Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis

Christopher Leckey ,[1,2] Nicol van Dyk ,[1,3] Cailbhe Doherty ,[1,4] Aonghus Lawlor ,[4] Eamonn Delahunt [1,5]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bjsports-2024-108576).

[1]School of Public Health Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland
[2]High Performance Unit, Irish Rugby Football Union, Dublin, Dublin, Ireland
[3]Section Sports Medicine, Faculty of Health Sciences, University of Pretoria, Pretoria, Gauteng, South Africa
[4]Insight Centre for Data Analytics, University College Dublin, Dublin, Dublin 4, Ireland
[5]Institute for Sport and Health, University College Dublin, Dublin, Ireland

Correspondence to
Mr Christopher Leckey; christopher.leckey@irfu.ie

Accepted 7 November 2024

To cite: Leckey C, van Dyk N, Doherty C, *et al*. *Br J Sports Med* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjsports-2024-108576

ABSTRACT

Objective This study reviewed the current state of machine learning (ML) research for the prediction of sports-related injuries. It aimed to chart the various approaches used and assess their efficacy, considering factors such as data heterogeneity, model specificity and contextual factors when developing predictive models.
Design Scoping review.
Data sources PubMed, EMBASE, SportDiscus and IEEEXplore.
Results In total, 1241 studies were identified, 58 full texts were screened, and 38 relevant studies were reviewed and charted. Football (soccer) was the most commonly investigated sport. Area under the curve (AUC) was the most common means of model evaluation; it was reported in 71% of studies. In 60% of studies, tree-based solutions provided the highest statistical predictive performance. Random Forest and Extreme Gradient Boosting (XGBoost) were found to provide the highest performance for injury risk prediction. Logistic regression outperformed ML methods in 4 out of 12 studies. Three studies reported model performance of AUC>0.9, yet the clinical relevance is questionable.
Conclusions A variety of different ML models have been applied to the prediction of sports-related injuries. While several studies report strong predictive performance, their clinical utility can be limited, with wide prediction windows or broad definitions of injury. The efficacy of ML is hampered by small datasets and numerous methodological heterogeneities (cohort sizes, definition of injury and dependent variables), which were common across the reviewed studies.

WHAT IS ALREADY KNOWN ON THE TOPIC

⇒ The increased rate of data collection relating to athlete load has led to interest in machine learning (ML) approaches for sports data analysis, including injury risk prediction.
⇒ Prior reviews have examined the application of ML for both performance analysis and injury risk prediction, although no charting of study characteristics were conducted.

WHAT THIS STUDY ADDS

⇒ Random Forest has been tested more than any other ML approach.
⇒ Extreme Gradient Boosting (XGBoost) provided the best performance in each paper in which it was tested.
⇒ Numerous heterogeneities exist, including definition of "injury", granularity of data and scope of prediction windows.
⇒ Lack of clinical relevance hampers the utility of some study findings.

INTRODUCTION

Injuries are commonly incurred by professional, amateur and recreational athletes.[1] They can have a significant impact on the short- and long-term health of individual athletes, as well as team performance. In men's elite-level European football (soccer), lower injury burden and higher match availability are associated with a higher final league ranking.[2] Similarly, injuries are negatively correlated to team success in elite-level rugby union in England.[3] Diminished on-field performance also leads to a loss of potential revenue for sporting organisations.[4] English Premier League teams lose, on average, an estimated amount of £45 million due to injury-related decrements in performance per season.[5]

To inform injury risk mitigation initiatives, sports practitioners and researchers leverage the wealth of player-related data made available through the proliferation of electronic performance and tracking systems. For instance, wearable Global Positioning System (GPS)-based units can provide a multitude of outcome metrics on a player's physical exertions with millisecond granularity.[6] In addition to these GPS-based outcome metrics, scientists collect data on self-perceived wellness scores,[7] ratings of perceived exertion (RPE),[8 9] musculoskeletal screening tests[10] and sleep quality[11] from players on a regular basis.

In an effort to explore the complex interaction of different player-related datasets, sports scientists increasingly apply machine learning (ML) models. As a subset of artificial intelligence, ML could establish previously unknown relationships in complex datasets[12] across a variety of sporting domains, including results prediction,[13] player scouting[14] and tactical team analysis.[15] To address the problem of injury risk prediction in sports, ML models could be a suitable solution due to their ability to harness the complex non-linearity associated with the physiological and biomechanical processes that precede an injury.[16]

These complexities extend to the highly individualised nature of sporting injury, influenced by numerous factors specific to an individual athlete. Given the abundance of player-related data from a multitude of sources, it is important to better understand the ML approaches that have been applied to sports injury risk prediction to date. We might anticipate that ML methods and approaches will be selected based on the available data and the purpose of the analysis, yet we do not know what models are more often included and how their performance is defined.

Therefore, the aim of this scoping review was to synthesise the findings of the published literature on the use of ML to predict sport-related injuries. Objectives were to:

1. Summarise the ML approaches that have been used to predict sport-related injuries.
2. Assess the efficacy of the ML approaches that have been used to predict sport-related injuries.

METHODS

Protocol and reporting

Our review protocol is publicly accessible via the Open Science Framework (https://osf.io/tyu5f/, DOI: 10.17605/OSF.IO/TYU5F). We have reported our review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.[17]

Inclusion criteria

We included studies published in a peer-reviewed journal as full-text English articles that described the use of an ML technique to predict in vivo sport-related injuries.

Exclusion criteria

We excluded studies if: (1) ML models were not used to predict in vivo sport-related injuries; (2) the full-text article published in a peer-reviewed journal was not available in the English language and (3) the article was a review (narrative, systematic and scoping), editorial or conference abstract/paper.

Search Strategy

We conducted an initial, exploratory Google Scholar search (March 2023) for thematically relevant studies. 20 relevant studies were visualised using the *wordcloud* library in the statistical programming language R (R Foundation for Statistical Computing, Vienna, Austria), allowing key search terms to be identified. These were reviewed for relevance, with additional terms included based on the authors' domain knowledge. This process led to the development of the following search strategy:

('sport*' OR 'athlet*' OR 'soccer' OR 'basketball' OR 'rugby' OR 'football' OR 'volleyball' OR 'skating' OR 'handball' OR 'distance running' OR 'hockey') AND ('injury' OR 'injuries' OR 'injured') AND ('artificial intelligence' OR 'bayesian logistic regression' OR 'boosting' OR 'decision tree*' OR 'deep learning' OR 'elastic net' OR 'k-means' OR 'lasso' OR 'learning algorithm*' OR 'machine learning' OR 'naive bayes' OR 'nearest neighbo*' OR 'neural network*' OR 'random forest*' OR 'ridge' OR 'support vector machine*' OR 'XGBoost'). This search strategy was applied across the following four electronic bibliographic databases: MEDLINE via PubMed, EMBASE via Ovid, SportDiscus via EBSCOhost and IEEEXplore. The databases were searched from the date of earliest publication to 20 May 2023.

Selection of sources of evidence

Two reviewers (CL and ED) independently screened the titles, abstracts and, where necessary, full text of all articles identified through the search strategy. Discrepancies in their identification of articles that satisfied the inclusion criteria were iteratively discussed with a third reviewer (NVD) until a consensus was reached.

The data items were initially charted following an iterative process using a customised Microsoft Excel spreadsheet. Collected data included, but was not confined to, publication year, cohort information (volume, age range and gender), study duration and the sporting discipline in which the research was conducted. The location and classification of the injuries being investigated were also collected, as well as the ML algorithm, which was found to yield the highest predictive performance. Information was also collected regarding the dependent variables used during each analysis. Data relating to the stated hypotheses were also collected to investigate the potential existence of 'Hypothesising After the Results are Known' (HARKing).

Synthesis of results

Studies were charted based on common topics and themes. These included the study cohort (ie, elite or amateur, youth or adult, male or female), the type of injuries they sought to predict (ie, contact or non-contact and the location of injury) and the sport in which they were conducted.

Equity, diversity and inclusion statement

The authorship group acknowledges a gender imbalance among the contributing researchers. The group was assembled based on the academic research studies panel, convened for the purposes of CL's progression through his PhD research. This panel was assembled based on everyone's expertise, availability and willingness to engage with the project.

During the data charting, the gender of each study cohort was recorded and reported within the results section of the following review.

RESULTS

Numerical analysis

The results of the search strategy and screening process are illustrated in figure 1. The initial electronic search yielded 1787 potentially relevant studies; 548 of these were duplicates and were removed. Title, abstract and, where necessary, full-text screening of the remaining 1241 articles yielded 37, which satisfied the inclusion criteria. The inter-rater agreement of the two reviewers who independently assessed the eligibility of all the studies was 98%. Following a manual reference list search, one additional article was identified as relevant and included, resulting in the inclusion of a total of 38 studies. A summary of the main features of each study is available in online supplemental table 1).

Publication year and country of origin

Studies were identified from 15 countries (figure 2), with the USA contributing 37% (n=14). Australia, the only Oceanic nation represented, contributed five articles. 10 European Union member states contributed 16 articles, with Spain contributing three articles. Studies by Asian countries, Iran (n=1) and China (n=1) accounted for two articles. No relevant articles were identified from South American or African countries.

Research in the domain of ML for injury risk prediction in sports has grown steadily over the last 10 years (figure 2). The most recent full calendar year included in our review was 2022, which represented the greatest volume of academic output in the field, with 12 articles, an increase of 140% on the previous year. We expect similar growth in the number of publications for 2023, given the number of articles included in the review up to 20 May 2023.

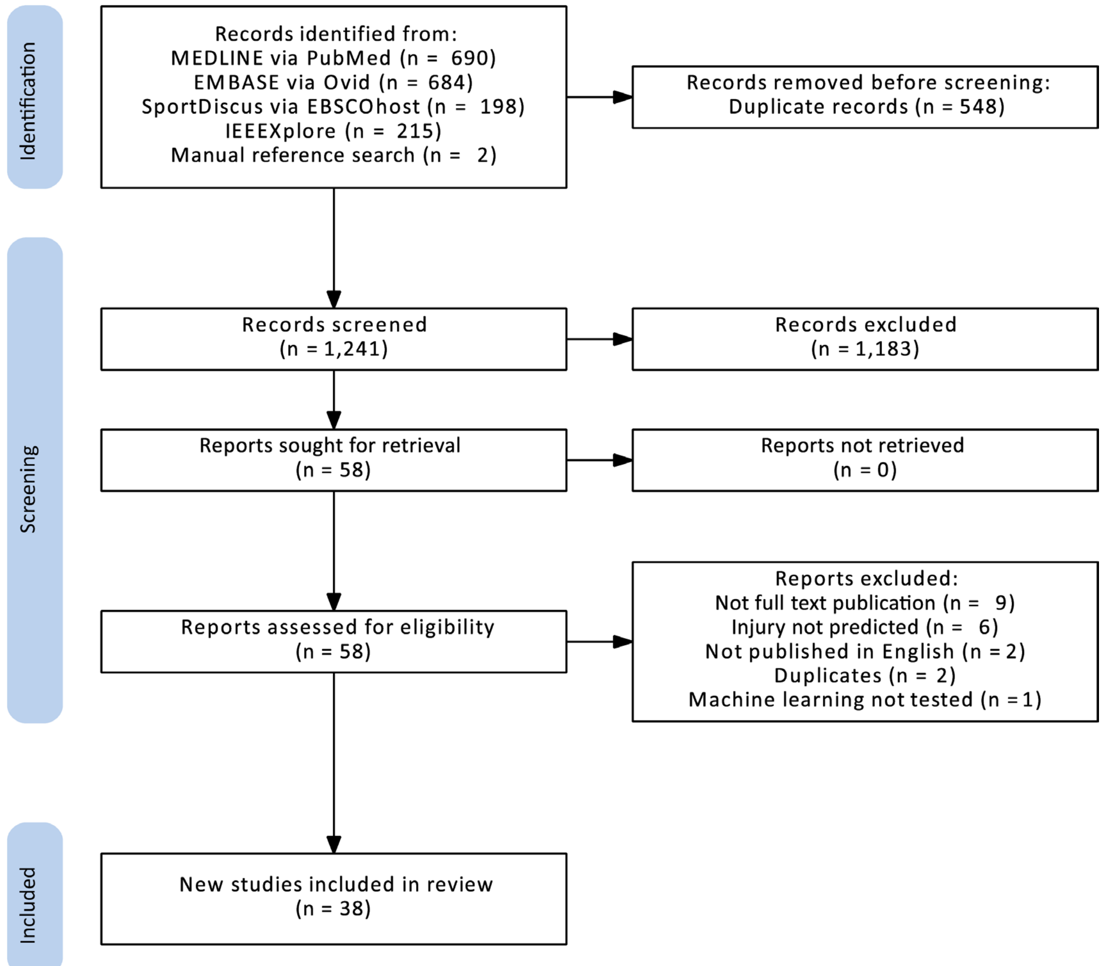

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses Flowchart.

Study cohorts and environments

Nine studies were undertaken in football, while three studies were conducted in both Australian football and basketball, and two studies were conducted in long-distance running. One study was conducted in each of the following sports: rugby league, baseball, speed skating, cricket, ice hockey, American football and futsal. 11 studies included more than one sport and/or military personnel. Three studies did not report in which sport the research was conducted.

Studies including male-only cohorts were most common (47%, n=18). Only 18% of studies included entirely female cohorts (n=7). 12 studies (32%) analysed mixed-gender cohorts, while one study did not report the sex of the participants. In most cases, studies included adult cohorts (n=26, 68%). Eight studies included youth athlete cohorts, and four studies included mixed-age cohorts. Furthermore, 68% of studies were conducted in elite or professional sports settings (n=26), while 12 studies (32%) were conducted in amateur or recreational sports settings.

Cohort sizes within the included studies varied significantly (table 1). The smallest study included 11 participants, while the largest study comprised 15 682 participants. The median cohort size of the included studies was 122 participants. The heterogeneity in cohort sizes equates to a coefficient of variation (CV) of 264%. This variation is further evidenced when considering the number of injuries included in each of the studies upon which the ML models were to be trained (table 1). The heterogeneity in the number of included injuries is large (CV of 299%).

Injury data collection

Injuries were diagnosed and reported by professional medical staff in 30 studies (79%), with four studies (11%) requiring MRI verification of included injuries. Five studies (13%) relied on self-reported survey responses and three studies (8%) used data from publicly available sources.

Nine studies conformed to the consensus statement of Fuller,[18] with regard to the reporting, classification and severity of injuries. Two studies followed the consensus statement of the Union of European Football Association.[19] 18 studies (47%) explicitly required that the injuries used to train the ML models were time-loss injuries. This time period was predominantly defined as a 1-day minimum (n=13). Other studies defined this minimum as two (n=1), five (n=1), seven (n=1), eight (n=1) and nine (n=1) day periods of absence from physical activity (table 2). Nine studies (24%) included non-time-loss injuries within the target variable. Two studies (5%) permitted the inclusion of medical illnesses within the definition of injury.

Contact versus non-contact

18 studies (47%) limited their analyses to non-contact injuries (injuries without any direct contact or collision with another individual).[20] Two studies considered only contact injuries: one relating to concussion and one to dental injuries. In 47% of the studies (n=18), contact and non-contact injuries were considered jointly.

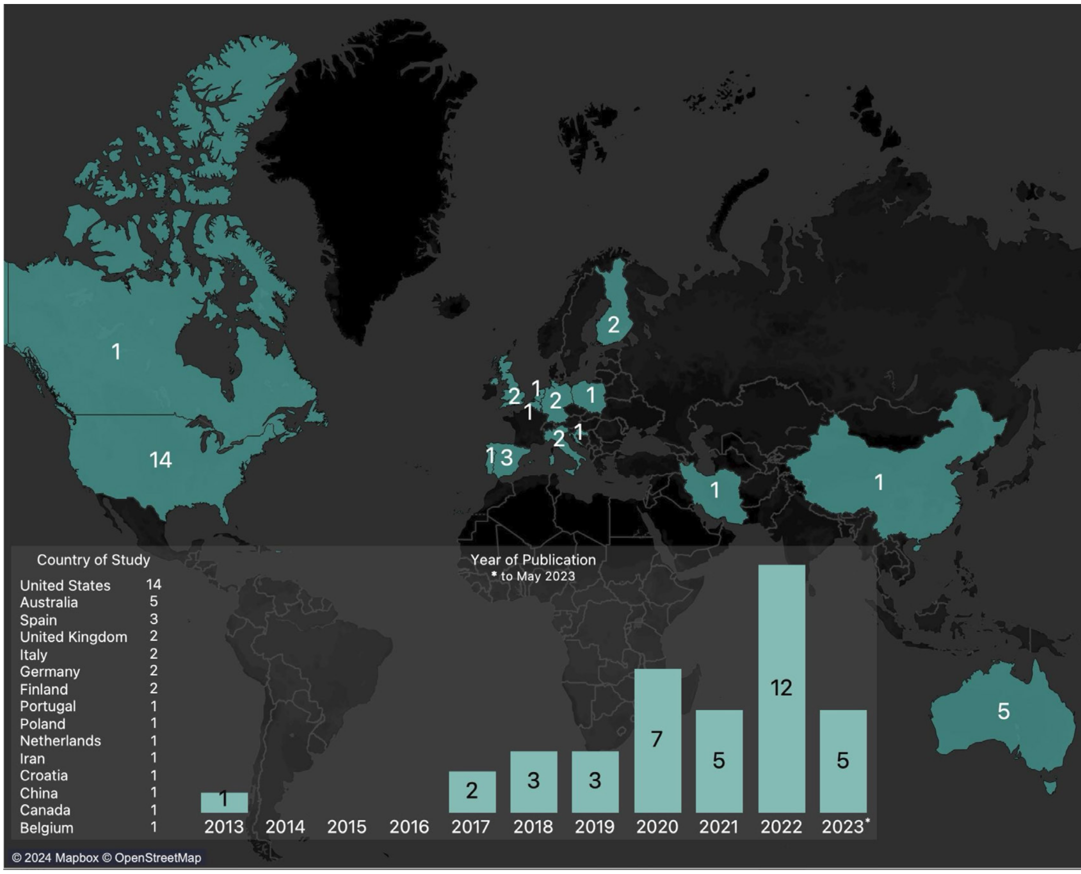

Figure 2 Included studies by year of publication and location. Number of studies added in columns.

Body part analysed

16 studies (42%) assessed the predictive capabilities of ML models relating to injuries sustained to any body part. 10 studies (26%) refined the target variable definition to any injury occurring to the lower extremity (ie, hamstring, calf, ankle, etc) as a whole, with another study (n=1) considering injuries to both the lower body and trunk. One study jointly predicted the onset of injuries to the knee joint and ankle joint of participants. The remaining 10 studies analysed injuries sustained to specific body parts; the most common were the knee (n=3) and hamstring muscles (n=3). Other injury locations under analysis included the mouth (n=1), shin (n=1), head (n=1, concussion) and kidneys (n=1).

Commonly applied ML models

Most studies (92%) related to classification tasks, whereby the ML models were developed to assign records to a particular class (ie, 'injury' or 'non-injury'). Two studies[21 22] used time-to-event analysis, with one other study[23] investigating regression analysis.

Table 1 Statistical summary of reviewed cohort sizes, volume of injuries analysed and dependent variables

	Mean (SD)	Median (IQR)	Range	Coefficient of variation
Cohort size	1417 (3,747)	122 (36–623)	11–15 682	264%
Injuries (n)	493 (1,471)	57 (28–194)	10–6982	299%
Dependent variables	61 (150)	31 (16–48)	7–957	246%

In 42% (n=16) of studies, only one ML model was assessed. Other studies applied two (n=3), three (n=8), four (n=4), five (n=4) and six (n=3) different methods to determine which achieved the highest statistical performance. Random Forest was the most investigated method of predictive ML algorithm, with variations of the method (including conditional, calibrated and survival) applied in 54% of studies (n=21), amounting to

Table 2 Data collection methods and injury definition criteria

Injury data collection (n(%))	
Medical staff	30 (79%)
Club/team medical staff	19
Other medical personnel	11
Self-reported survey	5 (13%)
Public data source	3 (8%)
Definition of injury (n(%))	
Time loss (number of days)	18 (47%)
≥1	13
≥2	1
≥5	1
≥7	1
≥8	1
≥9	1
Medical Evaluation	11 (29%)
Assessment by clinician	7
MRI diagnosis	4
Non-time-loss and time-loss	9 (24%)

Table 3 A summary of different machine learning and statistical approaches identified

Method	Applied	Highest performing method	Only method applied
Random Forest*[21 22 24–29]	23 (23%)	8 (21%)	6 (38%)
Extreme Gradient Boosting (XGBoost)[30–37]	8 (8%)	8 (21%)	4 (25%)
Logistic regression[9 44 47 48]	12 (12%)	4 (11%)	0 (0%)
Support Vector Machine[45 64 66 67]	9 (9%)	4 (11%)	1 (6%)
Decision tree+ [38–40]	15 (15%)	3 (8%)	1 (6%)
Artificial neural network[42 43]	6 (6%)	2 (5%)	1 (6%)
SmoteBoostM1 technique with a cost-sensitive AD tree[68 69]	2 (2%)	2 (5%)	0 (0%)
Naïve Bayes[46 70]	4 (4%)	1 (3%)	0 (0%)
LASSO regression[54]	2 (2%)	1 (3%)	1 (6%)
UnderBagging technique with a cost-sensitive SMO[56]	1 (1%)	1 (3%)	0 (0%)
Ridge regression[23]	1 (1%)	1 (3%)	0 (0%)
Gradient boosting algorithm[10]	1 (1%)	1 (3%)	1 (6%)
Dynamic Bayesian Network[49]	1 (1%)	1 (3%)	1 (6%)
Recursive partitioning and regression trees[41]	1 (1%)	1 (3%)	0 (0%)
K-nearest neighbours	6 (6%)	0 (0%)	0 (0%)
Elastic Net	2 (2%)	0 (0%)	0 (0%)
Generalised estimating equations	1 (1%)	0 (0%)	0 (0%)
Logit classifier	1 (1%)	0 (0%)	0 (0%)
Generalised linear mixed-effect models	1 (1%)	0 (0%)	0 (0%)
Linear discriminant	1 (1%)	0 (0%)	0 (0%)
Ordinary least squares regression	1 (1%)	0 (0%)	0 (0%)
Stepwise forward regression	1 (1%)	0 (0%)	0 (0%)
Total	**100**	**38**	**16**

For each group, the number of models using that method is presented as applied, identified as highest performing method or only method applied.
*Includes Conditional Random Forests, Calibrated Random Forests and Random Survival Forest,
†Includes Chi-square automatic interaction detector decision tree, C4.5, SimpleCart, J48 and random tree.

23 models. Nine studies developed and applied decision trees and variations thereof including Chi-square automatic interaction detector (CHAID), C4.5, SimpleCart, J48 and random tree, amounting to a total of 15 models. Logistic regression (n=12), Support Vector Machines (SVM, n=9) and Extreme Gradient Boosting (XGBoost, n=8) were also commonly applied.

Highest-performing ML models

XGBoost and Random Forest were both reported as the highest-performing ML methods in eight studies (table 3). A Random Forest solution achieved the highest AUC across the studies reviewed, with a reported value of 0.95.[24] XGBoost was reported as the statistically highest performing approach in each of the eight studies in which it was applied, although four of these instances considered it as the only predictive ML model. Four studies determined SVM as the highest-performing model, of which only one paper failed to compare it against other models. Decision trees were deemed the highest-performing approach in five studies.

23 studies (61%) determined that the statistically highest-performing approach was a tree-based solution. These solutions included Random Forest,[21 22 24–29] XGBoost[30–37] and decision trees,[38–40] referenced above, along with Recursive Partitioning and Regression Trees (n=1)[41] and Gradient Boosting Algorithms (n=1).[10] Artificial neural networks (ANN) provided the highest performance in two studies,[42 43] having been applied in six. Two ANN architectures included one hidden layer,[42 43] two used two hidden layers[44 45] and two failed to disclose the number of hidden layers.[33 46] Four studies determined that logistic regression outperformed the other ML models against which it was applied.[9 44 47 48] Bayesian approaches were reported to provide the highest predictive performance in two studies: Naive Bayes (n=1)[46] and Dynamic Bayesian Networks (n=1).[49] These results are summarised in table 3.

Class imbalance, cross-validation model explainability

15 studies (39%) identified an imbalance between classes within the data. This imbalance exists due to the infrequent nature of injuries, given that an athlete will experience more days uninjured than injured. All 15 studies applied oversampling techniques. 10 studies applied the synthetic minority oversampling technique (SMOTE), with others using random oversampling (n=4) and adaptive synthetic sampling (n=1). Four of these studies also applied undersampling techniques, with each randomly undersampling the majority class. Cross-validation techniques were performed in 32 studies (84%).

Seven studies (18%) applied model explainability methods. Five studies (13%) used SHapley Additive exPlanations (SHAP) values to improve model interpretability, with two studies using local-interpretable model-agnostic explanations.

Evaluation methods

10 evaluation metrics of predictive validity were reported across the 38 studies. 27 studies (71%) reported the area under the curve (AUC) of the receiver operating characteristic curve as a measure of model performance. An AUC of 0.5 represents a model with predictive performance no greater than random chance, while an AUC of 1 signifies a perfect predictor. The reported values of AUC, illustrated in figure 3, ranged between 0.57 and 0.95. The narrative assessment of ‘poor’, ‘acceptable’, ‘excellent’ and ‘outstanding’ conforms with the work of Sturdivant *et al*.[50] One-third of studies reported AUC values in the regions of ‘poor’ (between 0.50 and 0.69) (n=9), with 37% (n=10) achieving ‘acceptable’ (between 0.70 and 0.79) results. Five studies (19%) reported AUC scores between 0.80 and 0.89, considered ‘excellent’, and three (11%) attained AUC scores more than 0.9—classed as ‘outstanding’.

Of the remaining 10 studies that did not report AUC, sensitivity, also referred to as recall, was reported most often (n=6), with an average value of 0.81. Precision (n=4), specificity (n=2), F1 Score (n=2) and F2 Score (n=2) were the only other metrics stated more than once within these studies, achieving averages of 0.79, 0.85, 0.73 and 0.87, respectively. Across all 38 studies, measures of accuracy were reported in only six studies (16%) with an average score of 89.79%.

Dependent variables

A total of 1359 dependent variables were analysed across the 38 studies (see online supplemental table 1). The number of dependent variables available for predictive modelling in a single paper ranged from 7 to 957.[45 46] The average number of dependent variables was 61, with a coefficient of variation equal to 246%.

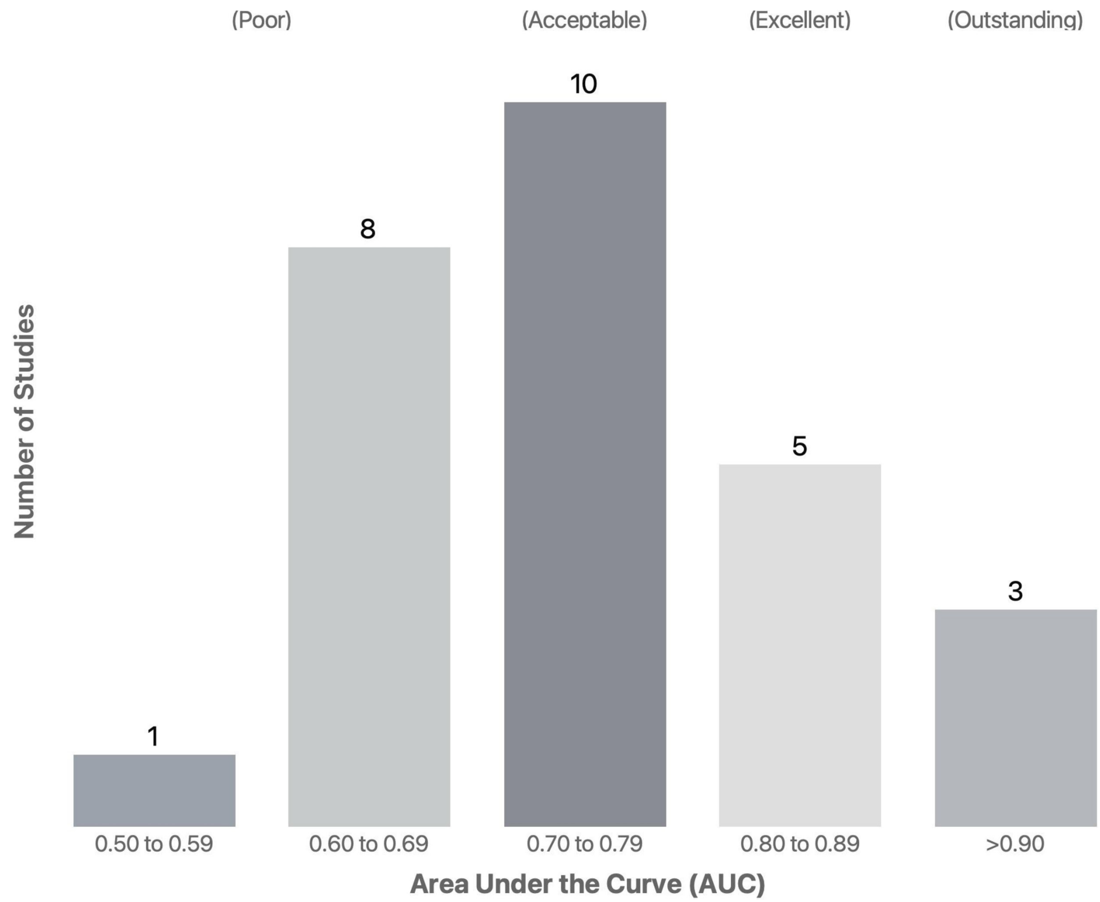

Figure 3 Area under the curve from receiver operator characteristic analysis.

Screening data, collected during the preseason or prior to study commencement, were the most frequently analysed. 28 studies (74%) collected data relating to non-modifiable risk factors, including age, injury history, playing position and years of competitive experience. 22 studies (58%) analysed musculoskeletal screening tests, and 18 (47%) analysed body composition and anthropometric measurements. Four studies screened for psychological risk factors, and one screened for self-perceived ankle instability. The most common measures of internal load used within the studies were a rating of perceived exertion (RPE) and session RPE (sRPE), which featured in 11 studies (29%), while 6 studies (16%) gathered scores related to individuals' self-perceived wellness (ie, sleep quality, stress, muscle soreness, etc). Seven studies (18%) analysed external load data gathered from wearable electronic performance and tracking systems (ie, GPS-based units). Less common data sources included urine samples (n=2), blood tests (n=1), menstrual cycle data (n=1) and sleep quality data from WHOOP straps (WHOOP, Boston, Massachusetts, USA) (n=1). A study hypothesis was declared in 29% of studies (n=11). One study (3%) explicitly defined their analyses as 'exploratory', with one further study reporting that their analyses were informed via an 'inductive' approach.

DISCUSSION

XGBoost and Random Forest were reported as the highest-performing methods, based on the published statistical evaluation metrics in 42% of the included studies. Random Forest was the most commonly investigated method of injury prediction. XGBoost was the only method found to provide the highest statistical performance in each of the studies in which it was applied. In 60% of studies, a tree-based solution was found to provide the greatest predictive performance of the methods applied. As AUC was the most commonly reported metric of model evaluation (71% of studies), it provides the most effective means of performance comparison between reviewed studies.

Predictive performance

Random Forest was the most commonly applied ML approach. In line with this, Random Forest was reported as the highest-performing ML model, alongside XGBoost. XGBoost was the only method found to provide the highest performance score in each of the studies in which it was featured. However, this performance must be caveated by the fact that four of the eight studies listed XGBoost as the only algorithm applied. In 60% of the studies, a tree-based solution achieved the highest statistical performance, with an average AUC score of 0.77, 12% higher than the average AUC score of 0.69 reported by other techniques. The three studies achieving the highest AUC, exceeding 0.9, each applied tree-based methods — CHAID decision tree, XGBoost and Calibrated Random Forest[24 34 40] — with another study achieving an F1 score of 94.4% using XGBoost.[36]

While these reported statistical metrics are impressive, with AUC above 0.9 indicative of near-perfect predictive models, the variation in approaches is notable when comparing the two highest-performing studies. Luu *et al*[34] sought to predict the risk of 'next-season injury' (to any body part, illnesses included). Based on publicly available datasets relating to National Hockey League match metrics (ie, minutes played, shots and offensive/defensive actions), the analysis included 2322 players, yielding an AUC of 0.948. Yet, the clinical utility of such a model, with a broad window of prediction that does not define the body part in which an injury might occur, is limited. Shaw *et al*[24] achieved the highest reported AUC score of 0.95, predicting the onset of medial tibial stress syndrome in a group of 99 military trainees. Due to the COVID-19 pandemic, the participants were not involved in 'structured physical training', instead conducting their own self-directed running, thus the load could not be accurately tracked. In this study, 35% of the participants developed medial tibial stress syndrome following a period of self-directed running. This is the same as the expected incidence of 35% reported in athletic populations.[51]

These examples emphasise the need for clinical relevance regarding ML research for sports injury risk prediction. Research groups should comprise an appropriate balance of technical and domain expertise. For instance, a methodologically sound ML model developed by a team of data scientists will be of little benefit if the question it is designed to answer is misaligned with the needs of the clinical practitioners. To more effectively disseminate ML research to sports medicine teams, a wider application of model interpretability techniques, used in only 18% of reviewed studies, should be adopted. These methods, namely SHAP values, allow previously deemed 'black-box' models (ie, complex models that are not easily understood by humans)[52] to be more easily understood, facilitating their real-world use by medical teams, coaches, and athletes.

Available datasets

Compared with traditional statistical methods, ML methods are capable of analysing larger datasets with more dependent variables to develop more complex models.[53] Conversely, this review finds no clear correlation between the number of dependent variables being analysed and ML algorithmic performance. For example, an SVM model with 957 dependent variables available for the prediction of concussion achieved an AUC of 0.73,[45] yet was statistically outperformed by an SVM trained with only 15 features, reporting AUC=0.840.[52] These results indicate that the volume of dependent variables alone will not guarantee improved predictive performance. The context-specific nature of ML model deployment, contingent upon both the question at hand and the datasets being investigated, is key to interpreting the clinical utility of these findings.

Furthermore, ML approaches do not always yield greater predictive capabilities than traditional methods, with four studies reporting that logistic regression outperformed the ML methods against which it was applied. However, these findings require additional context. For instance, Jauhiainen *et al*[48] reported that logistic regression outperformed Random Forest, with achieved AUC values of 0.65 and 0.63, respectively. The negligible difference in AUC of 0.02, with both techniques yielding 'poor' predictive performance, suggests that the dataset at the author's disposal was not conducive to the task of injury risk prediction. In another example, logistic regression achieved an AUC of 0.82, marginally outperforming Random Forest (0.8), as well as SVM (0.73) and K-Nearest Neighbours (0.7).[44] The authors acknowledge that 'the data set is too small to fit a more complex model', given that GPS data was collected for only a subset of 'important' athletes within the group, due to the high cost of electronic performance and tracking equipment.

As such, 53% of studies (n=20), used only data taken from screening tests to train the ML models. While less costly than more continuous methods of athletic monitoring,[54] prediction of sporting injury through the analysis of screening data alone has been deemed ineffective.[55] The average AUC score of models trained and tested using screening data alone was 0.73. Performance was increased when more granular methods of athletic monitoring (RPE/sRPE, self-perceived wellness questionnaires, GPS metrics, etc) were analysed, achieving an average AUC of 0.77. The analysis of external load using GPS tracking is insubstantial, with only seven studies using the increased granularity of data offered by such electronic performance and tracking systems (average AUC of 0.75).

These findings highlight the logistical and financial challenges that teams and medical staff face in relation to the large-scale data collection required to effectively train ML models.[44 56]

The inherently sporadic nature of sporting injury introduces numerous difficulties when developing and evaluating predictive models. First, to address the class imbalance issue, 39% of studies applied some form of sampling technique. These methods of either undersampling the majority class (ie, 'non-injury') or oversampling the minority class (ie, 'injury') may cause unforeseen data issues. For instance, the SMOTE technique, the most commonly applied oversampling method in the studies included in this review, creates synthetic 'injury' instances upon which models can be trained. This generation of synthetic data introduces the risk of overfitting, whereby the model learns from synthetic cases that are too similar to the original data, thus reducing the generalisation of the models. The use of undersampling techniques causes unwanted loss of data, whereby instances of the majority class are removed to even the class distribution.

The issue of class imbalance adds complexity when assessing model performance, as evidenced by the numerous approaches (n=10) to statistical evaluation presented within the studies included in this review (online supplemental table 2). The frequent use of AUC, reported in 71% of cases, given its ability to better reflect class imbalance than accuracy scores.[57] Conversely, recent research has suggested that AUC may not be as effective at handling skewed data as previously thought, instead proposing the use of the area under the precision-recall curve,[58] further demonstrating the difficulty in effectively benchmarking and comparing ML performance. Importantly, there is no consensus on the best single metric of model evaluation from a sports injury perspective.

In spite of the widely cited recommendations of Bittencourt *et al*[16] regarding the suitability of complex ANNs to the task of injury risk analysis in sports, such approaches are limited within the current research, applied in only 16% of studies. Although extensively utilised within other medical research domains,[59] ANN topologies are yet to be thoroughly investigated from a sports injury perspective. The limited application of ANN methods may be reflective of the small datasets being analysed in many cases. A small dataset, for example, compromised the development of a long short-term memory neural network by Lyubovsky *et al*.[44]

Methodological robustness

The definition of, 'injury', varies significantly across the included studies. Research has most commonly attempted to predict injuries sustained to any body part (42%), discounting the significant differences in injury mechanisms relating to each. Moreover, 47% of studies jointly considered both contact and non-contact injuries during their analysis. By attempting to predict such disparate injury types and locations, models are incapable of appreciating the different inciting events and risk factors surrounding each. Additionally, varying injury prediction windows have been analysed, ranging from day-to-day,[44] weeklong,[31] and season-long[34 35] granularity. Each of these definitions has a significant bearing on the nature of the research question and thus the suitability of particular ML models to the task.

Numerous methodological limitations are evident within the cohort of reviewed studies. In 42% (n=16) of studies, only one type of predictive modelling was applied. This methodological approach significantly hinders the evaluation of ML techniques by failing to offer a comparative analysis as to the most effective method for the dataset being analysed. Furthermore, by declaring only one ML method, these studies are susceptible to 'p-hacking' whereby researchers conduct numerous tests with multiple combinations of parameters until a statistically significant result is found, reporting it in isolation.[60]

HARKing, a questionable research practice whereby authors develop or modify a study's hypothesis after the data has been analysed,[61] has been identified in the fields of sports and exercise medicine[62] and ML.[63] Conversely, it has been argued that ML research is inherently inductive in nature, thus not dependent on the acceptance or rejection of a stated hypothesis, as is the case with deductive reasoning.[49] This is supported by the fact that only 29% of reviewed studies declared a hypothesis. On the other hand, and in spite of the exploratory nature of ML research, only one paper explicitly declared its analysis as such. We suspect widespread HARKing within the reviewed cohort, with one study stating, in the introduction, that 'an acute kidney injury model of 90% accuracy and sensitivity was hypothesised to be possibly constructed.'[64] The authors later report results of precisely 90% for both accuracy and sensitivity.

This review has identified several ML techniques and highlighted the large variation in predictive model performance in the included studies. This variability is indicative of inconsistencies in many aspects of the study design, including data collection methods, study populations, definition and scope of injury, sources of data and dependent variables analysed. This is an important consideration when assessing ML model performance beyond the stated evaluation metrics. Importantly, contextual factors under which the models have been developed and the questions they seek to answer must have real-world utility if they are to benefit the sporting community. There exists no 'one size fits all' approach to injury risk prediction. This is clearly demonstrated when considering the charted research relating to football, the most analysed sport. The nine football studies included reported seven different ML approaches as providing the best predictive performance, emphasising the individuality of each dataset and the need to select the appropriate ML method for the type of data being analysed. Therefore, sporting organisations should collaborate to produce and share larger and more granular datasets to develop and validate ML models based on complex interactions within the data.[65] Data sharing could bolster the transition from group analyses to more individualised risk models. By appreciating the personal nature of injury risk profiling, such as an individual's predisposition to certain injuries, ML models could provide improved insights and augment decision-making on an individual basis. Our results suggest future research should be based on higher-quality and more extensive injury data, including more diverse sources of data. Diverse sources of data will reduce ML model bias. Data from diverse populations (e.g., the Global South) and related to important contextual factors, such as playing surfaces, weather conditions and use of sporting equipment, coupled with regular athletic monitoring through electronic performance and tracking systems, could improve the overall efficacy of future models.

LIMITATIONS

Our study inclusion criteria did not include grey literature databases. As a result, dissertations, conference proceedings and unpublished studies were not included, and important data could have been missed. In addition, by not including non-English studies, other relevant research could have been missed too. As our database search was conducted in May 2023, studies in this fast-moving field published after this point were not considered. As with any review of this nature, potential publication bias is an issue, as noteworthy ML applications may be published more readily than those with underwhelming results.

CONCLUSION

The effective application of ML models for sports injury risk prediction is hampered by widespread variability in study design, methodological approach and statistical analysis. Instances in which ML models have failed to outperform traditional approaches to sports injury risk prediction demonstrate the need for more granular and integrated datasets. These datasets may be capable of supplying the volume, variety and quality of data required for the effective application of ML approaches.

X Christopher Leckey @chrisleckey_, Nicol van Dyk @NicolvanDyk, Cailbhe Doherty @dohertycailbhe and Eamonn Delahunt @EamonnDelahunt

Contributors CL and ED are the co-guarantors of the review. CL, ED and NvD identified the framework for the review and developed the scoping review protocol. All authors contributed to the development of the research questions and study design. All authors edited and approved the final version of the manuscript.

Funding CL is in receipt of funding under the Irish Research Council's Employment-Based Postgraduate Scholarship programme (funding ID: EBPPG/2022/228). The Irish Research Council was not involved in the design or conduct of the study or the analysis and interpretation of the results.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Christopher Leckey http://orcid.org/0000-0001-5614-3252
Nicol van Dyk http://orcid.org/0000-0002-0724-5997
Cailbhe Doherty http://orcid.org/0000-0002-5284-856X
Aonghus Lawlor http://orcid.org/0000-0002-6160-4639
Eamonn Delahunt http://orcid.org/0000-0001-5449-5932

REFERENCES

1 Zech A, Hollander K, Junge A, *et al*. Sex differences in injury rates in team-sport athletes: A systematic review and meta-regression analysis. *J Sport Health Sci* 2022;11:104–14.

2 Hägglund M, Waldén M, Magnusson H, *et al*. 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;47:738–42.
3 Williams S, Trewartha G, Kemp SPT, *et al*. Time loss injuries compromise team success in Elite Rugby Union: a 7-year prospective study. *Br J Sports Med* 2016;50:651–6.
4 Bradbury JC. DETERMINANTS OF REVENUE IN SPORTS LEAGUES: AN EMPIRICAL ASSESSMENT. *Econ Inq* 2019;57:121–40.
5 Eliakim E, Morgulev E, Lidor R, *et al*. Estimation of injury costs: financial damage of English Premier League teams' underachievement due to injuries. *BMJ Open Sport Exerc Med* 2020;6:e000675.
6 Wilkerson GB, Gupta A, Colston MA. Mitigating Sports Injury Risks Using Internet of Things and Analytics Approaches. *Risk Anal* 2018;38:1348–60.
7 Campbell PG, Stewart IB, Sirotic AC, *et al*. Does exercise intensity affect wellness scores in a dose-like fashion? *Eur J Sport Sci* 2020;20:1395–404.
8 Rogalski B, Dawson B, Heasman J, *et al*. Training and game loads and injury risk in elite Australian footballers. *J Sci Med Sport* 2013;16:499–503.
9 Carey DL, Ong K, Whiteley R, *et al*. Predictive Modelling of Training Loads and Injury in Australian Football. *Int J Comput Sci Sport* 2018;17:49–66.
10 Hecksteden A, Schmartz GP, Egyptien Y, *et al*. Forecasting football injuries by combining screening, monitoring and machine learning. *Sci Med Football* 2023;7:214–28.
11 Costa J, Figueiredo P, Nakamura F, *et al*. Intra-individual variability of sleep and nocturnal cardiac autonomic activity in elite female soccer players during an international tournament. *PLoS One* 2019;14:e0218635.
12 Nussbaum M. Machine learning and processing of large data. In: Goss MJ, Oliver M, eds. *Encyclopedia of soils in the environment*. Oxford: Academic Press, 2023: 509–20.
13 Bai L, Gedik R, Egilmez G. What does it take to win or lose a soccer game? A machine learning approach to understand the impact of game and team statistics. *J Oper Res Soc* 2023;74:1690–711.
14 Ćwiklinski B, Giełczyk A, Choraś M. Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers. *Entropy (Basel)* 2021;23:90.
15 Watson N, Hendricks S, Stewart T, *et al*. Integrating machine learning and decision support in tactical decision-making in rugby union. *J Oper Res Soc* 2021;72:2274–85.
16 Bittencourt NFN, Meeuwisse WH, Mendonça LD, *et al*. Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition-narrative review and new concept. *Br J Sports Med* 2016;50:1309–14.
17 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* 2018;169:467–73.
18 Fuller CW, Ekstrand J, Junge A, *et al*. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Br J Sports Med* 2006;40:193–201.
19 Hägglund M, Waldén M, Bahr R, *et al*. Methods for epidemiological study of injuries to professional football players: developing the UEFA model. *Br J Sports Med* 2005;39:340–6.
20 Waldén M, Krosshaug T, Bjørneboe J, *et al*. Three distinct mechanisms predominate in non-contact anterior cruciate ligament injuries in male professional football players: a systematic video analysis of 39 cases. *Br J Sports Med* 2015;49:1452–60.
21 Lu Y, Jurgensmeier K, Till SE, *et al*. Early ACLR and Risk and Timing of Secondary Meniscal Injury Compared With Delayed ACLR or Nonoperative Treatment: A Time-to-Event Analysis Using Machine Learning. *Am J Sports Med* 2022;50:3544–56.
22 Lu Y, Reinholz AK, Till SE, *et al*. Predicting the Risk of Posttraumatic Osteoarthritis After Primary Anterior Cruciate Ligament Reconstruction: A Machine Learning Time-to-Event Analysis. *Am J Sports Med* 2023;51:1673–85.
23 Martins F, Przednowek K, França C, *et al*. Predictive Modeling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players. *J Clin Med* 2022;11:4923.
24 Shaw A, Newman P, Witchalls J, *et al*. Externally validated machine learning algorithm accurately predicts medial tibial stress syndrome in military trainees; a multi-cohort study. *Sports Medicine* [Preprint].
25 Goggins L, Warren A, Osguthorpe D, *et al*. Detecting Injury Risk Factors with Algorithmic Models in Elite Women's Pathway Cricket. *Int J Sports Med* 2022;43:344–9.
26 Thornton HR, Delaney JA, Duthie GM, *et al*. Importance of Various Training-Load Measures in Injury Incidence of Professional Rugby League Athletes. *Int J Sports Physiol Perform* 2017;12:819–24.
27 Henriquez M, Sumner J, Faherty M, *et al*. Machine Learning to Predict Lower Extremity Musculoskeletal Injury Risk in Student Athletes. *Front Sports Act Living* 2020;2:576655.
28 Briand J, Deguire S, Gaudet S, *et al*. Monitoring Variables Influence on Random Forest Models to Forecast Injuries in Short-Track Speed Skating. *Front Sports Act Living* 2022;4:896828.
29 Farhadian M, Torkaman S, Mojarad F. Random forest algorithm to identify factors associated with sports-related dental injuries in 6 to 13-year-old athlete children in Hamadan, Iran-2018 -a cross-sectional study. *BMC Sports Sci Med Rehabil* 2020;12:69.
30 Rommers N, Rössler R, Verhagen E, *et al*. A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Med Sci Sports Exerc* 2020;52:1745–51.
31 Huang Y, Huang S, Wang Y, *et al*. A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. *Front Physiol* 2022;13:937546.
32 Lövdal SS, Den Hartigh RJR, Azzopardi G. Injury Prediction in Competitive Runners With Machine Learning. *Int J Sports Physiol Perform* 2021;16:1522–31.
33 Lu Y, Pareek A, Lavoie-Gagne OZ, *et al*. Machine Learning for Predicting Lower Extremity Muscle Strain in National Basketball Association Athletes. *Orthop J Sports Med* 2022;10.
34 Luu BC, Wright AL, Haeberle HS, *et al*. Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017. *Orthop J Sports Med* 2020;8.
35 Karnuta JM, Luu BC, Haeberle HS, *et al*. Machine Learning Outperforms Regression Analysis to Predict Next-Season Major League Baseball Player Injuries: Epidemiology and Validation of 13,982 Player-Years From Performance and Injury Profile Trends, 2000-2017. *Orthop J Sports Med* 2020;8.
36 Piłka T, Grzelak B, Sadurska A, *et al*. Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors. *Sensors (Basel)* 2023;23:1227:23:.
37 Senbel S, Sharma S, Raval MS, *et al*. Impact of Sleep and Training on Game Performance and Injury in Division-1 Women's Basketball Amidst the Pandemic. *IEEE Access* 2022;10:15516–27.
38 Rossi A, Pappalardo L, Cintia P, *et al*. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One* 2018;13:e0201264.
39 Oliver JL, Ayala F, De Ste Croix MBA, *et al*. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *J Sci Med Sport* 2020;23:1044–8.
40 Connaboy C, Eagle SR, Johnson CD, *et al*. Using Machine Learning to Predict Lower-Extremity Injury in US Special Forces. *Med Sci Sports Exerc* 2019;51:1073–9.
41 Bird MB, Koltun KJ, Mi Q, *et al*. Predictive utility of commercial grade technologies for assessing musculoskeletal injury risk in US Marine Corps Officer candidates. *Front Physiol* 2023;14:1088813.
42 Morse S, Talty K, Kuiper P, *et al*. Machine learning prediction of combat basic training injury from 3D body shape images. *PLoS One* 2020;15:e0235017.
43 McCullagh J, Whitfort T. *An Investigation into the Application of Artificial Neural Networks to the Prediction of Injuries in Sport*. . World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, 2013:7. 356–60.
44 Lyubovsky A, Liu Z, Watson A, *et al*. A pain free nociceptor: Predicting football injuries with machine learning. *Smart Health (2014)* 2022;24:100262.
45 Castellanos J, Phoo CP, Eckner JT, *et al*. Predicting Risk of Sport-Related Concussion in Collegiate Athletes and Military Cadets: A Machine Learning Approach Using Baseline Data from the CARE Consortium Study. *Sports Med* 2021;51:567–79.
46 Ruddy JD, Shield AJ, Maniar N, *et al*. Predictive Modeling of Hamstring Strain Injuries in Elite Australian Footballers. *Med Sci Sports Exerc* 2018;50:906–14.
47 Karuc J, Mišigoj-Duraković M, Šarlija M, *et al*. Can Injuries Be Predicted by Functional Movement Screen in Adolescents. *Appl Mach Learn J Strength Cond Res* 2021;35:910–9.
48 Jauhiainen S, Kauppi J-P, Leppänen M, *et al*. New Machine Learning Approach for Detection of Injury Risk Factors in Young Team Sport Athletes. *Int J Sports Med* 2021;42:175–82.
49 Peterson KD, Evans LC. Decision Support System for Mitigating Athletic Injuries. *Int J Comput Sci Sport* 2019;18:45–63.
50 Sturdivant R, Hosmer Jr DW, Sturdivant R. Applied Logistic Regression. John Wiley & Sons, 2013.
51 Deshmukh NS, Phansopkar P. Medial Tibial Stress Syndrome: A Review Article. *Cureus* 2022;14:e26641.
52 Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Can J Cardiol* 2022;38:204–13.
53 Ley C, Martin RK, Pareek A, *et al*. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc* 2022;30:753–7.
54 Kolodziej M, Groll A, Nolte K, *et al*. Predictive modeling of lower extremity injury risk in male elite youth soccer players using least absolute shrinkage and selection operator regression. *Scandinavian Med Sci Sports* 2023;33:1021–33.
55 Bahr R. Why screening tests to predict injury do not work—and probably never will…: a critical review. *Br J Sports Med* 2016;50:776–80.
56 Ruiz-Pérez I, López-Valenciano A, Hernández-Sánchez S, *et al*. A Field-Based Approach to Determine Soft Tissue Injury Risk in Elite Futsal Using Novel Machine Learning Techniques. *Front Psychol* 2021;12:610210.
57 Richardson E, Trevizani R, Greenbaum JA, *et al*. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patt (N Y)* 2024;5:100994.
58 Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* 2015;10:e0118432.
59 Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2:230–43.
60 Head ML, Holman L, Lanfear R, *et al*. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13:e1002106.

61 Kerr NL. HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* 1998;2:196–217.
62 Büttner F, Toomey E, McClean S, *et al*. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. *Br J Sports Med* 2020;54:1365–71.
63 Gencoglu O, Gils M, Guldogan E, *et al*. HARK Side of Deep Learning - From Grad Student Descent to Automated Machine Learning. 2019.
64 Hsu P-Y, Hsu Y-C, Liu H-L, *et al*. An Acute Kidney Injury Prediction Model for 24-hour Ultramarathon Runners. *J Hum Kinet* 2022;84:103–11.
65 Tedesco S, Scheurer S, Brown KN, *et al*. A survey on the use of artificial intelligence for injury prediction in sports. 2022 IEEE International Workshop on Sport, Technology and Research (STAR); 2022
66 Jauhiainen S, Kauppi J-P, Krosshaug T, *et al*. Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes. *Am J Sports Med* 2022;50:2917–24.
67 Mandorino M, Figueiredo AJ, Cima G, *et al*. Predictive Analytic Techniques to Identify Hidden Relationships between Training Load, Fatigue and Muscle Strains in Young Soccer Players. *Sports (Basel)* 2021;10:3.
68 Ayala F, López-Valenciano A, Gámez Martín JA, *et al*. A Preventive Model for Hamstring Injuries in Professional Soccer: Learning Algorithms. *Int J Sports Med* 2019;40:344–53.
69 López-Valenciano A, Ayala F, Puerta JM, *et al*. A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms. *Med Sci Sports Exerc* 2018;50:915–27.
70 Ruddy J, Shield A, Maniar N, *et al*. Predicting hamstring strain injury incidence in elite Australian footballers. *J Sci Med Sport* 2017;20:10–1.