

A review of machine learning applications in soccer with an emphasis on injury risk

AUTHORS: George P. Nassis^{1,2}, Evert Verhagen³, João Brito⁴, Pedro Figueiredo^{4,5}, Peter Krustrup^{2,6,7}

¹ Physical Education Department, College of Education, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

² Department of Sports Science and Clinical Biomechanics, SDU Sport and Health Sciences Cluster (SHSC), University of Southern Denmark, Odense, Denmark

³ Amsterdam Collaboration on Health and Safety in Sports, Department of Public and Occupational Health, Amsterdam UMC, Amsterdam Movement Sciences, Amsterdam, Netherlands

⁴ Portugal Football School, Portuguese Football Federation, Oeiras, Portugal

⁵ CIDEFES, Universidade Lusófona, Lisboa, Portugal

⁶ Danish Institute for Advanced Study (DIAS), University of Southern Denmark, Odense, Denmark

⁷ Sport and Health Sciences, College of Life and Environmental Sciences, University of Exeter, United Kingdom

ABSTRACT: This narrative review paper aimed to discuss the literature on machine learning applications in soccer with an emphasis on injury risk assessment. A secondary aim was to provide practical tips for the health and performance staff in soccer clubs on how machine learning can provide a competitive advantage. Performance analysis is the area with the majority of research so far. Other domains of soccer science and medicine with machine learning use are injury risk assessment, players' workload and wellness monitoring, movement analysis, players' career trajectory, club performance, and match attendance. Regarding injuries, which is a hot topic, machine learning does not seem to have a high predictive ability at the moment (models specificity ranged from 74.2%-97.7%. sensitivity from 15.2%-55.6% with area under the curve of 0.66–0.83). It seems, though, that machine learning can help to identify the early signs of elevated risk for a musculoskeletal injury. Future research should account for musculoskeletal injuries' dynamic nature for machine learning to provide more meaningful results for practitioners in soccer.

CITATION: Nassis GP, Verhagen E, Brito J et al. A review of machine learning applications in soccer with an emphasis on injury risk. *Biol Sport.* 2023;40(1):233–239.

Received: 2021-06-15; Reviewed: 2021-11-13; Re-submitted: 2022-01-27; Accepted: 2022-02-18; Published: 2022-03-16.

INTRODUCTION

As a branch of artificial intelligence, machine learning is being used in medicine and health sciences for some years [1, 2]. Though, machine learning is still quite new in sports science, sports medicine [3, 4] and particularly in soccer science. Machine learning use is based on the assumption that the computer and the algorithms will learn as we feed them with more data. Following data collection and data cleaning, the algorithms can build relationships among variables either without (unsupervised learning approaches) or with human assistance (supervised), who provide them with the cut-off values for specific variables. Through repeated data feeding, the computers and the algorithms will learn and will be able to identify and select, among the big number of variables, those that account for the dependent variable [3, 5, 6, 7].

One of the first studies with machine learning in soccer was published in 2014 [8]. It was focused on evaluating the technical and tactical abilities of teams during the UEFA EURO2012. Since that time, several papers have been published in soccer, focusing

on machine learning and injury prediction, physical performance prediction, training load and monitoring, players' career trajectory, club performance, and match attendance (Table 1). There are also papers on soccer match results prediction and betting, but these are not within the present review's scope. Despite the increasing number of research papers on machine learning use in soccer, it is still unclear what machine learning does, what it can offer to soccer clubs now and in the future, and how scientists and practitioners can prepare to take advantage of machine learning capabilities. This narrative review aims to bring together information from the past to better articulate what may happen in the future regarding machine learning use in soccer, emphasizing injury risk assessment. Finally, we aim to provide practical tips for the health and performance staff in soccer clubs (i.e., physicians, physiotherapists, coaches, strength and conditioning staff, sports scientists) on how they can pave the way within the club and take competitive advantage of machine learning use. Accordingly, this manuscript is

Corresponding author:

George P. Nassis

Physical Education
Department–College of Education
(CEDU), United Arab Emirates
University, Al Ain, Abu Dhabi,
United Arab Emirates

ORCID:

George P. Nassis
0000-0003-2953-3911

Evert Verhagen
0000-0001-9227-8234

João Brito
0000-0003-1301-1078

Peter Krustrup
0000-0002-1461-9838

Key words:

Machine learning
Soccer injury risk
Data analytics
Big data
Football

TABLE 1. Summary of studies using machine learning approaches exclusively with soccer players as their studied population.

Study	Participants Period of study	Main outcome studied	Main finding	Topic
Ayala et al. [5]	96 male professional players, 1 season	Hamstring strain injuries predicted	The prediction model showed moderate to high accuracy	Injury risk/Injury occurrence prediction
Oliver et al. [18]	355 elite youth players aged 10–18 years old, 1 season	Injuries predicted	The best performing decision tree model provided a specificity of 74.2% and sensitivity of 55.6% with an AUC of 0.663	
Rossi et al. [17]	26 professional male players, 1 season	Non-contact injuries	Machine learning technique can detect around 80% of the injuries with about 50% precision, far better than the baselines and state-of-the-art injury risk estimation techniques	
Rommers et al. [7]	734 players, under-10 to under-15 age categories, 1 season	Non-contact injuries	Machine learning algorithm was able to identify the injured players in the hold-out test sample with 85% precision, 85% recall (sensitivity) and 85% accuracy	
Bongiovanni et al. [35]	16 male under-15 team players, 1 season	Sprint CoD, CMJ & aerobic fitness performance prediction	Anthropometric features were predictors of sprint performance and aerobic fitness, not CoD and CMJ	Physical performance prediction
Campbell et al. [29]	As the authors state “The data encompassed multiple seasons (2013–2018) and was pooled across pre-season and in-season training sessions” without including information on the data size.	Internal (sRPE) and external load (total distance covered)	Very low predictive ability	Players’ monitoring
Geurkink et al. [28]	46 elite male players, 61 training sessions, 913 observations	Predicted sRPE	sRPE was predicted accurately	
Jaspers et al. [27]	38 professional male players, 2 seasons	External and internal training load	More accurate predictions of training Rate of Perceived Exertion from external workload data in combination with pre-session wellness	
Op De Beéck et al. [26]	One soccer team (no information for duration of the study)	Future wellness items (i.e., fatigue, sleep quality, general muscle soreness, stress levels, and mood)	Wellness was predicted based on internal and external workload data. Their effect sizes indicate that the external load and internal load, separately and in combination, do not have sufficient predictive ability	

TABLE 1. Continue.

Study	Participants Period of study	Main outcome studied	Main finding	Topic
Perri et al. [36]	28 sub-elite players, 1 season	Wellness index as predicted by internal training load	Machine learning technique predicted the wellness index based on previous training day internal load	
Dick et al. [37]	Tracking data consisting of a sequence of coordinates of all players and the ball for a set of soccer games	Successful attacks	Proposed an approach to learn valuations of multiplayer positioning using positional data	Performance analysis
Goes et al. [23]	Position tracking data of 118 Dutch Eredivisie matches, containing 12424 attacks	Successful attacks	Identified dynamic formations based on position tracking data, and identified dynamic subgroups for every timeframe in a match	
Link and Hoernig [24]	Data from 60 matches in the German Bundesliga, 1 season	Models for detecting individual and team ball possession based on position data	Match event were detected automatically	
Montoliu et al. [25]	Football videos including two regular league matches played by up to four professional teams	Team activity recognition and analysis	The proposed method performed the team activity recognition task with high accuracy	
Wang [8]	Teams playing in UEFA EURO2012	Technical and tactical analysis of teams	Key performance indicators were identified	
Zago et al. [38]	13 elite female players performed a shuttle run test, wearing 6-axes inertial sensor at the pelvis level	Prediction of turn direction, speed (before/after turn) and the related positive/negative mechanical work	Good predictive ability of the machine learning algorithms	Movement analysis
Barron et al. [39]	966 outfield male players, 1 season	Identify key performance indicators that influence player's career status	Specific technical characteristics correctly predicted 78.8% of the players' league status with a test error of 8.3%	Player's career trajectory
Matesanz et al. [40]	Soccer players' transfer network among 21 European first leagues between the seasons 1996/1997 and 2015/2016	Table rank, UEFA points	Clubs with the highest transfer spending achieve better performance	Club performance
Sahin and Erol [41]	Data of 236 soccer games, 1 season	Predict the attendance demand in European soccer games	A model was proposed to predict attendance with higher accuracy	Match attendance

Note: AUC: Area Under the Curve; CoD: change of direction; CMJ: countermovement jump; sRPE: session rating of perceived exertion

targeting the applied sports scientists and sports medicine staff who have limited knowledge and interest in the technical and computing aspects of machine learning.

Machine learning evolution in soccer

Decision-makers (i.e., club's owners and top managers, directors of health and performance) and team support staff/practitioners constantly strive for accurate and time-efficient methods to predict performance and assess the injury risk (or even predict the occurrence of a musculoskeletal injury). Performance prediction can help develop better training programs and shape effective game strategies, whereas injury risk prediction can protect athletes' health and eventually optimize performance [9].

To achieve these goals, practitioners are using the latest technological advances (e.g., tracking systems such as GPS technology and inertial movement sensors, fatigue and wellness-related biological and psychological markers, screening tests) and the best analytical methods [10, 11, 12]. In the past, regression analysis has been used to assess injury risk [3] and predict sports performance. The problem with this traditional statistical analysis is that it eliminates the variables that are not linearly associated with the dependent variable. This will create a bias when searching for the interactions among variables [5, 6, 7]. The secondary problem is that traditional statistics cannot account for the effect of multiple factors on the dependent variable. For example, to assess the risk of sustaining a musculoskeletal injury, multivariate analysis models take into account the independent effect of factors in isolation, like previous injury or muscle strength, and the potential interactions between a limited number of factors (i.e., 2 to 3 each time). However, injury is a complex phenomenon [13]. Many parameters may account independently or in combination to its occurrence, including a previous injury, muscle strength imbalances, aerobic fitness or workload [14]. To account for the etiological factors' complexity, machine learning algorithms are now being used in high-performance sports for injury prediction [7].

Machine learning analysis usually involves two phases. In phase one, an algorithm is developed based on the actual data. In phase two, this algorithm is applied to another group or a sub-group of the initial sample to assess its performance [1]. A well-known tool to assess the models' performance is the area under the curve (AUC) for the receiver operating characteristics (ROC) curve. The ROC curve is created with a true positive rate on the vertical axis and the false positive rate on the horizontal axis. The true-positive rate is known as the sensitivity, and the false-positive rate is the probability of a false alarm and is calculated as 1-specificity. The higher the AUC, the better the prediction model [5, 6, 7].

Although there is some evidence [5], we are unsure if musculoskeletal injuries can be predicted through performance and screening tests to an accuracy suitable for valid decision making. As stated by Bahr [14], for a screening test to predict injuries, at least two things are needed: a strong relationship between the test outcome

and the injury risk and the need for the test to be examined in relevant populations using appropriate tools. Both criteria have not been met with currently available tests, and it is unlikely that due to the nature of injuries, we ever will have sufficiently accurate tests available [15]. We believe these approaches (i.e., screening tests) and the statistical procedures can help identify early signs of elevated injury risk for the team supporting staff and the players to act in advance. One of the contributions of machine learning may be that it can develop the cut-off values needed from the sample data.

Despite the criticisms, the assessment of injury risk is a hot topic in soccer, as injury occurrence and the associated absence from training and match play are related to lower team ranking and lower club earnings [16]. This might be a reason explaining the growing number of research on machine learning to predict injuries [3]. If not the first study on the topic, one of the first was that by Rossi *et al.* [17]. The authors collected GPS data that described 26 professional male players' workload over one season and constructed injury prediction models. The best machine learning algorithm could detect around 80% of the injuries with 50% precision and an AUC of 0.76. As the authors argued, the algorithm was "far better than the baseline and state-of-the-art injury risk estimation techniques" [17]. It should be mentioned at this point that risk estimation is not equal to risk prediction. People tend to think that predicting involves saying something will happen, whereas estimation is thought to involve how likely something is bound to happen. However, risk prediction is nothing more than calculating a probability.

Recent studies claim that injuries can be predicted by measurements taken in the pre-season. This is a challenging concept, and at least 3 studies were published in the last 2 years on the topic [5, 7, 18]. Traditionally, sports medicine and sports science staff test athletes in the pre-season for various attributes (e.g. body composition, cardiovascular fitness, muscle strength, flexibility, sprinting, and change of direction ability) to analyze the data and assess the injury risk at specific time points in the course of the season [5, 6, 7, 18]. The assumption is that any disadvantage diagnosed with these tests will result in an elevated risk of musculoskeletal injury. Machine learning algorithms have shown to predict injuries with moderate to high accuracy (Table 1). Despite that, we are unsure of what this "moderate to high accuracy" means in clinical terms.

Ayala *et al.* [5] screened 96 players regarding their history of injuries, psychological and neuromuscular risk factors as part of their pre-season assessment. The best model showed moderate to high accuracy with an AUC score of 0.83, a true positives rate of 77.8% and true negatives of 83.8% in predicting hamstring strain injuries. Similarly, Oliver *et al.* [18] performed neuromuscular testing in pre-season and followed 355 youth soccer players for the entire soccer season. The machine learning model showed a specificity of 74.2% and sensitivity of 55.6%, with an AUC of 0.66. Interestingly, logistic regression provided a specificity of 97.7% and a sensitivity of 15.2%, with an AUC curve of 0.66, suggesting a much higher sensitivity with machine learning in predicting injuries than traditional statistics [18].

However, it must be noted that an AUC of 0.66 is not acceptable for an injury prediction tool [18]. Rommers et al. [7] tested 734 young players for strength, flexibility, agility, and endurance in pre-season. Anthropometric data and occurring injuries were also recorded throughout the entire season. The machine learning algorithm predicted injury with a precision of 86% in the training data set [7].

In conclusion, it seems that machine learning techniques can assess the injury risk with moderate to high accuracy based on pre-season screening data. Despite that, we are still unsure what this “moderate to high accuracy” means in clinical terms. Furthermore, it should be mentioned that injury risk estimation is not equally interpreted as injury prediction. Even though fundamentally, in prediction models, one looks for the probability of an outcome, it is easily (and most frequently) wrongly interpreted as the estimation and outcome ‘will’ occur, instead of ‘can’ occur.

Challenges with regards to injury prediction

The problem with the approaches mentioned above is that data collected months before an injury cannot account for the dynamic nature of soccer activities and the changes that may happen to the players. Athletes are exposed to a specific workload that may modify the relation between a parameter and the injury risk [14]. Indeed, it is assumed that the accuracy with machine learning can be improved with data (e.g., workload, readiness, physiological and contextual data like training ground condition, shoe type) collected as close to the event (the injury occurrence) possible. Real-time data on athlete’s physiology (e.g., heart rate, body temperature), mechanical responses (e.g., fatigue-induced alterations in mechanics of movements), and other contextual factors (e.g., pitch condition, shoe-surface properties) could add value and improve the accuracy of machine learning techniques in the future. This will be one of the biggest challenges soon with integrating big data and machine learning applications in a way that adds a competitive advantage in sports.

There are at least two more problems associated with the use of machine learning in soccer injury prediction: 1) the low incidence of specific injuries that prevent algorithms from reaching a higher accuracy [19]; and 2) the uncertainty when applying machine learning algorithms to another setting or using slightly different data collection procedure than the original one. Regarding the first point, in elite soccer, one can expect about 50 injuries per team during a season, a very small number to create prediction models with high accuracy in a single team [19]. The situation becomes more challenging when considering the different types of injuries within a soccer squad. The problem of low cases in most studies is called the imbalanced data-sets problem in data science [5, 6, 7, 20], because one class (in this case, the injured athletes) is underrepresented in the data-set [20].

Machine learning models are usually biased towards the majority class (in this case, the non-injured athletes), which means the algorithms may inaccurately predict the minority class (the injury cases). Experts in the field have suggested technical solutions to this

problem, and the readers may find the technical details elsewhere [5, 6, 7, 20]. However, another potential solution is to collect data from many teams and treat them as one sample. As an example, all teams of a national league could send their data to a common database. Nevertheless, what works in one team may not work in another. As stated before, injury risk is context-specific and essential factors like the team, medical support, and staff communication are crucial elements [21]. Therefore, we need more data specific to the players. Then, we move to the second challenge, which is the uncertainty when applying a machine-learning algorithm to a different setting [5, 6, 7]. It seems that even a slight deviation in the procedure of data collection may affect the outcome. As suggested, if a new test is added or an existing one is executed differently from the original one used for the algorithm development, this could affect the algorithm performance [5, 6, 7].

What can soccer learn from other sports and scientific fields?

Besides injuries, the occurrence of illnesses is of major concern in sports due to potential loss of training and hence the risk of under-performance. An earlier study has developed predictive models for illnesses based on workload data in rugby league players, and this is an interesting dimension with high practical importance [22]. Thirty-two professional rugby league players were recruited, their internal training and match load were recorded, and perceptual well-being ratings were collected for 29 weeks during a rugby league season. A decision-tree model was developed with the risk factors contributing to the self-reported illness and their cut-off values [22]. Overall, a reduction in well-being combined with an increased internal training load were the main contributors to the self-reported illness occurrence. The area under the ROC curve ranged from 46% to 80%, depending on the model [22]. The ROC range may indicate that more work is needed for the machine learning algorithms to provide risk assessments with higher accuracy based on this nature of data.

Machine learning use in other outcome measures in soccer

In soccer, performance analysis is the topic with most studies related to machine learning published so far (Table 1). Most of these studies are related to the automatic detection of match events [23, 24, 25]. The development of algorithms based on machine learning, which will detect match events with higher accuracy faster, is advantageous for high-performance teams who analyze big data (e.g., data of their team and those of the opponents) [24].

An interesting dimension is applying machine learning for predicting a player’s wellness [26]. The authors used machine learning algorithms to predict wellness based on external and internal load data during training sessions. Wellness was calculated every morning from scores on perceived fatigue, sleep quality, general muscle soreness, stress levels, and mood state. Machine learning was applied to develop predictive models for the next day’s wellness score based on the last day’s training load (e.g., external and internal load) and wellness score. Although this is a promising dimension, the analysis

showed that the data did not have sufficient predictive ability for the next day's wellness score [26]. In the same topic, Jasper *et al.* [27] and Geurkink *et al.* [28] have employed machine learning to predict training session's ratings of perceived exertion (sRPE) from external workload data (i.e., GPS-derived). The analyses showed that sRPE could be predicted accurately for sRPE values below 8 [28]. This work is very relevant because coaches may design better training methods using past GPS-derived data to predict sRPE and hence the internal workload of the players with higher accuracy.

If wellness data have a predictive value for subsequent training, practitioners would inform coaches to modify training volume and intensity when necessary. This was the aim of a recent study conducted by Campbell *et al.* [29], who analysed a pool of data captured and stored in an athlete's management system platform. As the authors state, "the data encompassed multiple seasons (2013–2018) and was pooled across pre-season and in-season training sessions" without including information on the data size. This study showed the very low predictive ability of wellness data on internal (s-RPE) and external load (total distance covered) in soccer players [29].

Ethical and technological considerations with the use of machine learning

As with every new technology, there are some ethical considerations regarding the unintended use of machine learning [30]. It is tempting to use existing data that was not directly gathered for machine learning studies. Consent, as such, is not always given for the use of data by the athletes. Practitioners should carefully review the process and ensure their actions obey the law and follow the authorities and organizational policies. For instance, the use of databases containing personal information collected without the consent of the individual(s) is a major issue, and practitioners should stick to the rules. This is extremely important regarding the use of medical records that contain sensitive personal data [31]. Therefore, data protection experts' contribution and approval are highly recommended when using machine learning in soccer.

Another consideration is the use of machine learning for decision-making. This brings management and leadership aspects to the front stage of successful management of risk using machine learning [32]. For instance, how does the supporting team staff deal with false-positive cases communicated to the coach or the true-positives not communicated? What is the acceptable threshold for the identified high-risk cases to be communicated to the coaches and the top management? Education of everyone involved with data analysis and decision-making is therefore of utmost importance [33].

In the future, the predictive models could integrate data related both to intrinsic factors (i.e. demographics, anatomical, physiological, psychological profile) and extrinsic factors (i.e. workload and environmental data). The addition of veracity metrics related to physical testing and workload monitoring could also help. A recent paper reported that data veracity (accuracy, reliability, and quality of data) was found for 54% of tools and 23% of the parameters used in testing and workload monitoring in soccer studies [34].

CONCLUSIONS

In this narrative review paper, we have attempted to shed some light on the use of machine learning in soccer, emphasizing injuries, given practitioners' interest on this topic. Machine learning does not seem to have high predictive ability in every setting. It seems that machine learning can help to identify early signs of elevated risk for a musculoskeletal injury. Future research should account for musculoskeletal injuries' dynamic nature for machine learning to provide more meaningful results. Performance analysis is the area with most research at the moment. Knowledge in this area is growing and is having practical applications in the club setting.

Conflict of interest

The authors declare that they have no competing interests.

REFERENCES

1. Deo RC. Machine learning in medicine. *Circulation*. 2015; 132(20):1920–1930.
2. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med*. 2001; 23(1):89–109.
3. Claudino JG, Capanema DO, de Souza TV, Serrão JC, Machado Pereira AC, Nassis GP. Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. *Sports Med Open*. 2019; 5(1):28.
4. Edouard P, Verhagen E, Navarro L. Machine learning analyses can be of interest to estimate the risk of injury in sports injury and rehabilitation. *Ann Phys Rehabil Med*. 2020; S1877-0657(20)30159-7.
5. Ayala F, López-Valenciano A, Gámez Martín JA, De Ste Croix M, Vera-García FJ, García-Vaquero MDP, Ruiz-Pérez I, Myer GD. A preventive model for hamstring injuries in professional soccer: learning algorithms. *Int J Sports Med*. 2019; 40(5):344–353.
6. López-Valenciano A, Ayala F, Puerta JM, De Ste Croix MBA, Vera-García FJ, Hernández-Sánchez S, Ruiz-Pérez I, Myer GD. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Med Sci Sports Exerc*. 2018;50(5):915–927.
7. Rommers N, Rössler R, Verhagen E, Vandecasteele F, Verstockt S, Vaeyens R, Lenoir M, D'Hondt E, Witvrouw E. A machine learning approach to assess injury risk in elite youth football players. *Med Sci Sports Exerc*. 2020; 52(8):1745–1751.
8. Wang M. Evaluating technical and tactical abilities of football teams in Euro 2012 based on improved information entropy model and SOM neural networks. *Int J Multimedia Ubiquitous Eng*. 2014; 9(11):293–302.
9. Gabbett HT, Windt J, Gabbett TJ. Cost-benefit analysis underlies training decisions in elite sport. *Br J Sports Med*. 2016; 50(21):1291–1292.

10. Gabbett TJ, Nassis GP, Oetter E, Pretorius J, Johnston N, Medina D, Rodas G, Myslinski T, Howells D, Beard A, Ryan A. The athlete monitoring cycle: a practical guide to interpreting and applying training monitoring data. *Br J Sports Med.* 2017; 51(20):1451–1452.
11. Nassis GP, Massey A, Jacobsen P, Brito J, Randers MB, Castagna C, Mohr M, Krstrup P. Elite football of 2030 will not be the same as that of 2020: Preparing players, coaches and support staff for the evolution. *Scand J Med Sci Sports.* 2020; 30(6):962–964.
12. Paul DJ, Nassis GP. Testing strength and power in soccer players: the application of conventional and traditional methods of assessment. *J Strength Cond Res.* 2015; 29(6):1748–1758.
13. Bittencourt NFN, Meeuwisse WH, Mendonça LD, Nettel-Aguirre A, Ocarino JM, Fonseca ST. Complex system approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concepts. *Br J Sports Med.* 2016; 50(21):1309–1314.
14. Bahr R. Why screening tests to predict injury do not work and probably never will...: a critical review. *Br J Sports Med.* 2016; 50(13):776–780.
15. Verhagen E, van Dyk N, Clark N, Shrier I. Do not throw the baby out with the bathwater: screening can identify meaningful risk factors for sports injuries. *Br J Sports Med.* 2018; 52(19):1223–1224.
16. Hägglund M, Waldén M, Magnusson H, Kristenson K, Bengtsson H, Ekstrand J. Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *Br J Sports Med.* 2013; 47(12):738–742.
17. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernández J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One.* 2018; 13(7):e0201264.
18. Oliver JL, Ayala F, De Ste Croix MBA, Lloyd RS, Myer GD, Read PJ. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *J Sci Med Sport.* 2020; 23(11):1044–1048.
19. Ekstrand J, Hägglund M, Waldén M. Injury incidence and injury patterns in professional football: the UEFA injury study. *Br J Sports Med.* 2011; 45(7):553–558.
20. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *EEE Trans. Syst. Man Cybern.* 2012; 42(4):463–484.
21. Ekstrand J, Lundqvist D, Lagerbäck L, Vouillamoz M, Papadimitiou N, Karlsson J. Is there a correlation between coaches' leadership styles and injuries in elite football teams? A study of 36 elite teams in 17 countries. *Br J Sports Med.* 2018; 52(8):527–531.
22. Thornton HR, Delaney JA, Duthie GM, Scott BR, Chivers WJ, Sanctuary CE, Dascombe BJ. Predicting self-reported illness for professional team-sport athletes. *Int J Sports Physiol Perform.* 2016; 11(4):543–550.
23. Goes FR, Brink MS, Elferink-Gemser MT, Kempe M, Lemmink KAPM. The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *J Sports Sci.* 2021; 39(5):523–532.
24. Link D, Hoernig M. Individual ball possession in soccer. *PLoS One.* 2017; 12(7):e0179953.
25. Montoliu R, Martín-Félez R, Torres-Sospedra J, Martínez-Usó A. Team activity recognition in Association Football using a Bag-of-Words-based method. *Hum Mov Sci.* 2015; 41:465–478.
26. Op De Beéck T, Jaspers A, Brink MS, Frencken WGP, Staes F, Davis JJ, Helsen WF. Predicting Future Perceived Wellness in Professional Soccer: The Role of Preceding Load and Wellness. *Int J Sports Physiol Perform.* 2019; 14(8):1074–1080.
27. Jaspers A, De Beéck T, Brink MS, Frencken WGP, Staes F, Davis JJ, Helsen WF. Relationship between the external and internal training load in professional soccer: what can we learn from machine learning? *Int J Sports Physiol Perform.* 2018; 13(5):625–630.
28. Geurkink Y, Vandewiele G, Lievens M, de Turck F, Ongenae F, Matthys SPJ, Boone J, Bourgeois JG. Modeling the Prediction of the Session Rating of Perceived Exertion in Soccer: Unraveling the Puzzle of Predictive Indicators. *Int J Sports Physiol Perform.* 2019; 14(6):841–846.
29. Campbell PG, Stewart IB, Sirotic AC, Drovandi C, Foy BH, Minett GM. Analysing the predictive capacity and dose-response of wellness in load monitoring. *J Sports Sci.* 2021; 39(12):1339–1347.
30. Liaw ST, Liyanage H, Kuziemsky C, Terry AL, Schreiber R, Jonnagaddala J, de Lusignan S. Ethical Use of Electronic Health Record Data and Artificial Intelligence: Recommendations of the Primary Care Informatics Working Group of the International Medical Informatics Association. *Yearb Med.* 2020; 29(1):51–57.
31. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 2019; 20(5):e262–e273.
32. Verhagen E, Mellette J, Konin J, Scott R, Brito J, McCall A. Taking the lead towards healthy performance: the requirement of leadership to elevate the health and performance teams in elite sports. *BMJ Open Sport Exerc Med.* 2020; 6(1):e000834.
33. Nassis GP. Leadership in science and medicine: can you see the gap? *Sci Med Football.* 2017; 1(3):195–196.
34. Claudino JG, Filho CAC, Boullosa D, Lima-Alves A, Carion GR, da Silva Gianoni RL, Guimaraes RS, Ventura FM, Araujo ALC, Rosso SD, Afonso J, Serrao JC. The role of veracity on the load monitoring of professional soccer players: a systematic review in the face of the Big Data Era. *Appl Sci.* 2021; 11:6479.
35. Bongiovanni T, Trecroci A, Cavaggoni L, Rossi A, Perri E, Pasta G, Iaia FM, Alberti G. Importance of anthropometric features to predict physical performance in elite youth soccer: a machine learning approach. *Res Sports Med.* 2020; 23:1–12.
36. Perri E, Simonelli C, Rossi A, Trecroci A, Alberti G, Iaia FM. Relationship Between Wellness Index and Internal Training Load in Soccer: Application of a Machine Learning Model. *Int J Sports Physiol Perform.* 2021; 16(5):695–703.
37. Dick U, Brefeld U. Learning to Rate Player Positioning in Soccer. *Big Data.* 2019; 7(1):71–82.
38. Zago M, Sforza C, Dolci C, Tarabini M, Galli M. Use of Machine Learning and Wearable Sensors to Predict Energetics and Kinematics of Cutting Maneuvers. *Sensors (Basel).* 2019; 19(14):3094.
39. Barron D, Ball G, Robins M, Sunderland C. Artificial neural networks and player recruitment in professional soccer. *PLoS One.* 2018; 13(10):e0205818.
40. Matesanz D, Holzmayer F, Torgler B, Schmidt SL, Ortega GJ. Transfer market activities and sportive performance in European first football leagues: A dynamic network approach. *PLoS One.* 2018; 13(12):e0209362.
41. Şahin M, Erol R. Prediction of Attendance Demand in European Football Games: Comparison of ANFIS, Fuzzy Logic, and ANN. *Comput Intell.* 2018; 7:5714872.