frontiers in Public Health

TYPE Systematic Review
PUBLISHED 27 November 2026
DOI 10.3389/fpubh.2026.1938641

Check for updates

OPEN ACCESS

EDITED BY
Yang Renhao,
Ruijin Hospital North, China

REVIEWED BY
Fabio Santacaterina,
Campus Bio-Medico University, Italy
Taifu Hou,
Luohe Medical College, China

*CORRESPONDENCE
Ulas Can Yildirim
ucyildirim@sinop.edu.tr;
ulascanyildirim@aybu.edu.tr
Selin Yildirim Tuncer
selin.yildirim@lokmanhekim.edu.tr

RECEIVED 16 July 2026
REVISED 20 October 2026
ACCEPTED 09 November 2026
PUBLISHED 27 November 2026

CITATION
Ozdenk S, Erkan D, Ermis A, Cevik A, Bayrakdar A, Yildirim S, Turgut Z, Gundem MC, Yildirim Tuncer S and Yildirim UC (2026) Machine learning–based injury prediction models in athletes: a systematic review of predictive performance and injury-related outcomes. *Front. Public Health* 14:1938641. doi: 10.3389/fpubh.2026.1938641

COPYRIGHT
© 2026 Ozdenk, Erkan, Ermis, Cevik, Bayrakdar, Yildirim, Turgut, Gundem, Yildirim Tuncer and Yildirim. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Machine learning–based injury prediction models in athletes: a systematic review of predictive performance and injury-related outcomes

Serhat Ozdenk[1], Dilara Erkan[2], Aydan Ermis[3], Adem Cevik[4], Akan Bayrakdar[5], Sila Yildirim[6], Zeynep Turgut[1], Mehmet Can Gundem[1], Selin Yildirim Tuncer[7]* and Ulas Can Yildirim[6,8]*

[1]Department of Sport Management, Faculty of Sport Sciences, Sinop University, Sinop, Türkiye, [2]Department of Physical Education and Sports, Institute of Health Sciences, Ankara University, Ankara, Türkiye, [3]Department of Foreign Languages, School of Foreign Languages, Ondokuz Mayıs University, Samsun, Türkiye, [4]Department of Physical Education and Sports, Faculty of Sport Sciences, Ondokuz Mayıs University, Samsun, Türkiye, [5]Department of Coaching Education, Faculty of Sport Sciences, Alanya Alaaddin Keykubat University, Antalya, Türkiye, [6]Department of Coaching Education, Faculty of Sport Sciences, Sinop University, Sinop, Türkiye, [7]Department of Coaching Education, Faculty of Sport Sciences, Lokman Hekim University, Ankara, Türkiye, [8]Department of Coaching Education, Faculty of Sport Sciences, Ankara Yildirim Beyazit University, Ankara, Türkiye

Background: Machine learning (ML), a subfield of artificial intelligence (AI), is increasingly used to predict injury risk in athletes. However, the predictive performance, methodological reliability, and generalizability of these models across sporting contexts are not yet clear. In this systematic review, ML-based injury prediction models developed for athlete populations were examined in terms of performance metrics, validation approaches, predictor domains, and risk of bias.
Methods: PubMed, Web of Science, and Scopus databases were searched from inception to January 15, 2026. Peer-reviewed studies that developed or evaluated machine-learning-based models predicting injury risk in competitive athletes were included in the review. Study selection and data extraction were performed independently by two researchers. Risk of bias was assessed using PROBAST. Due to heterogeneity among studies, findings were synthesized narratively in accordance with the Synthesis Without Meta-analysis (SWiM) guideline.
Results: Of the 791 records identified, 17 met the inclusion criteria. Among the 15 studies reporting AUC for observed injury outcomes, the descriptive median of the representative study-level AUC values was 0.80 (IQR, 0.75–0.83), with reported values ranging from 0.63 to 0.99. Sensitivity ranged from approximately 30–100%, and specificity from 39 to 99.5%. External training load, neuromuscular performance, and injury history were the most frequently used predictor domains. Risk of bias was high in eight studies. Validation relied predominantly on internal approaches. None of the included studies performed external validation using an independent dataset, and one study reported an author-described temporal evaluation based on an athlete-disjoint hold-out set. The analysis domain was the principal source of concern, particularly because of limited outcome-event counts, inadequate handling of data complexity, and insufficient control of overfitting and optimism.

Conclusion: ML-based injury prediction models generally reported moderate discrimination across individual studies, although the estimates were heterogeneous and were not pooled. However, the current evidence base is limited by high or unclear risk of bias, limited external validation, class imbalance, and substantial methodological heterogeneity. These models should therefore be regarded as investigational rather than ready for routine clinical or sports-performance decision support. Future studies should prioritize prospective external validation, representative multicenter samples, and reporting of calibration and class-specific performance metrics.

Systematic review registration: https://www.crd.york.ac.uk/PROSPERO/view/CRD420261377436, Identifier PROSPERO (CRD420261377436).

KEYWORDS

artificial intelligence, athletes, injury prediction, machine learning, sports injuries

1 Introduction

Musculoskeletal injuries, which directly affect athletes and sports stakeholders, are related to many parameters such as training continuity, team performance, and consistency in sport. In professional team sports, the injury rate can reach serious levels. In football, in-match injuries occur at a rate of approximately 20 injuries per 1,000 exposure hours. Similarly, a high risk of musculoskeletal injuries is observed in basketball (1, 2). Particularly in teams with inadequately structured organizational risk management, injury-related absences and deterioration in tactical flexibility can lead to a decline in team performance. This loss of performance brings with it serious economic problems. While injuries increase healthcare costs due to extensive rehabilitation and surgical interventions, factors such as early loss of productivity, decreased spectator and sponsorship revenues increase the financial losses of sport clubs (3). When musculoskeletal injuries are examined from an athlete-centered perspective, it is seen that they cause early-onset osteoarthritis, chronic pain, and a decrease in quality of life. In addition, these injuries also shape the post-career lives of athletes (4). Particularly anterior cruciate ligament (ACL) and hamstring tears are high-prevalence injuries that, due to their high re-injury rates, put athletes at risk of early retirement (5, 6). Developing proactive strategies against injuries that negatively impact athletes and sport clubs in both the short and long term is of critical importance.

Traditional injury-risk screening approaches have shown limited ability to predict which athletes will subsequently be injured (7). Risk estimation methods generally aim to assess the athlete's workload using session-based perceived exertion ratings (RPE) and GPS-based measurements. This assessment involves converting collected data into indicators such as the acute and chronic workload ratio (ACWR) (8, 9). Neuromuscular capacity, isokinetic dynamometry, jumping tests, and functional movement scan results are used as performance data in risk assessment. Furthermore, the results obtained are evaluated together with epidemiological factors such as the athlete's injury history and bilateral strength asymmetries (10, 11). Identifying risk factors, making individual-level predictions, and evaluating decision utility are related but distinct objectives. A variable may be associated with injury at the group level; however, this does not mean that it will significantly improve individual predictions for athletes. Similarly, adequate predictive performance alone does not demonstrate that the use of a model improves clinical or training decisions or reduces injury outcomes (7, 12).

In statistical analyses, logistic regression is one of the methods frequently used to mathematically reveal the relationships between variables that make up risk profiles (13, 14). While an AUC value of approximately 0.70 can be used in some studies to define acceptable discrimination performance, it should not be considered a universal threshold for clinical usefulness, especially in the field of sports medicine where calibration, class balance, validation, and clinical decision outcomes are also important (15, 16). It is increasingly accepted that sports injuries result from complex and nonlinear interactions between multiple risk factors (17). When properly structured, both regression-based approaches and machine learning approaches can model nonlinear relationships and interactions. While logistic regression requires these relationships to be explicitly defined, machine learning methods can provide greater flexibility in processing high-dimensional or multimodal data. However, neither approach is inherently more accurate, generalizable, or clinically useful; performance depends on data quality, sample size, model specification, validation design, and the intended context of use (18, 19). Therefore, focusing solely on prediction performance is insufficient when evaluating machine learning-based injury prediction models. The validation strategies, risk of bias, and practical applicability of these models must also be systematically addressed. The fact that a large proportion of ML-based injury prediction models are developed using datasets from a single sport, a specific level of competition, or a limited demographic group also limits their transferability to different athlete populations and their external validity (20). However, tree-based ensemble algorithms, particularly random forest and XGBoost, offer advantages in modeling complex and nonlinear relationships between variables and can support variable selection, but they do not inherently resolve class imbalance arising from the relatively low incidence of injuries (21). Similarly, support vector machines and artificial neural networks allow the inclusion of highly dimensional and heterogeneous data sources such as measurements, kinematic indicators, and physiological biomarkers from wearable technologies into prediction models (22). In previous studies, model performance has been evaluated using various discrimination and calibration measures such as sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and Brier score. AUC-ROC values ranging from approximately 0.70 to 0.85 have been reported for models predicting soft tissue injuries, indicating moderate to good discrimination performance (23, 24).

The applicability of these algorithms varies depending on the structure, dimensionality, and size of the available dataset. Tree-based

ensemble methods such as Random Forest and XGBoost can effectively model nonlinear relationships and complex interactions, while support vector machines and artificial neural networks can handle high-dimensional data. However, the performance of these approaches depends heavily on the characteristics of the dataset, the model's configuration, and hyperparameter settings. Furthermore, more complex models can present problems in terms of interpretability and overfitting, especially in small datasets (18, 21, 22). Despite advances in machine learning, current studies in the field of injury prediction appear to have significant methodological limitations (18). Models developed with small datasets, in particular, may perform well in the sample in which the study is conducted, but may not yield similar results across different teams, seasons, or groups of athletes. This demonstrates that high internal validation approach values do not always indicate strong generalizability, and the risk of overfitting should be considered (19). Appropriate internal, temporal, and external evaluation is therefore important for estimating optimism and determining whether model performance is maintained in new settings (12). Another significant problem is that injury cases are generally less frequent than injury-free cases. This class imbalance can make it difficult to accurately identify athletes who are likely to be injured and negatively impact positive class prediction performance (21). These limitations make it difficult to transfer models developed for a team or season to different groups of athletes and real-world application environments. Incomplete reporting of predictors, model specification, and evaluation procedures can also impede the critical appraisal and implementation of prediction models (25). In the current literature, there are very few studies that have prospectively tested the developed models by incorporating them into injury prevention programs. Therefore, it is not yet clear whether the increase in predictive performance translates into a significant reduction in the number of injuries, time loss from training or competition, or injury severity.

The current evidence regarding the reliability of ML-based injury prediction models and their potential role as a proactive tool in injury management is insufficient. This review aims to systematically examine the literature on ML-based injury prediction models developed for athlete populations. Accordingly, by evaluating the methodological quality of the models, their prediction performance, and their relationship with clinically significant injury outcomes, the goal is to provide an evidence-based framework that can guide future research and field applications.

2 Methods

2.1 Protocol and registration

This systematic review was conducted and reported in accordance with the PRISMA 2020 statement (26) (Supplementary Material 2). The review was registered with PROSPERO on April 23, 2026 (CRD420261377436). At the time of registration, pilot work, database searching, and eligibility screening had been completed, whereas data extraction, risk-of-bias assessment, and evidence synthesis had started but had not yet been completed. The initial PROSPERO record specified adult athletes as participants aged ≥18 years and included organized competitive-sport settings, including collegiate athletes. Thus, although the ≥18-year criterion was documented in the initial registered record and was applied during screening, registration occurred after searching and eligibility screening had been completed.

The PROSPERO record was amended on August 19, 2026 to align it with the methods implemented in the review. The amendments included the use of Web of Science instead of the initially listed Embase, restriction to published peer-reviewed English-language studies, removal of conference-proceedings searching, and omission of the initially planned certainty-of-evidence assessment. These amendments are reported in the updated PROSPERO record and the PRISMA checklist.

2.2 Eligibility criteria

Eligibility criteria for the studies were determined according to the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) framework. In terms of population, eligible populations comprised adult athletes (≥18 years) participating in organized team or individual sports involving structured training and formal competition. Eligible competition levels included university/NCAA, competitive amateur, sub-elite, elite, semi-professional, and professional levels. Studies involving recreationally active individuals, the general active population, military personnel, or students not involved in a clearly defined organized competitive sport were excluded. Samples consisting solely of young athletes (<18 years) were excluded. Mixed-sample studies including both adults and young people were considered acceptable only if data for adult athletes were presented separately or if it was reported that there were no participants under the age of 18.

In terms of intervention, machine learning-based prediction models, including supervised learning algorithms, ensemble methods, and deep learning architectures developed to predict or classify injury risk, were examined. In terms of comparison, studies comparing these models with traditional statistical methods, particularly approaches such as logistic regression, as well as studies that did not use any reference model, were included. As outcome variables, both clinical injury indicators such as injury frequency, severity, training or match loss, and relapse rate, as well as measures reflecting model performance such as AUC, sensitivity, specificity, precision, accuracy, F1 and F2 scores, AUPRC, and Brier score were considered. In terms of study design, only original peer-reviewed studies published in English with full text access were included; reviews, editorials, letters, conference proceedings, protocols, theses, and book chapters were excluded. Studies classifying biomechanically or clinically defined injury-risk states were considered eligible if the classification objective was explicitly defined as injury risk in an athlete population. However, these surrogate risk classifications were evaluated separately from models predicting prospectively observed injury events in the synthesis.

2.3 Information sources

Electronic databases (PubMed, Web of Science, and Scopus) were systematically searched from their inception dates to January 15, 2026. To identify eligible studies that might not have been detected in the electronic database searches, the bibliographies of eligible studies and related articles were also reviewed. To prevent potential data loss and broaden the scope of the relevant literature, the bibliographies of

articles meeting the inclusion criteria were manually cross-checked retrospectively.

2.4 Search strategy

The search strategy was structured around three core concepts: athlete populations, machine learning and related AI terminology, and injury prediction or injury risk. Search syntax and field constraints were tailored to the structure of each database. PubMed, Web of Science Core Collection, and Scopus databases were searched from their inception dates, with the last searches performed on January 15, 2026. No publication date restrictions were applied to the database search syntax. Eligibility was limited to original, peer-reviewed research articles published in English. The complete reproducible database-specific search strategies, including all search terms, Boolean operators, truncation, field restrictions, search dates, and the number of records retrieved from each database, are provided in Supplementary Material 1.

During revision, a *post hoc* known-item terminology check was performed using the included studies as a benchmark set to assess the terminological breadth of the reported search strategy. This check indicated that some eligible studies used alternative descriptions, including pattern recognition and classification and regression tree (CART), suggesting that broader predictive-modelling terminology such as risk assessment, forecasting, predictive modelling, classification, and pattern recognition could improve search sensitivity. However, because this was a *post hoc* known-item assessment rather than a rerun of the database searches, it cannot exclude the possibility that additional eligible studies using alternative terminology were missed. Both studies that were not recovered by the core search terminology were identified through reference-list screening, illustrating the practical value of supplementary citation searching while also confirming the terminological limitations of the database search.

2.5 Study selection

All collected data were transferred to the EndNote (v20, Clarivate Analytics, Philadelphia, PA) reference management system, and duplicate records were removed from the system before scanning. Study selection was conducted by two independent researchers (S. O. and M. C. G.) according to pre-defined eligibility criteria, with the first stage being a title/abstract search, followed by a full-text evaluation. Potential disagreements between researchers were resolved by consensus, and in cases where a decision could not be reached, the opinion of a third researcher (S. Y. T.) was sought. All details of the study selection process, including the number of records identified, rejected, and ultimately included, are detailed in the Results section and the relevant PRISMA flowchart (Figure 1).

2.6 Data extraction and data items

Data extraction was performed independently by two researchers (U.C.Y. and S.Y.) using a standardized extraction form that was pilot-tested on three included studies before formal extraction. The pilot study was used to refine the definitions of data extraction items and ensure consistent interpretation of these items; it was not a formal validity study. To ensure the reliability of the process, data were cross-checked, and disagreements were resolved by mutual agreement. In cases where agreement could not be reached, a third researcher (D.E.) acted as an arbiter. The data were systematically organized using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA).

The extracted data items were categorized into three main categories: Characteristics of studies and models (Study design, sports, competition level, sample size, sex, follow-up, specific Injury type, external validation), Predictive performance metrics (Models, indicators, sensitivity, specificity, accuracy, F1/F2 scores and validation approach) and Input variables (External training load, Internal training load, Biomechanical and movement analysis, Neuromuscular and physical performance, Injury history, Psychological and lifestyle factors, Physiological and cardiorespiratory trait and Demographic and anthropometric profiles).

2.7 Risk of Bias assessment

The risk of bias in the included studies was independently assessed by two researchers (S.O. and M.C.G.) using PROBAST (Prediction Model Risk of Bias Assessment Tool), developed for prediction model studies. Prior to the assessment, a pilot calibration was conducted on a limited number of studies to ensure consistent application of the metrics by the researchers. The original PROBAST framework was selected because its explanation and elaboration guidance explicitly permits its application to prediction models using tree-based, machine-learning, and artificial- intelligence methods. Its four-domain structure—participants, predictors, outcome, and analysis—also provided a consistent study-level framework for evaluating risk of bias and applicability across the heterogeneous statistical and machine-learning approaches included in this review. Although PROBAST+AI provides an updated framework that more explicitly distinguishes the quality of model development from the risk of bias in model performance evaluation, the original PROBAST was retained to ensure a consistent assessment across all included studies (27).

The four core domains defined within PROBAST—participants, predictors, outcomes, and analysis—were examined separately. The analysis domain was evaluated in detail in terms of the risk of overfitting, the appropriateness of the validation methods used, and the adequacy of reporting performance metrics. Each domain's signaling questions were initially answered using PROBAST's standard response options: "Yes," "Probably Yes," "Probably No," "No," or "No Information." Following these evaluations, the overall risk of bias for each domain and each study was classified as "Low," "High," or "Unclear." A domain was considered low-risk when all applicable signaling questions were answered "Yes" or "Probably Yes." A high-risk decision was given when one or more methodological issues likely to cause bias in reported model performance were identified. An unclear-risk judgment was used when inadequate reporting prevented a reliable assessment and no clear issue related to high risk could be identified. Decisions in the analysis domain were primarily based on the number of outcome events relative to model complexity (signalling question 4.1), handling of missing data (4.4), consideration of complexities in the data structure (4.6), appropriateness of model performance evaluation (4.7), and methods used to control for overfitting and optimism (4.8). The absence of external validation was considered a limitation regarding the portability of the model; however, it was not used as a justification for a high risk of bias decision alone. Disagreements between researchers were resolved through consensus, and in cases where consensus could not be reached, the opinion of a third researcher (D.E.) was sought. Only the final consensus decisions were retained. The initial independent scores of the evaluators were

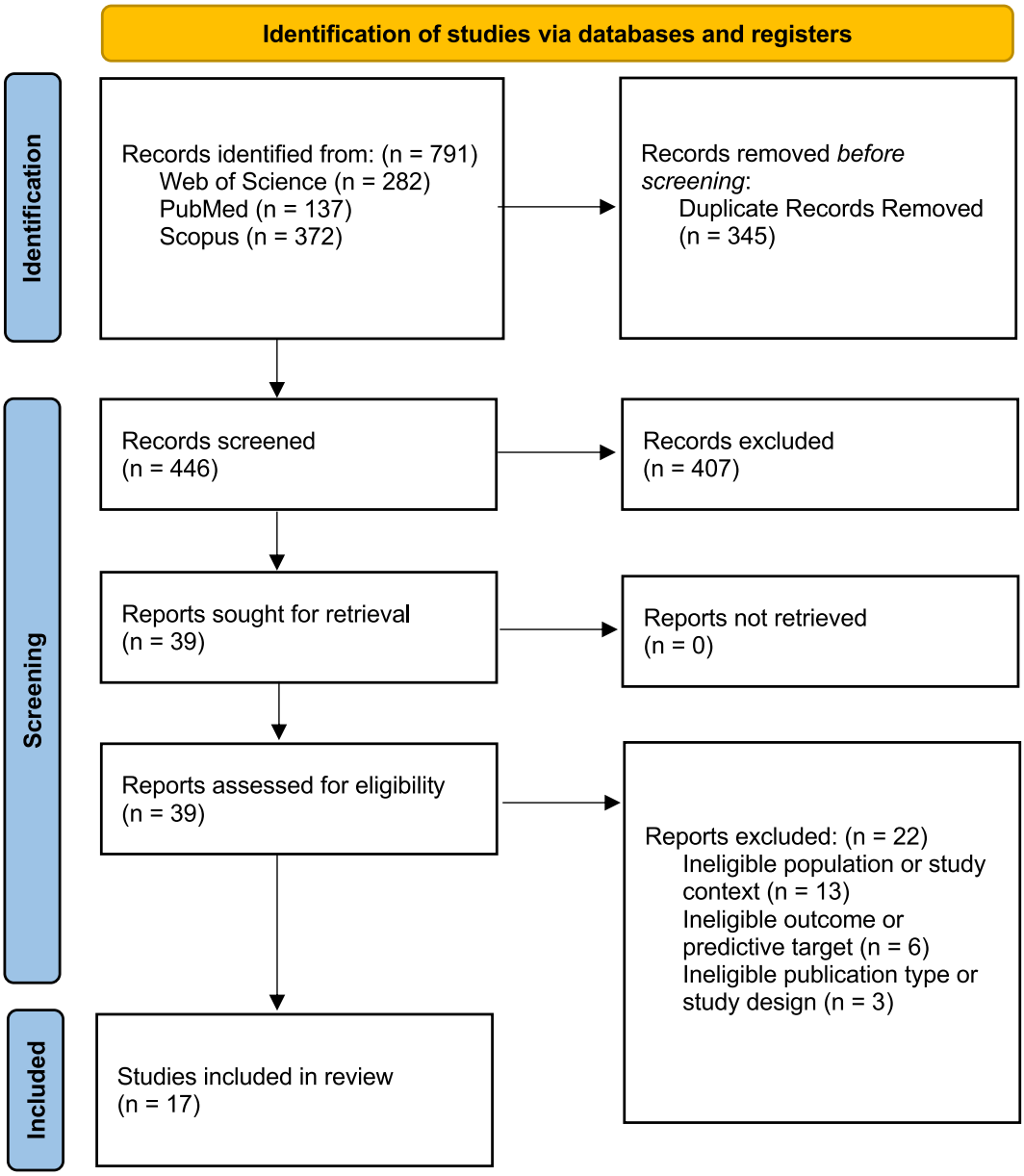


FIGURE 1
PRISMA 2020 flow diagram of the study selection process (Adapted with permission from Page et al. (26), licensed under CC BY).

not kept separately. Therefore, inter-rater agreement statistics such as Cohen κ or percentage agreement could not be calculated retrospectively without creating an artificial estimate.

2.8 Data synthesis

The included studies showed heterogeneity in terms of sample characteristics, injury definitions, machine learning methods used, variables included in the model, and the reporting format of performance metrics. Due to this heterogeneity, it was decided that a quantitative meta-analysis was not appropriate, and the findings were combined using narrative synthesis in accordance with the Synthesis Without Meta-Analysis (SWiM) reporting guideline proposed by Campbell et al. (28). In the synthesis process, the degree to which the models distinguish between injured and non-injured athletes was first evaluated using AUC-ROC values, and then the data types and predictor variables used in model development were classified, and the distribution and frequency of use of these variables among the included studies were examined. For descriptive summarization of AUC values, one representative AUC value was used per study. When multiple AUC values or an AUC range were reported within a study, the median or midpoint of the reported values, respectively, was used to avoid disproportionate weighting of individual studies. Median and interquartile range (IQR) were then calculated across studies reporting AUC. Performance metrics from models classifying surrogate biomechanical risk states rather than observed injury outcomes were described separately and were not included in the descriptive AUC summary of injury-outcome prediction models. All outcome- and prediction-horizon-specific AUC values, the representative value selected for each study, and inclusion in the median/IQR calculation are reported in Supplementary Table S4B. The median and IQR were treated solely as descriptive study-level summaries and

should not be interpreted as pooled estimates of model performance. Because AUPRC and Brier scores are influenced by outcome prevalence, these measures were summarized as reported but were not considered directly comparable across studies with different injury prevalence.

3 Results

3.1 Study selection

A total of 791 records were identified through database searches. Two hundred and eighty-two of these were from Web of Science, 137 from PubMed, and 372 from Scopus. After removing 345 duplicate records, 446 records were screened by title and abstract. Four hundred and seven records that did not meet the predefined eligibility criteria were excluded at this stage, and the full text of the remaining 39 reports was reviewed for eligibility.

As a result of the full-text evaluation, 22 reports were excluded due to reasons related to ineligible population or study context ($n = 13$), ineligible outcome or predictive target ($n = 6$), and ineligible publication type or study design ($n = 3$) (Supplementary Material 3). Consequently, 17 studies that met the inclusion criteria were included in the final synthesis. The details of the study selection process are presented in Figure 1.

3.2 Characteristics of included studies

This systematic review included a total of 17 studies published between 2018 and 2026, covering different sports branches, athlete populations, study designs, and injury outcomes (Table 1; Supplementary Table S4A).

Prospective cohort designs were the most common ($n = 7$) (2, 29–34), followed by retrospective analyses ($n = 5$) (35–39), case–control designs ($n = 2$) (40, 41), two prospective observational studies (42, 43), and one cross-sectional study (44). The smallest study included 25 athletes, while the largest dataset contained over 2000 observations.

A variety of sports were represented, with soccer/football being the most frequently studied ($n = 9$) (2, 30, 32, 33, 37–39, 41, 43), followed by basketball ($n = 4$) (31, 32, 35, 40). While the majority of studies focused on team sports, research examining individual sports such as endurance running and track and field was more limited (36). Samples consisted predominantly of male athletes. Only two studies included only female participants (30, 41) while five studies included mixed-sex groups (29, 32, 34, 36, 44). The level of competition among athletes ranged from sub-elite to professional and elite levels. Follow-up periods also varied considerably, ranging from a few weeks to multiple seasons. There was also significant diversity in terms of injury outcomes examined. Most studies focused on lower extremity musculoskeletal injuries, particularly non-contact muscle injuries and anterior cruciate ligament injuries. Some studies assessed re-injury risk and time loss injuries. Sixteen studies evaluated models against observed injury or reinjury outcomes, whereas one study (44) classified expert-defined biomechanical injury-risk patterns rather than prospectively observed injury events. Model validation predominantly relied on internal approaches such as cross-validation and training-test splits. None of the included studies reported external validation using an independent dataset.

3.3 Predictive performance

The findings regarding the predictive performance of the included studies are presented in Table 2. Reported AUC values ranged from 0.63 to 0.99 (37, 41). Across the 15 studies reporting AUC for observed injury outcomes, the descriptive median of the representative study-level AUC values was 0.80 (IQR, 0.75–0.83); this was not a pooled estimate. Most individual studies reported AUC values between 0.70 and 0.84 (34, 40, 42). However, some studies reported lower AUC values. For example, a study aiming to predict anterior cruciate ligament injuries had an average AUC value of 0.63 (41). In contrast, several studies reported higher AUC values above 0.90, including values of 0.97 and 0.99 in different cohorts (37, 39). Alzahrani et al. reported an AUC of 0.93 for classification of expert-defined biomechanical injury-risk patterns rather than prospectively observed injury events (44).

Sensitivity and specificity were reported in a subset of studies. Across the reported injury outcomes, sensitivity ranged from 30% for the lower-limb injury model reported by Evans et al. (42) to 100% in the study by Fältström et al. (30). Specificity ranged from 39% for the severe lower-limb injury model reported by Evans et al. (42) to 99.5% in the study by Guo et al. (31). Considerable variability was observed both between studies and across different injury outcomes within the same study.

Accuracy and F1-scores were inconsistently reported across studies. Ordinary accuracy values ranged from 73.6% to 96.2% among studies reporting this metric (31, 42), whereas Martins et al. (33) reported a balanced accuracy of 72%. F1-scores ranged from 0.02 to 0.957 (2.0%–95.7%) (35, 37). Notably, Cohan et al. (35) reported an accuracy of 93.4% alongside an F1-score of 0.02 and an AUPRC of 0.009. Similarly, Guo et al. (31) reported an accuracy of 96.2% but a substantially lower F1-score of 0.613. These discrepancies indicate that overall accuracy may provide an overly optimistic impression of model performance in highly imbalanced injury datasets.

Validation approaches varied across studies. The majority of studies employed validation methods, including cross-validation (31, 37, 40, 42), stratified cross-validation (2, 31, 33), and repeated cross-validation (38). Other approaches included hold-out validation (29, 32, 35), bootstrap resampling (40), and leave-one-out cross-validation (38, 42). One study reported an author-described temporal evaluation based on an athlete-disjoint hold-out set (36). Supplementary Table S4B reports whether the performance estimates were obtained from apparent/in-sample evaluation, cross-validation, or held-out testing and identifies estimates potentially affected by preprocessing-related data leakage. Because these evaluation designs provide different levels of protection against optimism, their performance estimates were not interpreted as methodologically equivalent. Potential preprocessing-related data leakage was identified as a separate methodological concern. In Abasi et al. (39), imputation, scaling, polynomial feature construction, feature selection, and SMOTE were described without clear evidence that all preprocessing steps were fitted separately within each cross-validation training fold. Guo et al. (31) implemented data augmentation with Gaussian noise before the initial data splitting operation; however, SMOTE was then limited to the training subsets of cross-validation. Martins et al. (33) reported SMOTE before model training, but its placement relative to the individual stratified cross-validation folds was not clearly described. These workflows may allow information from preprocessing operations or synthetic samples to

10.3389/fpubh.2026.1938641

TABLE 1 Methodological, population, and injury-related characteristics of the included studies.

Author(s)/ Year	Study design	Sport(s)	Competition level	Sample size	Sex	Follow-up	Injury type targeted	Model(s) used	External validation
M. Henriquez et al. (2020) (32)	Prospective cohort	Basketball, soccer, football, volleyball	Sub-elite	122	51F, 71 M	12 months	Lower extremity MSK injuries	Random Forest	No
S. Lövdal et al. (2021) (36)	Retrospective analysis of prospective cohort	Middle/ long-distance running	National to international	74	27F, 47 M	7 years	Any training-limiting injury	Bagged XGBoost	No
A. López-Valenciano et al. (2018) (2)	Prospective cohort	Soccer, Handball	Professional	132	M	9 months	Lower extremity muscle injuries	Cost-sensitive SmoteBagging (CS-SBAG) with ADTree	No
T. Tsilimigkras et al. (2024) (38)	Retrospective analysis	Soccer	Professional	25	M	3 seasons	Non-contact muscle injuries	SVM with RBF kernel	No
A. Rossi et al. (2018) (43)	Prospective observational	Soccer	Professional	26	M	23 weeks	Non-contact time-loss injuries	Decision Tree, Random Forest, Logistic Regression	No
C. C. Claros et al. (2025) (29)	Prospective cohort	Multiple NCAA	Sub-elite	194	121F, 73 M	12 months post-RTP	Post-concussion MSK injuries	L2-regularized logistic regression	No
Y. Lu et al. (2022) (40)	Case–control	Basketball	Professional	2,103	M	20 years	Lower extremity muscle strains	XGBoost	No
A. Cohan et al. (2021) (35)	Retrospective analysis	Basketball	Professional	856	N/R	10 seasons	Multiple injury types	METIC (deep learning / transformers)	No
S. Jauhiainen et al. (2022) (41)	Case–control	Handball, soccer	Elite	791	F	8 years	ACL injuries	Linear SVM	No
I. Ruiz-Pérez et al. (2021) (34)	Prospective cohort	Futsal	Elite	139	72 M, 67F	1 season	Lower extremity non-contact soft tissue injuries	UnderBagging + cost-sensitive SMO	No
J. Ma et al. (2025) (37)	Retrospective	Football	Sub-elite	800	N/R	6 months	Time-loss MSK injuries (≥7 days)	SVM	No
A. Abasi et al. (2025) (39)	Retrospective analysis	Soccer	Elite	256	N/R	8 months	Reinjury risk	CatBoost, SVM	No
A. Alzahrani et al. (2026) (44)	Cross-sectional	Sprint, weightlifting, jumping	Semi-professional to professional	50	35 M, 15F	N/R	Biomechanical injury-risk patterns	Hybrid IMU–sEMG Bi-LSTM	No
L. Guo et al. (2025) (31)	Prospective cohort	Basketball	Sub-elite	104	M	12 months	ACL injuries	Random Forest	No
A. Fältström et al. (2021) (30)	Prospective cohort	Soccer	Sub-elite to elite	117	F	2 years	Second ACL injury	CART	No

(Continued)

TABLE 1 (Continued)

Author(s)/ Year	Study design	Sport(s)	Competition level	Sample size	Sex	Follow-up	Injury type targeted	Model(s) used	External validation
S. Evans et al. (2024) (42)	Prospective observational	Rugby	Semi-professional	36	M	2 years	Non-contact lower limb injuries	Bayesian pattern recognition	No
F. Martins et al. (2025) (33)	Prospective cohort	Football	Professional	96	M	4 seasons	Time-loss muscle injuries	Kstar	No

MSK, musculoskeletal; ACL, anterior cruciate ligament; RTP, return to play; SVM, support vector machine; SMO, sequential minimal optimization; LSTM, long short-term memory; IMU, inertial measurement unit; sEMG, surface electromyography; N/R, not reported; F, female; M, male.

influence the validation fold, thus leading to an optimistic estimate of internal validation performance.

3.4 Input features and predictors

The types of input features varied across studies, reflecting differences in data sources, sport contexts, and injury outcomes (Table 3). The most frequently used feature domains included external training load ($n = 5$) (33, 36, 38, 42, 43), neuromuscular performance ($n = 6$) (2, 30, 32, 34, 37, 42), and injury history ($n = 7$) (29, 30, 32, 35, 40, 42, 43). External load variables commonly consisted of distance, speed zones, acceleration, and session-based workload metrics, whereas internal load measures (33, 36, 38) included heart rate, perceived recovery and perceived exertion ($n = 3$). Biomechanical variables ($n = 4$) (2, 31, 41, 44) encompassed movement patterns, joint kinematics, strength, and functional performance tests. Psychological and lifestyle variables were evaluated as candidate predictors in five studies (2, 29, 30, 34, 37), whereas physiological ($n = 2$) (41, 42) and demographic variables ($n = 4$) (32, 40, 42, 43) were less frequently reported.

The temporal scope of prediction varied across studies. Some models were based on short-term windows (36), including daily or weekly predictions using recent training data. Other studies used longer observation periods, such as multi-week pre-injury windows or season-level predictions based on pre-season assessments. Real-time prediction using wearable sensor data was reported in a limited number of studies (44).

3.5 Risk of Bias

The risk of bias in the included studies was assessed using the PROBAST tool, and the distribution of decisions by domain is presented in Figure 2. Eight of the included studies were classified as high risk (31, 34, 35, 37, 38, 40, 42, 43), and seven studies were assessed in the unclear risk category (2, 30, 32, 33, 36, 39, 44). Only two studies were assessed as low risk (29, 41). The highest high risk ratio at the domain level was observed in the Analysis domain. Seven studies were classified as high risk, six as unclear risk, and four as low risk. The main recurring concerns in the analysis domain included limited sample sizes or outcome-event counts relative to model complexity, feature selection or hyperparameter tuning without clearly nested validation, inadequate or unclear handling of missing data and class imbalance, resampling or augmentation before data partitioning or outside the training folds, limited assessment of calibration, incomplete reporting of relevant performance measures, and insufficient control of overfitting and optimism. The absence of independent external validation was considered separately as a limitation of model transportability rather than an automatic source of high risk of bias. In contrast, in the Predictors domain, 16 studies were assessed as low risk, one as unclear risk, and no studies received a high risk rating. In the Outcome domain, 11 studies were assessed as low risk, six as unclear risk, and no high risk studies were found. Assessments in the participant domain were more heterogeneous. Twelve studies were classified as low risk, three as unclear risk, and two as high risk. The unclear or high risk decisions were primarily related to concerns about sample representativeness and the participant selection process. Detailed study-by-study assessments are presented in the traffic light plot in Figure 2, and the aggregate bias risk distribution by domain is presented in Figure 3.

4 Discussion

This review synthesizes evidence on machine learning-based injury prediction models in athlete populations. The findings show that reported AUC-ROC values varied widely from 0.63 to 0.99, with most individual studies reporting values between 0.70 and 0.84. Although values of $\geq$0.70 are sometimes interpreted as indicating acceptable discrimination, they do not necessarily demonstrate clinical utility (15). Reviewing the included evidence suggests that the high AUC values reported in some studies may be associated with small sample sizes, limited injury types, or restricted population diversity. This could increase the risk of algorithmic overfitting, leading to optimistic performance predictions that do not reflect the actual risks on the field (19). In contrast, current evidence suggests that models evaluated on large and heterogeneous datasets have reported lower discriminatory power but greater overlap with actual risk conditions. This is particularly evident in studies predicting low-frequency outcomes such as anterior cruciate ligament injuries (41).

4.1 Predictive performance

4.1.1 Outcome type and performance

The nature and epidemiological frequency of the targeted injury may influence model performance (21). Models predicting broad outcomes such as running injuries, lower-extremity injuries, or muscle complaints leading to training loss reported AUC values between 0.70 and 0.84 (2, 32–34, 40, 42). However, because these studies differed in outcome definition, prevalence, sample size, and validation design, this pattern should not be interpreted as evidence

TABLE 2 Predictive performance and validation characteristics of the included models.

Author(s)/ Year	Model(s)	AUC	Sensitivity	Specificity	Other metrics	AUPRC	Brier score	Validation approach
M. Henriquez et al. (2020) (32)	Random Forest	0.79	N/R	N/R	N/R	N/R	N/R	CV + Hold-out
S. Lövdal et al. (2021) (36)	Bagged XGBoost	0.72 (day) 0.68 (week)	58.4% (day) 50.4% (week)	74.1% (day) 74.6% (week)	N/R	N/R	N/R	Athlete-disjoint hold-out
A. López-Valenciano et al. (2018) (2)	Cost-sensitive SmoteBagging (CS-SBAG) + ADTree	0.75	65.9%	79.1%	N/R	N/R	N/R	Stratified CV
T. Tsilimigkras et al. (2024) (38)	SVM-RBF	N/R	73%	85%	Accuracy: 78%	N/R	N/R	LOOCV
A. Rossi et al. (2018) (43)	Decision Tree	0.76	80%	87%	Precision: 50% F1: 64%	N/R	N/R	Repeated stratified 2-fold CV
C. C. Claros et al. (2025) (29)	L2-regularized LR	0.82	79%	N/R	N/R	0.85	N/R	Held-out
Y. Lu et al. (2022) (40)	XGBoost	0.84	N/R	N/R	N/R	N/R	XGBoost: 0.030	Bootstrap
A. Cohan et al. (2021) (35)	METIC	0.80	N/R	N/R	Accuracy: 93.4% F1: 0.02	0.009	N/R	Stratified Hold-out
S. Jauhiainen et al. (2022) (41)	Linear SVM	0.63	N/R	N/R	N/R	N/R	N/R	Repeated CV
I. Ruiz-Pérez et al. (2021) (34)	UnderBagging + cost-sensitive SMO	0.77	65.9%	62%	N/R	N/R	N/R	Repeated 5-fold stratified CV
J. Ma et al. (2025) (37)	SVM	0.99	N/R	N/R	Accuracy: 95.6% F1: 95.7%	0.993	SVM: 0.044	Hold-out + 5-fold CV
A. Abasi et al. (2025) (39)	CatBoost / SVM	0.97	N/R	N/R	Accuracy: 91.38% F1: 91.48%	N/R	CatBoost: 0.065; SVM: 0.060	10-fold CV
A. Alzahrani et al. (2026) (44)	Hybrid IMU–sEMG Bi-LSTM	0.93	90.5%	N/R	Accuracy: 92.3%	N/R	N/R	Subject-wise 10-fold CV
L. Guo et al. (2025) (31)	Random Forest	0.80	N/R	99.5%	Accuracy: 96.2% F1: 61.3%	N/R	N/R	Stratified CV
A. Fältström et al. (2021) (30)	CART	0.89	100%	76%	Accuracy: 82%	N/R	N/R	CV
S. Evans et al. (2024) (42)	Bayesian pattern recognition	0.70; 0.76; 0.71	30%; 96%; 86%	93%; 39%; 52%	Accuracy: 73.6%; 81.9%; 75.7%	N/R	N/R	LOOCV
F. Martins et al. (2025) (33)	Kstar	0.81	N/R	N/R	Balanced Accuracy: 72% F1: 70%	0.83	N/R	Stratified CV

AUC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; CART, classification and regression tree; CV, cross-validation; F1, F1-score; IMU, inertial measurement unit; LOOCV, leave-one-out cross-validation; LR, logistic regression; LSTM, long short-term memory; N/R, not reported; RBF, radial basis function; sEMG, surface electromyography; SMO, sequential minimal optimization; SVM, support vector machine. Sensitivity values for Lopez-Valenciano et al. (2) and Ruiz-Perez et al. (34) are those reported in the abstracts of the original articles. The values reported for the selected models in the full texts are provided in Supplementary Table S4B.

TABLE 3 Feature domains used in the included machine learning models.

Feature domain	Studies (*n*)	Variables
External training load	5	Distance, speed zones, acceleration, ACWR
Internal training load	3	Heart rate, RPE, perceived recovery
Biomechanical/movement	4	3D motion capture, joint angles, ground reaction forces, EMG
Neuromuscular/physical performance	6	Strength, ROM, postural stability, hop tests, sprint times
Injury history	7	Previous injury count, recent injury type, concussion history
Psychological/lifestyle	5	Stress level, sleep, burnout, stress susceptibility, adventure seeking
Physiological (cardiorespiratory)	2	HRmax, ventilatory thresholds, HR zones
Demographics/ anthropometrics	4	Age, body mass, position

ACWR, acute: chronic workload ratio; RPE, rating of perceived exertion; ROM, range of motion; EMG, electromyography; HRmax, maximal heart rate.

that broadly defined injuries inherently provide a more stable basis for model development. Studies examining more specific and less frequent outcomes, such as ACL injuries, reported AUC values ranging from 0.63 to 0.89 (30, 31, 41). Furthermore, studies reporting high AUC values present inconsistent findings in secondary performance measures. For example, although Guo et al. (31), reported an AUC of 0.80, the F1 score, a more robust indicator of positive predictive success, was found to be contradictorily low at 0.61. Similarly, in the study by Fältström et al. (30), while the model provided 100% sensitivity, the specificity dropped to 76%. This wide range of sensitivity and specificity values reflects inter-study differences in injury outcome definitions, classification thresholds, and the severity of class disparity (45). In low-frequency injury datasets, classification thresholds optimized for overall accuracy may favor the predominantly non-injury class, while thresholds prioritizing sensitivity may improve injury detection at the expense of specificity. Similarly, inconsistent use of class imbalance strategies, including resampling and cost-sensitive learning, may partly explain inter-study differences in class-dependent performance metrics. On the other hand, studies reporting relatively higher AUC values appear to include models targeting secondary injury or re-injury outcomes in high-risk groups. Therefore, secondary injury and re-injury outcomes may represent a transition zone between common and rare injury events (30, 39). The findings show that the discriminative power of machine learning algorithms cannot be evaluated independently of the epidemiological background of the targeted injury. Considering that data volume, the frequency of injury, and its multidimensionality are related to the statistical results of the model, making predictions based solely on performance metrics risks methodological bias.

4.1.2 Insufficient validation, data leakage, and class imbalance

One of the methodological limitations in the examined evidence is the lack of reliable validation strategies. None of the 17 included studies performed external validation using an independent dataset. This indicates that the generalizability of the reported models beyond the development samples has not been adequately tested. Furthermore, only one study reported an author-described temporal evaluation, implemented as an athlete-disjoint hold-out within the same running team (36). While methods such as LOOCV, bootstrap resampling, and hold-out used in other studies provided estimates of model stability, their generalizability to independent populations has not been sufficiently demonstrated (19). Temporal evaluation can provide information on how model performance changes across prediction horizons. Lövdal et al. (36) showed that the AUC value of a model, which was 0.72 in daily predictions, decreased to 0.68 in the weekly prediction horizon. It is unclear to what extent the performance obtained by the models with other validation methods can be transferred to field applications (18). The performance observed in the development sample may not be maintained in different teams, seasons, competition levels, or settings with different injury frequencies. Class imbalance can exacerbate this problem by encouraging models to learn patterns where the uninjured class is dominant, reducing reliability in populations with different outcome distributions. The absence of external validation limits confidence in the model's portability; however, it does not, by itself, indicate a high risk of bias in model development efforts. High or unclear risk judgments in the analysis domain are instead based on issues identified by relevant PROBAST signaling questions, such as a limited number of outcome events relative to model complexity, inadequate or insufficiently reported handling of missing data and class imbalance, deficiencies in performance evaluation, and insufficient control of overfitting and optimism. Data leakage due to preprocessing poses a separate threat to model validity. Operations such as missing data assignment, scaling, feature generation, feature selection, and synthetic oversampling should be calculated and implemented separately within each training fold. When these operations are performed before data splitting or cross-validation, information from the entire dataset can indirectly affect the validation data, compromising its independence and overstating performance estimates such as AUC, accuracy, and AUPRC. This problem was identified or could not be excluded in the workflows reported by Abasi et al. (39), Guo et al. (31), and Martins et al. (33). Therefore, the performance estimates of these studies should be interpreted as potentially optimistic internal validation results, not as evidence of superior or portable model performance.

Another methodological problem closely related to insufficient validation is class imbalance resulting from low injury frequency. When rare injury events are examined during model optimization, a tendency to prefer the "uninjured" class, which constitutes the majority of the dataset, may arise. This can reduce the model's capacity to accurately identify true injury cases, which is its primary objective (21). In particular, in datasets where the class distribution is unbalanced, the overall accuracy rate may give an optimistic impression of model performance. Therefore, AUPRC, which better reflects the success in predicting the positive class, and the Brier score, which provides information about the overall accuracy of probabilistic estimates, are more informative measures in such data structures (45). These measures were reported in only a small fraction of the included

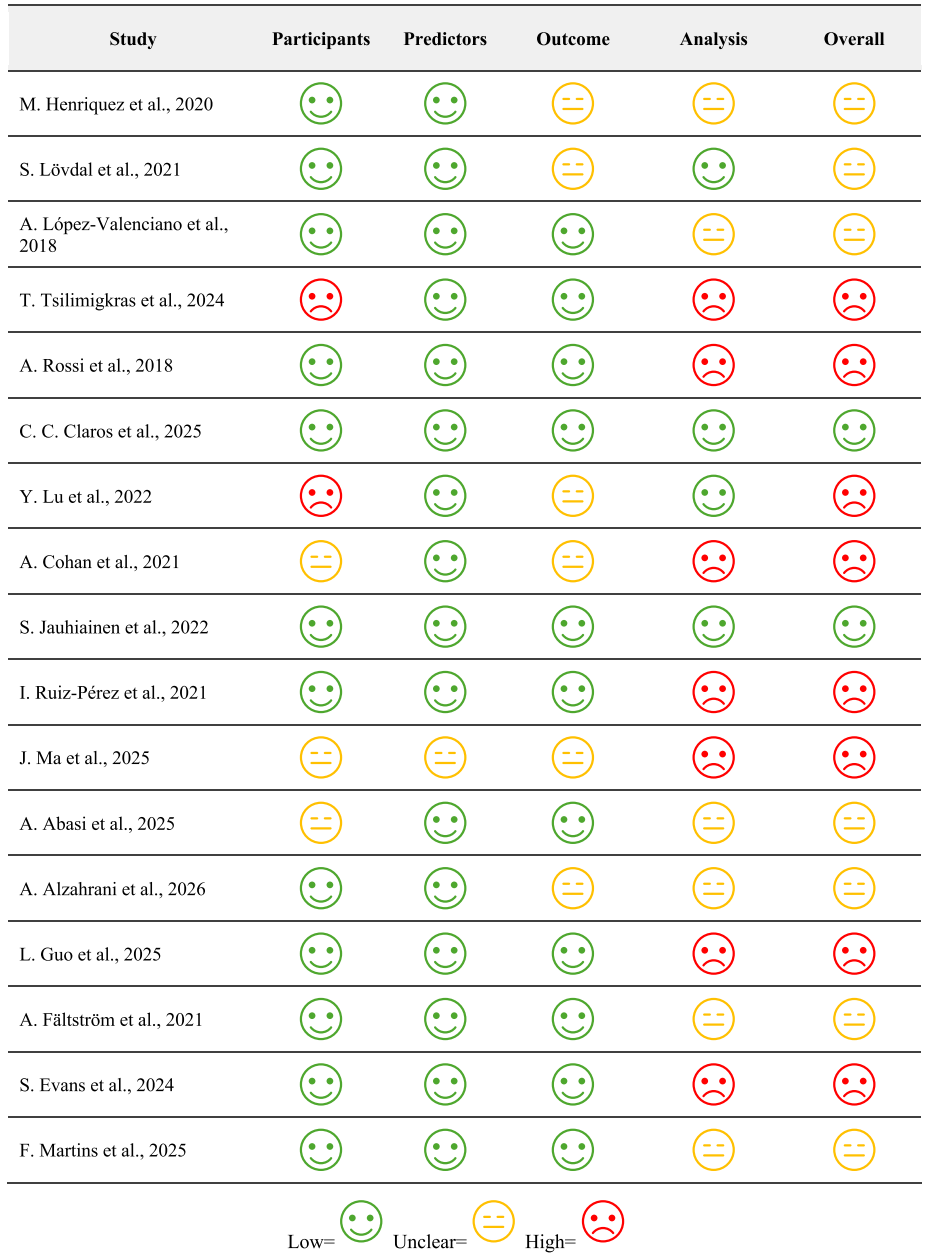


FIGURE 2
Risk of bias assessment for included studies.

studies, limiting the comparison of class-specific performance and probabilistic prediction quality among studies. The fact that Guo et al. (31) reported an overall accuracy rate of 96.2% while the F1 score remained at 0.61 clearly shows that evaluation based solely on accuracy rate can be misleading.

Tree-based ensemble methods, including Random Forest (31, 32), XGBoost (36, 40), and CatBoost (39), were evaluated in the included studies, with predictive performance varying across datasets. Their ability to model nonlinear relationships and interactions and process high-dimensional data may partly explain their frequent use in injury prediction research (21). Support vector machines have also been evaluated in various athlete populations (34, 37–39, 41). However, since the models were not directly compared using common datasets, outcomes, and validation procedures, algorithm-level performance cannot be reliably compared between these studies. Therefore, the observed differences may reflect differences in sample size, injury prevalence, predictor structure, class distribution, and validation design, rather than the structural superiority of a particular algorithm.

Deep learning and hybrid approaches have only been evaluated in a limited number of studies. Cohan et al. (35) compared a Transformer-based model with traditional approaches on a single dataset, while Alzahrani et al. (44) reported an AUC-ROC value of 0.93 for a hybrid LSTM model that classifies biomechanical injury risk patterns defined by experts. However, the target in the second study is a surrogate biomechanical risk state rather than a prospectively observed injury event, and these findings do not demonstrate that deep learning approaches are superior across different athlete populations or injury outcomes. In terms of model interpretability, SHAP-based analyses help to explain the contribution of less frequently assessed variables such as psychological and demographic characteristics to the prediction process. Such

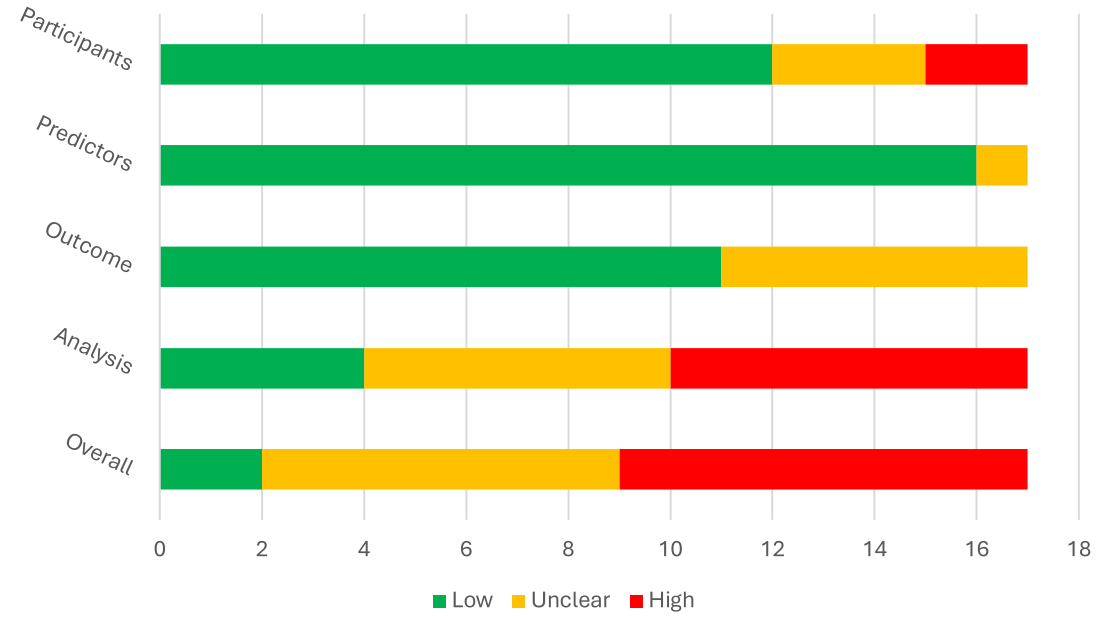


FIGURE 3
Risk of bias summary (PROBAST).

interpretability methods can increase the confidence of coaches and sports scientists in model results by showing how individual predictors contribute to model outputs (37).

4.2 Input features and predictors

The analysis of input variables is examined across various feature domains. In the available evidence, external training load, neuromuscular performance, and injury history are the three most frequently used feature domains. Neuromuscular performance was included in six models, injury history in seven models, and external training load in five models (Table 3). Additionally, these feature domains are consistent with established epidemiological evidence as they encompass parameters routinely monitored by clinicians. Four included studies used biomechanical and movement data as model inputs (2, 31, 41, 44). The scalability of these models in practice may be limited by the need for expensive equipment such as three-dimensional motion analysis or surface electromyography. Beyond biomechanical variables, some studies evaluated psychological and lifestyle variables, including sleep, burnout, stress, psychological readiness, and personality-related characteristics, as candidate predictors during model development (2, 30, 34, 37). This emerging approach highlights the multifactorial nature of injury risk and extends predictions beyond traditional biomechanical and physiological parameters.

4.2.1 Data temporality and multimodal integration

Both the type of input data and its temporal resolution can affect model performance. Models using longitudinal GPS or training load data can account for changes in exposure, fatigue, and physical condition over time (33, 36). In contrast, models based on assessments at a single time point, such as preseason screening batteries, cannot capture changes in these factors that occur throughout the season (2, 31, 32). However, whether the inclusion of time-varying information improves prediction accuracy is an empirical question and likely depends on the injury outcome, the available predictors, and the intended prediction horizon.

While these issues were not directly assessed in most of the included studies, broader application considerations for real-time prediction include sensor and communication latencies, computational requirements of models used in wearable or mobile devices, data security, and the management of sensitive physiological, biomechanical, and location information. Therefore, these considerations should be treated as general considerations for future applications, rather than as direct findings from the evidence synthesized in this review.

Studies have begun to combine longitudinal workload measures with athlete-level characteristics and multimodal wearable-sensor data. Rossi et al. (43) combined GPS-derived external-load features with previous-injury and player-profile variables; the primary decision-tree model achieved an AUC of 0.76 ± 0.12, sensitivity of $80 \pm 7\%$, precision of $50 \pm 11\%$, and an F1-score of $64 \pm 10\%$, whereas Alzahrani et al. (44) reported an AUC of 0.93 for a hybrid model integrating wearable IMU and surface-EMG data to classify expert-defined biomechanical injury-risk patterns. Furthermore, considering that these findings are obtained from small sample sizes and narrowly defined populations, adapting the obtained data to different contexts becomes difficult. Moreover, despite the increasing prevalence of wearable technologies in professional sports environments, the real-time prediction capacity and field applicability of these models have not yet been sufficiently investigated. The fact that only one study in the reviewed literature integrated wearable sensor technologies into real-time prediction models (44) highlights the gap between theoretical model success and field applications.

4.3 Limitations

This review has several limitations. Significant heterogeneity in sample characteristics, injury definitions, model structures, and performance measures among the included studies prevented a quantitative meta-analysis and limited direct comparisons between algorithms.

The inclusion of only English-language peer-reviewed articles may have led to language and publication bias. Furthermore, the focus of only two studies on female athletes and the predominance of male participants raise significant uncertainty regarding the validity and predictive performance of the models in female athlete populations. The conclusions reached in this review are limited by the methodological limitations of the primary studies and the availability of methodological details reported in those studies. In particular, problems identified in the analysis domain of the PROBAST assessment made it difficult to determine the extent to which reported model performance translated into clinical utility. The incomplete or inadequate reporting of class-specific performance measures, including sensitivity, specificity, and F1 score, in many studies also limited comprehensive comparisons between models. A structured TRIPOD+AI assessment was not performed as it was not pre-planned; instead, the review focused on predictive performance, validation features, and the risk of bias. Therefore, observations regarding underreporting relate only to specific items from which data was extracted and should not be interpreted as a formal assessment of overall reporting quality or TRIPOD+AI compliance (25). The use of the original PROBAST tool instead of the updated PROBAST+AI framework should also be considered a limitation. While the original tool is applicable to machine learning-based predictive models and covers the same four core domains, it does not include all the improvements introduced by PROBAST+AI, particularly the clear distinction between model development quality and the risk of bias in evaluating model performance. Therefore, the bias risk findings should be interpreted within the scope of the original PROBAST framework.

Although three complementary databases were searched and reference lists were manually reviewed, the relatively specific search terminology and the lack of use of additional specialist databases such as SPORTDiscus, Embase, and IEEE Xplore may have resulted in eligible studies using alternative terminology being missed. Therefore, the completeness of the identified evidence base cannot be guaranteed.

Temporal dataset shifts can occur when injury patterns vary between seasons due to differences in competition schedules, training exposure, recovery opportunities, or other contextual factors not represented in model predictors. The COVID-19 pandemic exemplifies how infection and changes in training and competition conditions can alter the underlying injury environment (46, 47). However, pandemic-specific effects were not directly assessed in the prediction models included in this review. This underscores the importance of performing temporal and external validation before models are ported to different seasons, teams, or athlete populations.

5 Conclusion

This review shows that while the current evidence base on machine learning-based injury prediction in athletes is promising, it is not yet sufficiently mature methodologically. Although some models exhibit moderate to high discrimination performance, small and often homogeneous samples, class imbalance, limited reporting of calibration, lack of external validation, and very limited temporal validation restrict their practical value. Model performance appears to depend not only on algorithm selection but also on injury prevalence, outcome definition, temporal structure of prediction, and validation design. These limitations restrict the portability of current models across different teams, seasons, competition levels, and athlete populations, hindering their routine use as independent injury prevention decision tools. Future research should prioritize prospective multicenter studies, independent external and temporal validation, cohorts with sufficient statistical power and demographic diversity, transparent reporting of calibration and class-specific performance, and evaluation of whether model-based decisions ultimately reduce injury frequency, severity, or time loss.

6 Practical implications

For clinicians and sports medicine specialists, current machine learning-based injury prediction models should be viewed not as established decision support systems or standalone systems, but as research-stage tools that may complement future decision-making processes. Predictions regarding rare outcomes should be interpreted with particular caution when models are developed using small samples, unbalanced datasets, or without external validation. Model outputs should not be used in isolation if evaluated in research or exploratory applications; they should be interpreted in conjunction with clinical assessment, injury history, and other athlete-specific information. Future research should investigate whether incorporating prediction outputs into longitudinal monitoring processes that combine training load, neuromuscular status, recovery indicators, and injury history provides additional benefit. None of the included studies have shown that model-driven interventions reduced injury incidence, severity, or downtime. Therefore, risk predictions should not automatically trigger training or clinical decisions unless their validity and usefulness have been prospectively demonstrated in similar athlete populations and competitive settings. Future studies should prioritize multicenter cohorts with adequate statistical power, independent external and temporal validation, appropriate handling of class imbalance, and transparent reporting of discrimination, calibration, and class-specific performance metrics.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

SO: Project administration, Data curation, Writing – original draft. DE: Writing – review & editing, Validation, Conceptualization, Methodology, Writing – original draft, Data curation. AE: Conceptualization, Writing – review & editing. AC: Validation, Writing – review & editing. AB: Conceptualization, Visualization, Writing – review & editing. SY: Data curation, Writing – review & editing, Conceptualization. ZT: Writing – review & editing, Investigation, Validation, MG: Validation, Data curation, Writing – review & editing. SYT: Writing – original draft, Conceptualization, Data curation. UY: Methodology, Project administration,

Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. Generative AI was used solely to improve language, grammar, readability, and clarity. It was not used to generate, analyze, or interpret data. All AI-assisted content was critically reviewed, revised, and approved by the author(s). The author(s) verify and take full responsibility for the use of generative AI in the preparation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2026.1938641/full#supplementary-material

SUPPLEMENTARY MATERIAL 1
Database-specific search strategies (PubMed, Scopus and Web of Science Core Collection).

SUPPLEMENTARY MATERIAL 2
PRISMA 2020 checklist.

SUPPLEMENTARY MATERIAL 3
Full-text reports excluded, with reasons.

SUPPLEMENTARY MATERIAL 4
Study-level data extraction matrix (Supplementary Tables S4A and S4B).

References

1. Rubio-Jiménez A, Peña J, Sabata A, Caparrós T, Casals M. Impact and costs of injuries in professional basketball: insights from a four-season analysis. *Apunts Sports Med*. (2025) 60:100482. doi: 10.1016/j.apunsm.2025.100482

2. López-Valenciano A, Ayala F, Puerta JM, De Ste Croix M, Vera-Garcia FJ, Hernández-Sánchez S, et al. A preventive model for muscle injuries: a novel approach based on learning algorithms. *Med Sci Sports Exerc*. (2018) 50:915–27. doi: 10.1249/mss.0000000000001535

3. Cumps E, Verhagen E, Annemans L, Meeusen R. Injury rate and socioeconomic costs resulting from sports injuries in Flanders: data derived from sports insurance statistics 2003. *Br J Sports Med*. (2008) 42:767–72. doi: 10.1136/bjsm.2007.037937

4. Whittaker JL, Roos EM. Risk profile for sport-related post-traumatic knee osteoarthritis. *Br J Sports Med*. (2020) 54:362–3. doi: 10.1136/bjsports-2019-100877

5. Ardern CL, Taylor NF, Feller JA, Webster KE. Fifty-five per cent return to competitive sport following anterior cruciate ligament reconstruction surgery: an updated systematic review and meta-analysis including aspects of physical functioning and contextual factors. *Br J Sports Med*. (2014) 48:1543–52. doi: 10.1136/bjsports-2013-093398

6. Ekstrand J, Hägglund M, Waldén M. Injury incidence and injury patterns in professional football: the UEFA injury study. *Br J Sports Med*. (2011) 45:553–8. doi: 10.1136/bjsm.2009.060582

7. 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. doi: 10.1136/bjsports-2016-096256

8. Cousins BEW, Morris JG, Sunderland C, Bennett AM, Shahtahmassebi G, Cooper SB. Match and training load exposure and time-loss incidence in elite Rugby union players. *Front Physiol*. (2019) 10:1413. doi: 10.3389/fphys.2019.01413

9. Seshadri DR, Li R, Voos JE, Rowbottom J, Alfes CM, Zorman CA, et al. Wearable sensors for monitoring the internal and external workload of the athlete. *npj Digit Med*. (2019) 2:71. doi: 10.1038/s41746-019-0149-2

10. Hägglund M, Waldén M, Ekstrand J. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *Br J Sports Med*. (2006) 40:767–72. doi: 10.1136/bjsm.2006.026609

11. Croisier JL, Ganteaume S, Binet J, Genty M, Ferret JM. Strength imbalances and prevention of hamstring injury in professional soccer players: a prospective study. *Am J Sports Med*. (2008) 36:1469–75. doi: 10.1177/0363546508316764

12. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. (2024) 384:e074819. doi: 10.1136/bmj-2023-074819

13. Gogoi H, Rajpoot YS, Borah P. A prospective cohort study to predict running-related lower limb sports injuries using gait kinematic parameters. *Phys Educ Theory Methodol*. (2021) 21:69–76. doi: 10.17309/tmfv.2021.1.09

14. Zaslav KR, editor. *An International Perspective on Topics in Sports Medicine and Sports Injury*. Rijeka: InTech (2012). (InTech eBooks) (2012). doi: 10.5772/1503

15. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. (2010) 5:1315–6. doi: 10.1097/JTO.0b013e3181ec173d

16. Goggins L, Warren A, Osguthorpe DJ, Peirce N, Wedatilake T, McKay C, et al. Detecting injury risk factors with algorithmic models in elite women's pathway cricket. *Int J Sports Med*. (2022) 43:344–9. doi: 10.1055/a-1502-6824

17. Bittencourt NFN, Meeuwisse WH, Mendonça LD, Nettel-Aguirre A, Ocarino JM, Fonseca ST. 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. doi: 10.1136/bjsports-2015-095850

18. Leckey C, van Dyk N, Doherty C, Lawlor A, Delahunt E. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. *Br J Sports Med*. (2025) 59:491–500. doi: 10.1136/bjsports-2024-108576

19. Ruddy JD, Cormack SJ, Whiteley R, Williams M, Timmins RG, Opar DA. Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Front Physiol*. (2019) 10:829. doi: 10.3389/fphys.2019.00829

20. Nuyts L, De Brabandere A, Van Rossom S, Davis J, Vanwanseele B. Machine-learned-based prediction of lower extremity overuse injuries using pressure plates. *Front Bioeng Biotechnol*. (2022) 10:987118. doi: 10.3389/fbioe.2022.987118

21. Majumdar A, Bakirov R, Hodges D, Scott S, Rees T. Machine learning for understanding and predicting injuries in football. *Sports Med Open*. (2022) 8:73. doi: 10.1186/s40798-022-00465-4

22. Wang P, Wang A-Y, Wang S. Integrating multimodal AI technologies for sports injury prediction and rehabilitation. *J Hum Sport Exerc*. (2026) 21:22–37. doi: 10.55860/w6j5wc21

23. Van Eetvelde H, Mendonça LDM, Ley C, Seil R, Tischer T. Machine learning methods in sport injury prediction and prevention: a systematic review. *J Exp Orthop*. (2021) 8:27. doi: 10.1186/s40634-021-00346-x

24. Seow D, Graham I, Massey A. Prediction models for musculoskeletal injuries in professional sporting activities: a systematic review. *Transl Sports Med*. (2020) 3:505–17. doi: 10.1002/tsm2.181

25. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. (2024) 385:e078378. doi: 10.1136/bmj-2023-078378

26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann T, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *PLoS Med*. (2020) 18:e1003583. doi: 10.1371/journal.pmed.1003583

27. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. (2019) 170:51–8. doi: 10.7326/M18-1376

28. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. (2020) 368:l6890. doi: 10.1136/bmj.l6890

29. Claros-Olivares CC, Anderson MN, Qian W, Brockmeier AJ, Buckley TA. A machine learning model for post-concussion musculoskeletal injury risk in collegiate athletes. *Sports Med*. (2025) 55:1971–82. doi: 10.1007/s40279-025-02196-4

30. Fältström A, Kvist J, Bittencourt NFN, Mendonça LD, Hägglund M. Clinical risk profile for a second anterior cruciate ligament injury in female soccer players after anterior cruciate ligament reconstruction. *Am J Sports Med*. (2021) 49:1421–30. doi: 10.1177/0363546521999109

31. Guo L, Cui Z, Loh WP, Shaharudin S. Prediction of ACL injury incidence and analysis of key features in basketball players based on multi-algorithm models. *PeerJ*. (2025) 13:e20141. doi: 10.7717/peerj.20141

32. Henriquez M, Sumner J, Faherty M, Sell T, Bent B. Machine learning to predict lower extremity musculoskeletal injury risk in student athletes. *Front Sports Act Living*. (2020) 2:576655. doi: 10.3389/fspor.2020.576655

33. Martins F, Sarmento H, Gouveia ÉR, Saveca P, Przednowek K. Machine learning-based prediction of muscle injury risk in professional football: a four-year longitudinal study. *J Clin Med*. (2025) 14:8039. doi: 10.3390/jcm14228039

34. Ruiz-Pérez I, López-Valenciano A, Hernández-Sánchez S, Puerta-Callejón JM, De Ste Croix M, Sainz de Baranda P, 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. doi: 10.3389/fpsyg.2021.610210

35. Cohan A, Schuster J, Fernandez J. A deep learning approach to injury forecasting in NBA basketball. *J Sports Anal*. (2021) 7:277–89. doi: 10.3233/JSA-200529

36. Lövdal SS, Den Hartigh RJ, Azzopardi G. Injury prediction in competitive runners with machine learning. *Int J Sports Physiol Perform*. (2021) 16:1522–31. doi: 10.1123/ijspp.2020-0518

37. Ma J, Liu S, Pei Y. SHAP-based interpretable machine learning for injury risk prediction in university football players: a multi-dimensional data analysis approach. *Sci Rep*. (2025) 15:40252. doi: 10.1038/s41598-025-24144-y

38. Tsilimigkras T, Kakkos I, Matsopoulos GK, Bogdanis GC. Enhancing sports injury risk assessment in soccer through machine learning and training load analysis. *J Sports Sci Med*. (2024) 23:537–47. doi: 10.52082/jssm.2024.537

39. Abasi A, Nazari A, Moezy A, Fatemi Aghda SA. Machine learning models for reinjury risk prediction using cardiopulmonary exercise testing (CPET) data: optimizing athlete recovery. *BioData Min*. (2025) 18:16. doi: 10.1186/s13040-025-00431-2

40. Lu Y, Pareek A, Lavoie-Gagne OZ, Forlenza EM, Patel BH, Reinholz AK, et al. Machine learning for predicting lower extremity muscle strain in National Basketball Association Athletes. *Orthop J Sports Med*. (2022) 10:23259671221111742. doi: 10.1177/23259671221111742

41. Jauhiainen S, Kauppi JP, Krosshaug T, Bahr R, Bartsch J, Äyrämö S. 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. doi: 10.1177/03635465221112095

42. Evans SL, Owen R, Whittaker G, Davis OE, Jones ES, Hardy J, et al. Non-contact lower limb injuries in Rugby union: a two-year pattern recognition analysis of injury risk factors. *PLoS One*. (2024) 19:e0307287. doi: 10.1371/journal.pone.0307287

43. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernàndez J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One*. (2018) 13:e0201264. doi: 10.1371/journal.pone.0201264

44. Alzahrani A, Aljohany M, Alsirhani H. Real-time wearable biomechanics framework for sports injury prevention and rehabilitation optimization. *Sci Rep*. (2026) 16:4436. doi: 10.1038/s41598-025-34551-w

45. Owusu-Adjei M, Ben Hayfron-Acquah J, Frimpong T, Abdul-Salaam G. Imbalanced class distribution and performance evaluation metrics: a systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*. (2023) 2:e0000290. doi: 10.1371/journal.pdig.0000290

46. Miccinilli S, Bravi M, Conti G, Bressi F, Sterzi S, Santacaterina F, et al. SARS-CoV-2 infection increases the risk of muscle injury in professional male soccer players—a retrospective analysis of the Italian and Spanish major leagues. *Infect Dis Rep*. (2023) 15:425–35. doi: 10.3390/idr15040043

47. Corsini A, Bisciotti A, Canonico R, Causarano A, Del Vescovo R, Gatto P, et al. Are football players more prone to muscle injury after COVID-19 infection? The "Italian injury study" during the Serie a championship. *Int J Environ Res Public Health*. (2023) 20:5182. doi: 10.3390/ijerph20065182