of the machine learning models indifferent blood sample groups

Blood sample group	Algorithm	Classes	Precision	Recall	F1-score
Blood group as variable	DT	No-injury	0.95	0.91	0.93
		Injury	0.20	0.34	0.25
	XGB	No-injury	0.96	0.94	0.95
		Injury	0.32	0.42	0.36
	Dummy	No-injury	0.94	0.94	0.94
		Injury	0.07	0.07	0.07
No-split	DT	No-injury	0.92	0.90	0.93
		Injury	0.19	0.35	0.25
	XGB	No-injury	0.96	0.93	0.94
		Injury	0.27	0.39	0.32
	Dummy	No-injury	0.94	0.94	0.94
		Injury	0.08	0.07	0.07

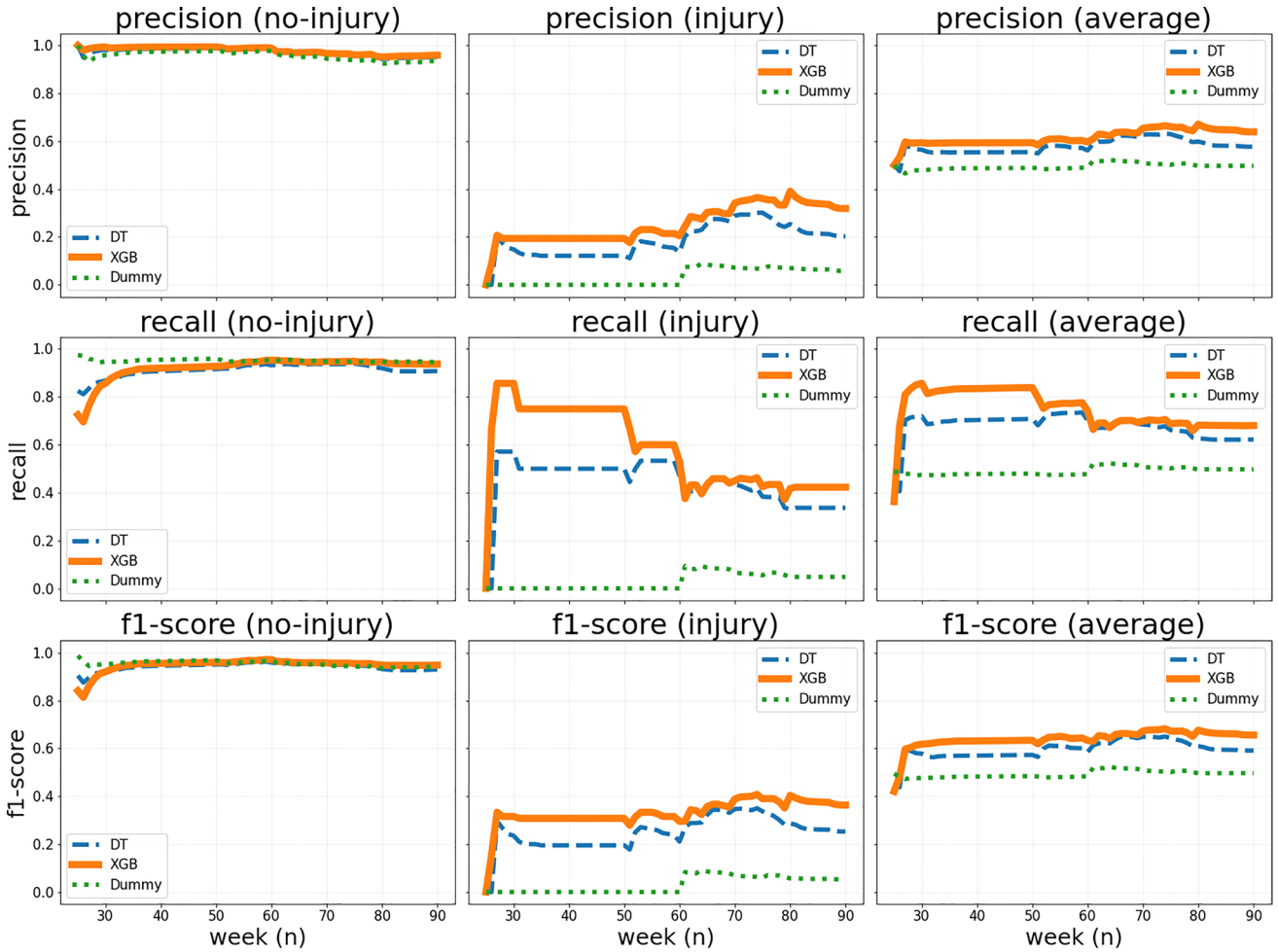

Fig. 3 Real scenario prediction performance

Discussion

Profiling the players in accordance with blood sample analysis helps to personalize the machine learning model increasing its ability to detect players' risk of injury. As a matter of fact, in the train and test scenario, providing information about blood samples permitted to reach an accuracy (f1-score) of about 63% in "Blood group as

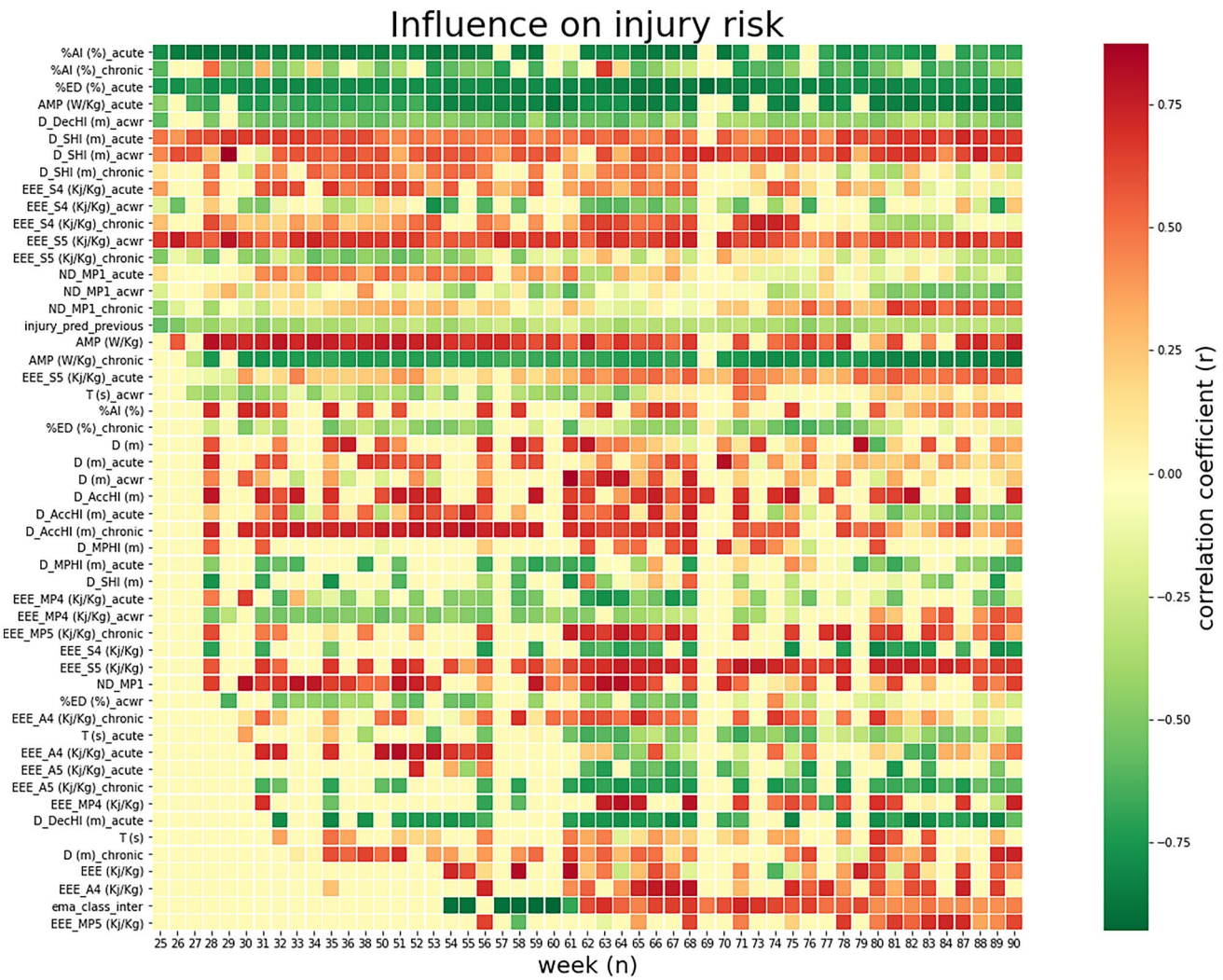

Fig. 4 Heatmap of the influence of each feature in the real scenario. Each column refers to a single week, while the columns refer to the input features. The darker the red is the higher the correlation between SHAP and features values is. Positive relationships indicate that the higher the values of a specific feature is, the higher the risk of injury is. Otherwise, for a negative relationship (the intensity of the green indicates the strength of the relationship)

variable" increasing the accuracy of injuries prediction of about 15% compared to "No-Split" (Table 3). Actually, the injury class precision increases of about 18% (precision in injury class for XGB in "Blood group as variable" and in "No-split" dataset is equal to 40% and 58%, respectively), while the recall was increased of about 4% (recall in injury class for XGB in "Blood group as variable" and in "No-split" dataset is equal to 65% and 61%, respectively). Similarly, in the real scenario, the highest predictive performance was detected by the XGB model in the "Blood group as variable" dataset (Table 4 and Fig. 3). These results corroborate the fact that blood sample variables permit to better assess the players status and consequently permit to personalize the algorithm decision-making rules in accordance with players' needs. Furthermore, Fig. 4 shows that the influence of the workload's features on injury prediction varies week-by-week, suggesting that the players differently respond to external stimuli in different parts of the soccer season in accordance with their physiological needs of that period of time. Hence, adding information about the players' wellness status (e.g., strength, balance, motor skills, and sleep quality) could increase more and more the injury prediction ability due to the fact that machine learning could learn from an increasingly comprehensive view of the player's status. Future works are needed to understand which information derived from different tests (e.g., physical and psychological) is the most sensitive to discriminate against players with a high risk of injury.

In this study, it was found that it is possible to split players in accordance with blood sample analysis into two main groups by a data-driven approach. In particular, hematocrit, hemoglobin, number of red blood cells, ferritin, and testosterone were the main blood sample features that show a statistical difference between the two main groups (Table 2). Actually, these features were found to be linked with aerobic capacity and over-training syndrome [21, 24]. In particular, the reduction of hematocrit, hemoglobin, and number of red blood cells is related to a high external workload performed on the previous days resulting in high physical stress [23]. Actually, subjects with the higher hematocrits (> 44.6%) were frequently overtrained resulting in iron deficiency and increased blood viscosity, while players with low hematocrit (< 40%) were associated with a higher aerobic capacity [21]. Hence, due to the fact that players in the two blood sample profiles show different characteristics, it is plausible to suppose that the two groups of players had different physiological demands resulting in a different response to external stimuli. As a matter of fact, Fig. 2a and b shows that the external workload features allowing to discriminate players that will get injured in the next week were different or shows different importance between High and Low blood sample groups. Moreover, to corroborate the fact that grouping players in accordance with blood samples are useful to increase the prediction ability, Fig. 2c shows that the most important feature in the "Blood sample groups as variable" dataset was the binary variable that provides the information about the player profile. Furthermore, the important features extracted from the "Blood sample groups as variable" dataset (Fig. 2c) were a mix of the ones obtained analyzing the "Blood Sample—High" and "Blood Sample—Low" datasets independently (Fig. 2a and b, respectively). Moreover, similar predictive performance was detected in "endorse groups" and "Blood sample groups as variable" models (Table 3), indicating that providing information about blood sample profile as a categorical feature allows to accurately creation rules in accordance with individuals' player profiles.

To be noted that the results of this study are valid only for the soccer team that is analyzed. This could be considered both as a limitation and strength of the approach proposed in this study, because it is not possible to generalize the predictive results and feature importance, because it is personalized on this soccer team and different periods of the soccer season. Different players, training programs, and soccer season demands could affect the players' physiological demands and, consequently, the rules and features that permit to predict injuries. Future works are needed to assess if this approach and the results obtained are realizable for all the soccer teams. Another limitation of this study is the low number of blood samples per player recorded during the soccer season. Even if only 3.3 blood samples per player were recorded during the entire season, the players' blood sample profiles permitted to increase the prediction ability of the machine learning models compared to the one trained using only external workload features. It is presumably that if the players will be often profiled as the season goes by, the player's status evaluation will be more accurate and, consequently, the injury prediction goodness will increase.

The results of this study can help coaches and athletic trainers to improve the decision-making process when scheduling the training program keeping in mind their players' biomedical status. Actually, the prediction of the players' wellness status provides important insight on individual psychophysiological responses to training allowing to maximize the training effect while reducing unwelcome detrimental effects associated with poor readiness.

Conclusion

Blood sample analysis is a proxy of the health status of the soccer players that allows profiling players and personalizing the rules that predict the individual injury risk. Field experts in soccer clubs should not only monitor the external workloads to assess the status of the players, but additional information derived from individuals' characteristics could help to have a complete overview of the players' well-being enabling a better training schedule, maximizing the training effect and minimizing the risk of injuries.

Author contribution All authors contributed to the study conception and design. CF: performed material preparation and data collection. AR: performed formal analysis and investigation. AR: wrote the first draft of the manuscript and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement. This work is supported by the European Community's H2020 Program under the funding scheme H2020-INFRAIA-2019-1 Research Infrastructures grant agreement 871042, www.sobigdata.eu, SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study.

Availability of data and materials All the relevant data are provided into the text. The owner of the data is an elite soccer club in Italy which wants to remain anonymous and did not give the permission to make the original data publicly available. The club has the right to choose which information, results, and data can be made public and has granted the access to these data to the authors only for research aims.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest.

Ethics approval Not applicable. The soccer club recorded the data during its daily routine and shared it with the researchers involved in this study through a Non-Disclosure Agreement only for research purposes.

Consent to participate Before starting the data recording, the players sign in the informed consent and the consent to the processing of personal information with the soccer club.

References

1. Hurley OA (2016) Impact of player injuries on teams' mental states, and subsequent performances, at the rugby world cup 2015. Front Psychol. https://doi.org/10.3389/fpsyg.2016.00807
2. Hägglund M, Waldén M, Magnusson H et al (2013) Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA champions league injury study. Br J Sports Med 47:738–742. https://doi.org/10.1136/bjsports-2013-092215
3. Seow D, Graham I, Massey A (2020) Prediction models for musculoskeletal injuries in professional sporting activities: a systematic review. Transl Sports Med 3:505–517. https://doi.org/10.1002/tsm2.181
4. Van Eetvelde H, Mendonça LD, Ley C et al (2021) Machine learning methods in sport injury prediction and prevention: a systematic review. J Exp Orthop 8:27. https://doi.org/10.1186/s40634-021-00346-x
5. Meeuwisse WH, Tyreman H, Hagel B, Emery C (2007) A dynamic model of etiology in sport injury: the recursive nature of risk and causation. Clin J Sport Med 17:215–219. https://doi.org/10.1097/JSM.0b013e3180592a48
6. Bittencourt NFN, Meeuwisse WH, Mendonça LD et al (2016) Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concept. Br J Sports Med 50:1309–1314. https://doi.org/10.1136/bjsports-2015-095850
7. Rossi A, Pappalardo L, Cintia P et al (2018) Effective injury forecasting in soccer with GPS training data and machine learning. PLoS ONE 13:e0201264. https://doi.org/10.1371/journal.pone.0201264
8. López-Valenciano A, Ayala F, JosM P et al (2018) A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms. Med Sci Sports Exerc 50:915–927. https://doi.org/10.1249/MSS.0000000000001535
9. Ruddy JD, Shield AJ, Maniar N et al (2018) Predictive modeling of hamstring strain injuries in elite Australian footballers. Med Sci Sports Exerc 50:906–914. https://doi.org/10.1249/MSS.0000000000001527
10. Ayala F, López-Valenciano A, Gámez Martín JA et al (2019) A preventive model for hamstring injuries in professional soccer: learning algorithms. Int J Sports Med 40:344–353. https://doi.org/10.1055/a-0826-1955
11. Carbuhn AF, Sanchez Z, Fry AC et al (2020) A simplified prediction model for lower extremity long bone stress injuries in male endurance running athletes. Clin J Sport Med 30:e124–e126. https://doi.org/10.1097/JSM.0000000000000661
12. Connaboy C, Eagle SR, Johnson CD et al (2019) Using machine learning to predict lower-extremity injury in US special forces. Med Sci Sports Exerc 51:1073–1079. https://doi.org/10.1249/MSS.0000000000001881
13. Gabbett TJ (2010) The development and application of an injury prediction model for noncontact, soft-tissue injuries in elite collision sport athletes. J Strength Cond Res 24:2593–2603. https://doi.org/10.1519/JSC.0b013e3181f19da4
14. Carey DL, Crossley KM, Whiteley R et al (2018) Modeling training loads and injuries: the dangers of discretization. Med Sci Sports Exerc 50:2267–2276. https://doi.org/10.1249/MSS.0000000000001685
15. Talukder H, Vincent T, Foster G, Hu C, Huerta J, Kumar A et al (2016) Preventing in-game injuries for NBA players. In: MIT Sloan Analytics Conference, Boston
16. Rossi A, Perri E, Trecroci A et al (2016) Characterization of in-season elite football trainings by GPS features: the identity card of a short-term football training cycle. In: 2016 IEEE 16th International conference on data mining workshops (ICDMW). pp 160–166
17. Rossi A, Perri E, Pappalardo L et al (2019) Relationship between External and internal workloads in elite soccer players: comparison between rate of perceived exertion and training load. Appl Sci 9:5174. https://doi.org/10.3390/app9235174
18. Rossi A, Perri E, Trecroci A et al (2017) GPS data reflect players' internal load in soccer. In: 2017 IEEE international conference on data mining workshops (ICDMW). pp 890–893
19. Murray NB, Gabbett TJ, Townshend AD, Blanch P (2017) Calculating acute:chronic workload ratios using exponentially weighted moving averages provides a more sensitive indicator of injury likelihood than rolling averages. Br J Sports Med 51:749–754. https://doi.org/10.1136/bjsports-2016-097152
20. Hulin BT, Gabbett TJ, Blanch P et al (2014) Spikes in acute workload are associated with increased injury risk in elite cricket fast bowlers. Br J Sports Med 48:708–712. https://doi.org/10.1136/bjsports-2013-092524
21. Gaudard A, Varlet-Marie E, Bressolle F et al (2003) Hemorheological correlates of fitness and unfitness in athletes: moving beyond the apparent "paradox of hematocrit"? Clin Hemorheol Microcirc 28:161–173
22. Rossi A, Pappalardo L, Cintia P (2022) A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer. Sports 10:5. https://doi.org/10.3390/sports10010005
23. Schumacher YO, Grathwohl D, Barturen JM et al (2000) Haemoglobin, haematocrit and red blood cell indices in elite cyclists. Are the control values for blood testing valid? Int J Sports Med 21:380–385. https://doi.org/10.1055/s-2000-3785
24. Brun JF, Bouchahda C, Chaze D et al (2000) The paradox of hematocrit in exercise physiology: which is the "normal" range from an hemorheologist's viewpoint? Clin Hemorheol Microcirc 22:287–303
25. Rampinini E, Alberti G, Fiorenza M et al (2015) Accuracy of GPS devices for measuring high-intensity running in field-based team sports. Int J Sports Med 36:49–53. https://doi.org/10.1055/s-0034-1385866
26. Hägglund M, Waldén M, Bahr R, Ekstrand J (2005) Methods for epidemiological study of injuries to professional football players: developing the UEFA model. Br J Sports Med 39:340–346. https://doi.org/10.1136/bjsm.2005.018267

27. Ekstrand J, Gillquist J (1983) Soccer injuries and their mechanisms: a prospective study. Med Sci Sports Exerc 15:267–270. https://doi.org/10.1249/00005768-198315030-00014
28. Ekstrand J, Gillquist J (1983) The avoidability of soccer injuries. Int J Sports Med 4:124–128. https://doi.org/10.1055/s-2008-1026025
29. Anderberg MR (2014) Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Academic Press, Cambridge