RESEARCH ARTICLE

DOI: 10.54392/ijpefs2633

International Journal of
PHYSICAL EDUCATION, FITNESS AND SPORTS

Artificial Intelligence in Sports Science: Evidence Maturity, Athlete Digital Twins, Multimodal Analytics, and a Translation Agenda for 2026–2035

S. Akila [a, *]

[a] Department of Physical Education and Sports, Central University of Tamil Nadu, Thiruvarur-610005, Tamil Nadu, India
* Corresponding Author E-mail: akilas@cutn.ac.in
DOI: https://doi.org/10.54392/ijpefs2633
Received: 25-05-2026; Revised: 09-08-2026; Accepted: 17-08-2026; Published: 29-08-2026

Abstract: Artificial intelligence is increasingly used in sports science, sports medicine, athlete monitoring, physical education, biomechanics, coaching, and performance analysis. However, many studies report technical success without demonstrating whether an approach is transportable beyond the development setting, useful in professional practice, or adequately protective of athlete interests. This systematic scoping review examined peer-reviewed research published from January 2016 to May 1, 2026 on machine learning, deep learning, computer vision, wearable and multimodal analytics, large language models, generative artificial intelligence, and athlete digital twins. Six bibliographic databases were searched, with Google Scholar and citation tracking used as supplementary discovery methods. The search identified 1,132 records; 901 records underwent title and abstract screening, 287 reports were assessed for eligibility, and 32 publications were included. Primary empirical and technical evidence was examined separately from review, conceptual, perspective, and governance literature. Ten fully verified primary empirical studies enabled direct study-level comparison. The strongest evidence came from narrowly defined tasks with explicit reference standards, including motor-skill assessment, biomechanical estimation, markerless motion capture, and athlete tracking. Evidence for injury prediction, wearable systems, generative AI, and athlete digital twins was less mature. Recurring limitations included small or homogeneous samples, single-dataset validation, limited calibration, incomplete reporting of model behaviour, restricted reproducibility, and a lack of prospective or independent evaluation. None of the included publications demonstrated sustained decision impact or attributable improvement in athlete, educational, sports-medicine, or organisational outcomes. Progress will require temporal and external validation, calibrated uncertainty, transparent reporting, representative datasets, interoperable multimodal systems, human oversight, and safeguards for athlete autonomy, privacy, fairness, and contestability.

Keywords: Sports Analytics, Machine Learning, Athlete Monitoring, Computer Vision, Digital Twins, Generative AI, Implementation, Responsible AI.

1. Introduction

Artificial intelligence (AI) is an increasingly important component of sports science, sports medicine, athlete development, physical education, and performance management. The term includes machine learning, deep learning, computer vision, natural language processing, generative AI, and intelligent decision-support systems. These methods can model complex, nonlinear, and multidimensional relationships in structured and multimodal datasets containing competition statistics, training load, physiological signals, wearable-sensor measurements, biomechanical variables, medical records, video, movement trajectories, and contextual information. AI systems are therefore being researched for performance prediction, tactical analysis, injury-risk estimation, movement recognition, rehabilitation, athlete monitoring, coaching, and educational assessment (Claudino *et al.*, 2019; Reis *et al.*, 2024; Souaifi *et al.*, 2025).

Wearable sensors, video-analysis systems, global positioning technologies, inertial measurement units, force platforms, and digital training platforms have all become more available, supporting the expansion of AI in sports. Earlier applications were mostly based on structured performance or medical

data, however more recent studies progressively incorporate physiological, biomechanical, spatial, visual, and behavioral information. This advancement has changed sports analytics away from retrospective description and toward continuous monitoring, pattern identification, and personalized prediction. AI-enhanced sensing and wearable technology reviews show that these systems can help with workload assessment, movement analysis, physiological monitoring, and return-to-sport decision-making, but their practical effectiveness is dependent on sensor validity, data quality, and appropriate human interpretation (Seshadri *et al.,* 2021; Chidambaram *et al.,* 2022). Wearable models have also been investigated for estimating core body temperature and other physiological variables that cannot be measured continuously during training or competition, but methodological differences in sensor selection, reference standards, and validation procedures continue to limit system comparability (Dolson *et al.,* 2022).

Machine learning is being used extensively in research to predict injury risk. AI models can use training exposure, previous injury, neuromuscular features, movement mechanics, recovery indicators, and athlete-profile information to identify individuals who may be at elevated risk. Systematic reviews indicate that machine-learning models can detect relationships that conventional analysis cannot, but the underlying evidence is heterogeneous in terms of sample size, outcome definition, feature selection, and validation strategy (Van Eetvelde *et al.,* 2021; Musat *et al.,* 2024). The distinction between statistical discrimination and effective injury prevention is very crucial. In a case-control study of 791 female elite handball and football players and 60 anterior cruciate ligament injuries, an extensive screening battery provided only limited predictive utility, with the best-performing model achieving a mean receiver operating characteristic area under the curve of approximately 0.63 (Jauhiainen *et al.,* 2022). This research illustrates that technically complex models may not always produce relevant sports-medicine predictions, particularly when injury events are rare, multifactorial, and influenced by factors not recorded during baseline screening.

Recent injury prediction studies have reported substantially higher internal performance. Ma *et al.* (2025) used ten machine-learning algorithms on a dataset of 800 university football players and found high accuracy, F1 score, and receiver operating characteristic area under the curve, with stress, sleep duration, and balance emerging as important predictors through SHAP-based interpretation. Raju *et al.* (2026) employed supervised machine learning and explainability analysis to predict injury risk based on workload, recovery, and athlete characteristics. Nonetheless, both studies depended on a single dataset and lacked independent external validation. Such findings should be viewed as evidence of retrospective model performance rather than proof that the models will be accurate across institutions, sports, competitive seasons, or athlete populations. The lack of external testing, prospective evaluation, and calibration analysis is a persistent problem in sports-injury AI research.

Biomechanics and movement analysis have also progressed quickly. Computer vision and deep learning can extract human pose, joint position, and movement phase information from standard video, thus removing the need for costly marker-based laboratory equipment. Machine- and deep-learning algorithms have already been applied to sport-specific movement recognition, although reported accuracy is highly affected by preprocessing, experimental design, sensor placement and evaluation methods. Recent reviews of sports biomechanics and motion-capture technologies indicate that markerless systems could broaden access to biomechanical assessment, but their validity remains dependent on the sporting task, camera arrangement, occlusion, movement speed and anatomical region being measured (Souaifi *et al.,*2025;Adlou *et al.,*2025).

The empirical studies demonstrate both their promise and current limitations. Mundt *et al.* (2022) employed artificial neural networks to estimate three-dimensional ground reaction forces from two-dimensional position data collected during running, sidestepping, and crossover movements. Their investigation of 1,474 usable movement samples from 14 participants revealed that the pose outputs from some estimation systems were somewhat interchangeable, while others yielded poorer detection performance. Johnson *et al.* (2019) used a convolutional neural network and multivariate regression to estimate knee-joint moments during sports-related movements in 30 people, finding the strongest mean correlation of 0.8895 with reference biomechanical estimations. Ishida *et al.* (2024) tested two-dimensional pose-estimation artificial intelligence on 18 healthy male participants to estimate peak vertical ground reaction force during single-leg landings. These investigations show that video- and sensor-derived models can approximate laboratory biomechanical measures, but

they were carried out under controlled conditions with small, homogeneous samples.

Sport-specific markerless motion-capture studies show that performance varies depending on the joint and movement under study. Lahkar *et al.* (2022) tested a markerless system against a 12-camera optoelectronic reference during standardised boxing movements. Shoulder and wrist joint-center errors were frequently less than 2.5 cm, whereas elbow errors exceeded 3 cm. Schortgen *et al.* (2026) applied markerless analysis to authentic competition conditions by tracking 22 elite amateur boxers over 18 rounds with monocular video and self-supervised detection. The study confirmed the possibility of athlete-position monitoring despite repeated occlusion, however it focused on technical robustness rather than improving coaching decisions or performance outcomes. Collectively, these studies show that computer vision is moving beyond laboratory feasibility to field deployment, while external validation, standardised error thresholds, and proof of decision impact are still restricted.

AI is also being used in physical education and motor skill assessments. Digital-intelligent technologies are being employed to help with instructional visualisation, personalized feedback, movement assessment, and adaptive learning. Systematic and comprehensive reviews indicate that these systems have the potential to create a more continuous instructional cycle by linking performance capture, automated evaluation, and targeted feedback, but teacher competence, privacy, infrastructure, and digital equity remain significant barriers (Wang *et al.,* 2024; Zhong *et al.,* 2025). Empirical studies provide preliminary evidence of practical value. Makaruk *et al.* (2025) conducted an AI-enhanced assessment of jumping-rope performance on 236 individuals and found excellent agreement with expert assessment, with an intraclass correlation coefficient of 0.96 and weighted kappa of 0.87. He *et al.* (2024) demonstrated better 400-meter running performance and positive user satisfaction responses after implementing interactive visual AI in a small group of university students. However, these studies were carried out in small educational settings and do not yet prove long-term learning effects, transferability across institutions, or equitable performance across varied student groups.

The research is also shifting toward foundational models and generative AI. Large language models can generate conversational feedback, training regimens, instructional explanations, and data summaries for athletes. Their apparent adaptability makes them ideal for coaching, rehabilitation, sports education, and athlete support services. However, unlike traditional predictive models trained for specific outcomes, generative systems may deliver plausible but inaccurate, incomplete, or contextually unsafe recommendations. Dergaa *et al.* (2024) discovered that GPT-4 could produce apparently useful exercise-prescription content, but expert verification was still required because the suggestions were not consistently specified or sufficiently contextualized. Vitale *et al.* (2026) observed that large language models might provide potentially useful advice on sleep and jet lag, but expert evaluation identified errors that precluded uncritical athlete-facing application. As such, generative AI should be viewed as supervised decision support rather than an autonomous replacement for coaches, therapists, or sports scientists.

Athlete digital twins extend this move towards individualisation. A digital twin is more than a digital record or a static athlete profile. In theory, it is an individualised computational representation that is updated with athlete data and can simulate future physiological, biomechanical, or performance responses. Existing research includes digital coaching structures, artificial sport trainers, and physiology-based simulations (Gámez Díaz *et al.,* 2020; Fister Jr. *et al.,* 2021). Boillet *et al.* (2024) created an individualised physiology-based digital twin of a cyclist by reinterpreting the Margaria-Morton model and comparing simulated behaviour with field performance and physiological measurements. Related modelling has been used to investigate how variations in rowing distance may affect energy-system contributions and performance needs (Boillet *et al.,* 2025). Although these studies show predictive and simulation potential, many systems referred to as digital twins continue to function as digital models or digital shadows due to a lack of continuous updating, bidirectional interaction, prospective testing, and closed loop adaptation.

Technical advancements have been rapid, but their implementation in routine sports practice remains limited. The existing evidence base consists primarily of reviews, retrospective studies, laboratory validations, small educational interventions, and technical prototypes. Cross-validation inside a single dataset is commonly seen as strong validation, despite the fact that it does not demonstrate transportability to new teams, competitions, or devices. Calibration is rarely reported, thus a model may appropriately rank athletes despite generating inaccurate absolute risk estimations.

Explainability is also inconsistent. Feature-importance approaches, such as SHAP, can discover variables that influence a prediction, but they do not demonstrate causality or ensure that coaches and athletes can understand or act on the results.

Representation also requires careful attention. Claudino *et al.* (2019) discovered that the evidence synthesized across team-sport AI studies was largely male-dominated, with male athletes accounting for nearly 97% of the pooled athlete sample. Small single-sport cohorts, limited female representation, narrow age ranges, and insufficient inclusion of paraathletes or lower-resource sporting environments can all contribute to systematic performance discrepancies in AI systems. According to ethical analysis, major unresolved issues include privacy, fairness, transparency, and accountability (Kim *et al.,* 2025). Continuous monitoring may help performance management, but it can also increase surveillance, weaken meaningful consent, and raise questions about athlete data ownership, secondary usage, and commercial control. AI deployment should be evaluated not only based on predictive performance, but also on athlete autonomy, human oversight, contestability, and organizational accountability.

Existing reviews have focused on team sports prediction, injury prevention, wearables, physical activity, biomechanics, movement recognition, physical education, and AI ethics (Claudino *et al.,* 2019; Van Eetvelde *et al.,* 2021; An *et al.,* 2024; Kim *et al.,* 2025). These topics are typically explored in isolation. A comprehensive synthesis is required to connect AI methods with stakeholders, data modalities, decision types, validation quality, real-world application, and evidential maturity. Such an approach is especially crucial for distinguishing technically interesting prototypes from systems that have undergone external validation, field evaluation, proven decision impact, or enhanced athlete, educational, sports-medicine, or organizational outcomes.

This systematic scoping review synthesises peer-reviewed evidence on AI applications in sports science, sports medicine, and physical education published between 2016 and 2026. It examines performance prediction, injury risk assessment, wearable analytics, biomechanics, computer vision, educational applications, foundation models, generative AI, and athlete digital twins. The review categorizes the collected literature by application domain, stakeholder, decision type, data modality, and evidence maturity. It also investigates scientific validity, predictive performance, out-of-sample evaluation, real-world value, transparency, athlete rights, and implementation readiness using the SPORT-AI Translation Framework. By distinguishing internal technical accuracy from external transportability, decision impact, and meaningful outcomes, the review seeks to establish the methodological, ethical, and implementation needs for responsible sports AI development between 2026 and 2035.

2. Materials and Methods

2.1 Review Design and Reporting Framework

This study was conducted as a systematic scoping review incorporating descriptive evidence mapping, design-specific methodological appraisal, technical classification, and evidence-maturity assessment. Reporting followed PRISMA-ScR and used the PRISMA 2020 flow structure where applicable. The purpose was to map heterogeneous evidence and assess translational maturity rather than to estimate pooled intervention effectiveness. A scoping approach was appropriate because the corpus included prediction and validation studies, educational interventions, technical prototypes, simulation studies, expert output-evaluation studies, evidence syntheses, perspectives, and governance publications. The review was not prospectively registered and no public protocol was released; this is acknowledged as a limitation. Before final screening and synthesis, the review team predefined the review period, information sources, eligibility criteria, selection procedures, extraction variables, validation taxonomy, evidence-maturity definitions, and synthesis approach. To improve reproducibility, the final manuscript should be accompanied by the screening rules, data-extraction form, coding definitions, operational SPORT-AI coding manual, and an amendment log documenting any departures from the predefined procedures. These materials should be treated as part of the review record and made available with the supplementary files or in an open repository where feasible.

2.2 Information Sources and Search Period

The search period began in January 2016 and ended in 2026. This time period was chosen to reflect the rapid development of deep learning, wearable sensing, computer vision, multimodal data integration, large language models, generative artificial intelligence, and digital-twin systems in modern sports science. The

six primary databases were Scopus, Web of Science Core Collection, PubMed/MEDLINE, SPORTDiscus, IEEE Xplore, and the ACM Digital Library. Additional searches were carried out using Google Scholar, backward reference screening, forward citation tracking, and related-article searching. These additional approaches helped identify eligible studies that would not have been found through the primary database searches, such as technical conference papers and multidisciplinary studies indexed outside of conventional sports-science databases. The final search was finished on May 1, 2026. The database specific search documentation is listed in Supplementary Table S1.

2.3 Search Strategy

Database specific strategies used controlled vocabulary, free-text keywords, Boolean operators, phrase searching, truncation, and field constraints. Three concept groups were defined to represent AI technologies, sport and human-performance contexts, and application areas. The core structure was: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "computer vision" OR "large language model" OR "generative AI" OR "wearable sensor" OR "multimodal analytics" OR "digital twin") AND ("sport" OR "athlete" OR "sports science" OR "physical education" OR "sports medicine" OR "exercise" OR "human performance") AND ("performance" OR "injury" OR "rehabilitation" OR "biomechanics" OR "coaching" OR "monitoring" OR "decision support"). The syntax was tailored to each database. Google Scholar searches were limited to the first 200 relevance-ranked results for each search setting, and backward and forward citation tracking was applied to all included papers. Table 1 listed the Search strategy, PCC eligibility framework, and study-selection summary.

2.4 Eligibility Criteria

Eligibility was determined using the Population-Concept-Context framework, as recommended for scoping reviews (Peters *et al.,* 2020). Athletes, sports participants, physical education students, coaches, sports scientists, clinicians, rehabilitation specialists, sports organizations, and technical systems meant to assist human-performance decisions were among the populations and data sources considered eligible. Eligible concepts included machine learning, deep learning, neural networks, computer vision, natural language processing, large language models, generative AI, wearable analytics, multimodal systems, and digital twins. Eligible contexts included competitive or recreational sports, physical education, sports medicine, rehabilitation, athlete development, training, biomechanics, field monitoring, and organizational sport settings. Peer-reviewed journal articles and technically substantial peer-reviewed conference papers published in English were eligible. A publication had to include, at a minimum, the sport or human-performance environment, the AI method or system, the expected output or decision, and enough methodological information to establish the design and validation status. Preprints, theses, dissertations, editorials, letters, patents, abstracts without full papers, non-peer-reviewed reports, fan-engagement-only studies, conventional statistical analyses lacking a substantive AI component, and publications failing to meet the minimum reporting threshold were all excluded.

Table 1. Search strategy, PCC eligibility framework, and study-selection summary

Review element	Specification
Design	Systematic scoping review with evidence mapping, design-specific appraisal, and translation-maturity assessment
Purpose	Map heterogeneous evidence and evaluate translation; not estimate pooled effectiveness
Reporting	PRISMA-ScR and PRISMA 2020 flow structure; no registered protocol
Period	January 2016 to 1 May 2026
Sources	Six bibliographic databases plus Google Scholar and citation tracking
Evidence sets	Primary evidence separated from contextual review, conceptual, perspective, and governance evidence
Selection	Two independent manual reviewers; disagreements resolved by discussion
Flow	1,132 identified; 901 screened; 287 assessed; 32 publications included

2.5 Study Selection Process

Two reviewers independently examined titles and abstracts before evaluating potentially eligible reports in detail. The journal and author information were not blinded. Disagreements were addressed through discussion, and no screening software was used. The searches identified 1,132 records. After 231 duplicate and other pre-screening removals, 901 records underwent title and abstract screening, and 614 were excluded. The remaining 287 reports were assessed in detail, and 255 were excluded due to scope mismatch, lack of a substantive AI component, failure to meet the minimum methodological reporting threshold, unavailable full text or unverifiable study details, duplicate or overlapping publication, or insufficient evidence for the review question. A total of 32 publications were included. Each included publication was retrieved independently and evaluated using the same eligibility criteria. The same 32 selected studies are listed in supplementary table S2.

2.6 Data Extraction

To prevent double counting and unequal weighting, the corpus was divided into mutually exclusive evidence sets. Primary evidence comprised prediction or validation studies, experimental or educational interventions, technical prototypes or simulations, and expert evaluations of AI outputs. Contextual evidence comprised systematic, scoping, and narrative reviews, surveys, perspectives, conceptual papers, and governance publications. Primary studies were used for study-level conclusions about methodology, validation, performance, and translation maturity. Contextual publications were used to map domains and interpret the broader research landscape but were not treated as independent primary evidence in direct quantitative comparisons. A primary study entered the fully verified empirical subset only when the population or dataset, sample size, AI approach, validation procedure, primary metric, and principal finding could all be confirmed from the publication and associated materials. If any of these core fields could not be verified, the publication remained eligible for qualitative or contextual interpretation when appropriate but was excluded from direct study-level quantitative comparison. Missing information was coded as not reported and was not inferred. Reviewer-derived appraisal was recorded separately from source-reported information. Supplementary Table S2 should identify the evidence-set assignment and verification status of every included publication so that readers can reproduce this distinction.

2.7 Validation Taxonomy and Methodological Appraisal

Validation was classified into eight categories: internal validation within the development dataset; temporal validation on later observations; organisational or geographic external validation; device external validation; validation in an independently assembled dataset; criterion validity against a reference standard; prospective validation; and authentic field evaluation. Cross-validation within a single dataset was not considered external validation. Criterion validity and transportability were recorded separately because agreement with a reference standard does not establish performance across sites, devices, seasons, or populations. Because the included evidence spanned prediction models, measurement-validation studies, interventions, technical prototypes, simulations, reviews, and expert evaluations, no single risk-of-bias instrument was appropriate across the entire corpus. The review therefore used a structured, design-specific methodological appraisal rather than a pooled quality score. The appraisal examined sample adequacy, outcome definition, feature selection, class imbalance, leakage prevention, train-test separation, calibration, uncertainty, model comparison, sensor and reference validity, prompt and model-version transparency, reproducibility, subgroup assessment, and human oversight. Where relevant, these domains correspond to concerns addressed by established design-specific appraisal frameworks, but the present review did not retrospectively claim formal scoring with tools that were not prospectively applied. Methodological quality was kept conceptually separate from translation maturity.

2.8 Application, Digital-Twin, and Evidence-Maturity Classifications

Application domain, stakeholder, decision type, data modality, design category, and translation evidence maturity were used to classify the studies. Two separate classification systems were intentionally maintained. First, athlete digital twins were classified by functional capability as DT Functional Stages 1–6: digital record, digital model, digital shadow, predictive athlete twin, interactive athlete twin, and closed-loop athlete twin. Second, the strength of translational evidence was classified independently as Evidence Maturity Level 0 (conceptual proposal), Level 1 (prototype or technical

feasibility), Level 2 (retrospective internal validation), Level 3a (independent criterion validation), Level 3b (temporal, device, organisational, geographic, or independently assembled dataset validation), Level 4 (prospective validation or authentic field evaluation), Level 5 (demonstrated decision impact), and Level 6 (demonstrated athlete, educational, sports-medicine, or organisational outcome impact). The prefixes "DT Functional Stage" and "Evidence Maturity Level" are used throughout to prevent readers from interpreting the two scales as interchangeable. Translation maturity reflects stage of evidence, not overall methodological quality; therefore, a technically sound laboratory study may remain at an early evidence-maturity level, whereas a field study may reach Level 4 while still retaining residual risk of bias.

2.9 SPORT-AI Translation Framework

The proposed SPORT-AI Translation Framework assessed scientific validity, predictive performance, out-of-sample validation, real-world utility, transparency, athlete rights, and implementation readiness. The framework is not offered as a validated measurement instrument; rather, it was created as an author-generated analytical structure based on implementation science, responsible-AI principles, prediction-model evaluation, and sport-specific governance problems. Because an aggregate score can conceal a critical weakness, each dimension was assessed independently. The SPORT-AI operational coding manual is shown in supplementary table S3. The translational assessment is listed in supplementary Table S4.

2.10 Evidence Synthesis

The heterogeneity of sports, populations, modalities, model types, outcomes, and validation processes made a meta-analysis inappropriate. The evidence was synthesised thematically and descriptively. The validated primary empirical subset served as the basis for quantitative claims, while the larger research landscape was interpreted through contextual articles. Technical output, predictive or measurement performance, criterion validity, transportability, field feasibility, decision change, intermediate management outcomes, and attributable athlete or organisational outcomes were all distinguished by the outcome hierarchy. Prospective evidence showing AI-informed decisions were applied repeatedly over time and improved decision quality in comparison to an explicit comparator was needed for "sustained decision impact."

3. Results

3.1 Study-Selection Results

A total of 1,132 records were identified through database and supplementary searching. Before title and abstract screening, 231 records were removed as duplicates or other clearly ineligible pre-screening records, leaving 901 records for screening. Of these, 614 were excluded at the title/abstract stage because they did not address a substantive application of AI in sport, exercise, physical education, sports medicine, or human-performance decision support. The remaining 287 reports were assessed for eligibility, and 255 were excluded after detailed assessment.

Full-text exclusion reasons included scope mismatch, absence of a substantive AI component, insufficient methodological reporting to establish the study design or validation status, unavailable or unverifiable study details, duplicate or overlapping publication, and an insufficient evidence base for the review question. Thirty-two publications met the final eligibility criteria and were included in the synthesis. The available review record supports the total number of pre-screening removals but does not preserve a reliable numerical subdivision of the 231 removals by individual reason; therefore, no post hoc breakdown was created. Figure 1 presents the study-selection flow, and the excluded full-text records with reasons should be supplied in the supplementary material.

3.2 Characteristics of the Included Evidence

The 32 publications were published between 2019 and 2026. There were three published in 2019, one in 2020, three in 2021, five in 2022, nine in 2024, eight in 2025, and three in 2026. The rapid acceleration of AI research in sports science is demonstrated by the seventeen papers that were published in 2024 and 2025. The corpus included multiple study designs. Systematic, scoping, narrative, or survey-based evidence syntheses comprised seventeen publications. Predictive and validation research, field and educational applications, technological prototypes, simulations, and expert evaluations of generative-AI output comprised the main body of evidence. One publication was primarily conceptual or perspective-based. Because they did not provide comparable types of effectiveness evidence, these categories were examined independently, as listed in table 2.

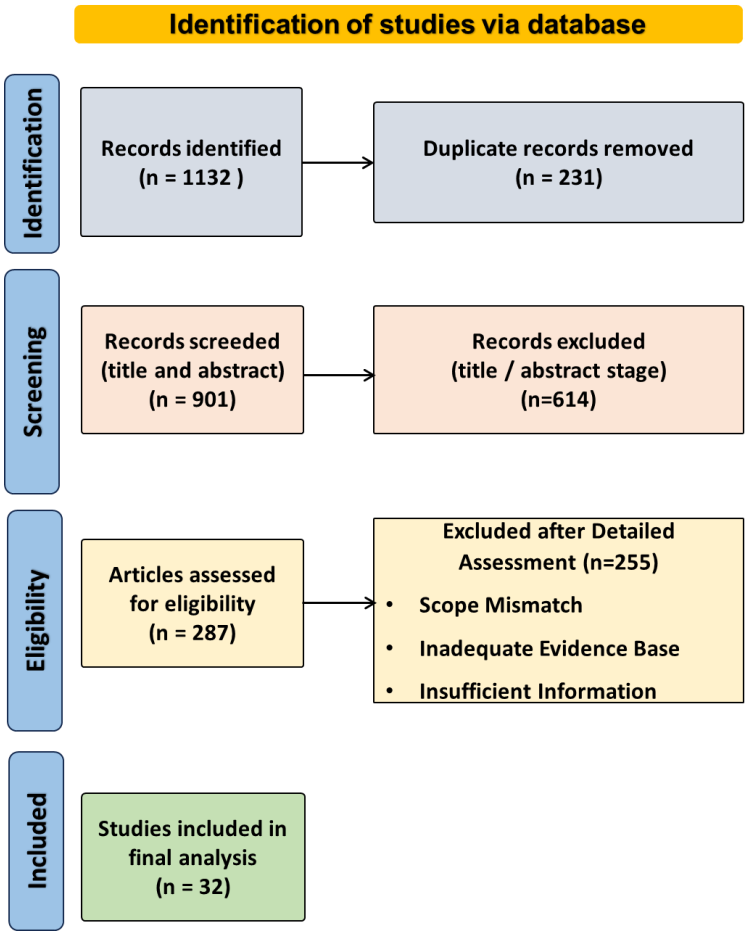

Figure 1. PRISMA-style flow diagram of study identification and selection.

Table 2. Characteristics of the included studies

Characteristic	Finding
Total included publications	32
Evidence-synthesis publications	17 systematic, scoping, narrative, or survey publications
Fully verified empirical subset	10 studies used for direct study-level quantitative comparison
Other primary evidence	Technical prototypes, simulations, field deployments, and expert output evaluations
Publication period	2019–2026; concentrated in 2024–2025
Dominant validation	Within-dataset internal validation and laboratory criterion validation
Highest demonstrated stage	Authentic field feasibility/evaluation; no Level 5 decision-impact or Level 6 attributable outcome-impact evidence

Team sports, football, boxing, cycling, rowing, physical education, general physical activity, sports medicine, injury prevention, sports biomechanics, exercise prescription, monitoring of heat-related illnesses, and athlete recovery from travel were all included in the studies. The populations included schoolchildren, university students, elite athletes, and competitors at the national level. Instead of focusing on a single athlete cohort, several articles synthesised evidence from several sports. Female elite football and handball players, college football players, elite boxers, national-level cyclists, healthy individuals engaging in

sport-related activities, and students taking part in physical education evaluations were all included in the empirical investigations.

Ten studies comprised the verified primary empirical subset. The population or dataset, sample size, AI system, validation process, primary metric, and primary finding were verified for each study. Additional primary publications contributed technical, simulation, prototype, or expert-evaluation evidence, but where a core field could not be validated, they were not used for direct quantitative comparison. The included studies used structured athlete-profile and training-load variables, video, two- and three-dimensional motion data, ground reaction forces, joint moments, physiological measurements, wearable-sensor signals, educational performance data, language-model outputs, and physiology-based simulation parameters. Computational methods included traditional supervised machine learning, support vector machines, random forests, artificial neural networks, convolutional neural networks, pose-estimation systems, markerless motion capture, self-supervised detection, explainable AI, large language models, and mechanistic athlete simulations.

3.3 Distribution of AI Application Domains

3.3.1 Performance Prediction and Tactical Analysis

Performance-related applications were covered in both general reviews and sport-specific technical research. Claudino *et al.* (2019) highlighted performance prediction as one of the most common applications of AI in team sports, specifically football, basketball, handball, and volleyball. Other studies examined movement recognition, physiological simulation, pacing, athlete tracking, and automated performance assessment. The empirical data was more focused on technical measurement than direct tactical decision assistance. Schortgen *et al.* (2026), for example, performed markerless monitoring of 22 best amateur boxers across 18 real competitive rounds. Despite frequent occlusion, the system maintained planar athlete position tracking, demonstrating that monocular computer vision can function under real-world competition situations. However, the study did not assess if the output enhanced tactical decision-making, coaching quality, or match results. Boillet *et al.* (2024) employed an individualised physiological model to mimic cycling performance, whereas a related rowing study investigated changes in energy-system contributions over different race distances. These studies demonstrated that athlete-specific simulation can answer performance-planning concerns, but they were predictive or exploratory rather than decision-impact assessments. Figure 2 summarizes the integrated sports-AI ecosystem produced by the included studies, which linked stakeholders, data modalities, computational models, decision contexts, and intended outcomes.

3.3.2 Injury-Risk Prediction and Prevention

Injury prediction was one of the most frequently represented application domains. Reviews indicated that, machine-learning algorithms can discover combinations of workload, biomechanical, physiological, and athlete-profile characteristics that are associated with injury risk. Nonetheless, the examined data was distinguished by variable results, a small number of injury events, inconsistent validation, and insufficient testing across distinct populations. Jauhiainen *et al.* (2022) examined 791 female elite handball and football players, including 60 with subsequent anterior cruciate ligament injuries. Four classifiers were tested with 100 repeated cross-validation cycles. The highest-performing linear support vector machine had a mean area under the receiver operating characteristic curve of around 0.63. The results showed statistically detectable predictive information but insufficient performance for dependable individual injury screening.

In contrast, Ma *et al.* (2025) reported substantially higher internal performance in a public dataset of 800 university football players. The best support vector machine had an accuracy of 95.6%, an F1 score of 95.7%, and an area under the receiver operating characteristic curve of 99.2%, with SHAP indicating stress, sleep duration, and balance as influential predictors. These statistics should be interpreted with caution because outcome construction, class balancing, duplicate or correlated observations, train-test separation, feature leakage, and hyperparameter selection can all significantly increase single dataset performance. The lack of independent external validation and prospective evaluation precludes the conclusion that the reported performance will generalise to other institutions, seasons, or athlete groups. Johnson *et al.* (2019) focused on injury-related biomechanics rather than direct injury incidence. A convolutional neural network was utilized to estimate three-dimensional knee-joint moments from motion-capture data of 30 athletes executing sports-related tasks.

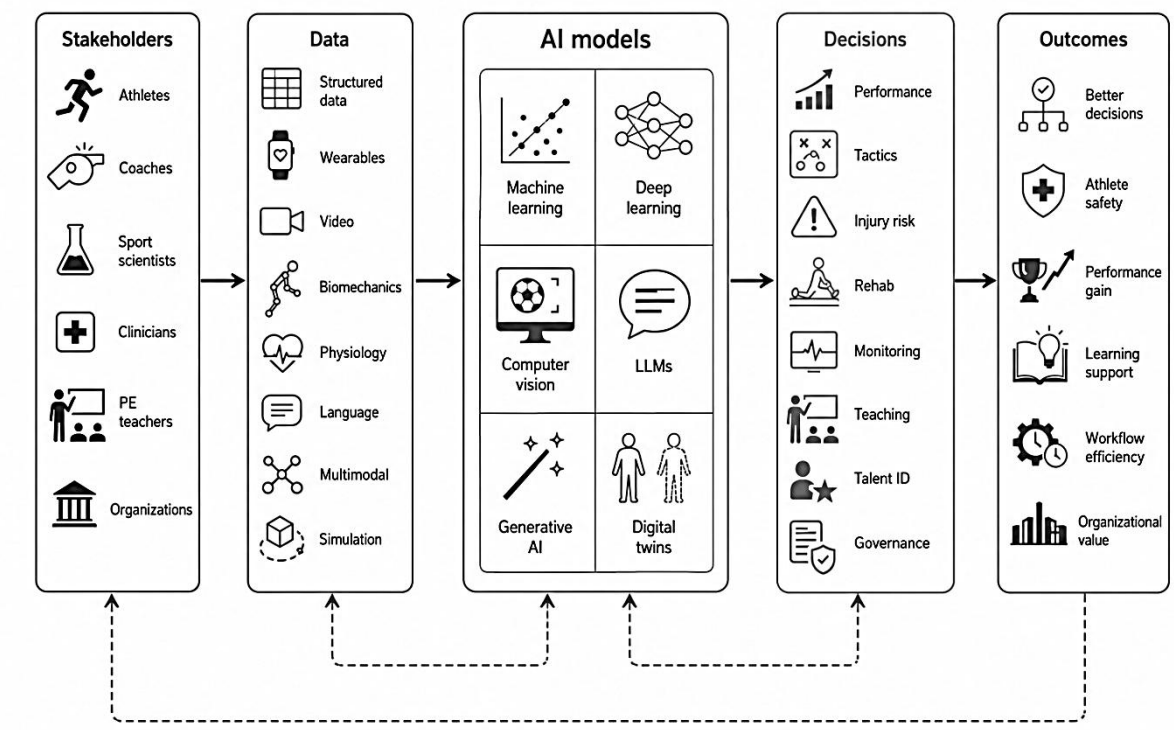

Figure 2. AI application ecosystem in sports science.

The initial phase of stance in sidestepping had the highest mean correlation with reference estimations at 0.8895. This finding confirmed the technical feasibility of portable biomechanical-risk assessment, but no prospective evidence linking the estimates to injury reduction was presented.

3.3.3 Rehabilitation and Sports Medicine

Rehabilitation and sports medicine applications were represented mainly in review-level evidence. The literature discussed possible applications for return to play monitoring, exercise prescription, workload management, physiological surveillance, and movement assessment. Seshadri *et al.* (2021) and Chidambaram *et al.* (2022) emphasized the potential of AI-enhanced wearables to aid in clinical and performance decisions, but they also noted varying sensor validity, inconsistent reporting, and the need for interdisciplinary interpretation. Generative AI was also investigated in relation to exercise prescription. Dergaa *et al.* (2024) discovered that GPT-4 could generate plausible exercise recommendations, but the results were not consistently specific, individualised, or safe enough for unsupervised use. AI was thus positioned primarily as a complement to professional judgment rather than an autonomous clinical or rehabilitation system in the sports-medicine literature. No included article provided Level 6 evidence of persistent, attributable benefits in rehabilitation outcomes, injury incidence, return-to-play duration, educational retention, sports medicine safety, or organizational performance.

3.3.4 Biomechanics and Computer Vision

Biomechanics and computer vision provided the most comprehensive empirical evidence in the verified subset, with the highest concentration of studies using explicit reference standards such as force platforms, marker-based motion capture, expert appraisals, and authentic competition tracking. This distinction refers to validation design and evidence completeness, rather than uniformly high methodological quality or verified athlete-outcome advantage.

Mundt *et al.* (2022) examined 1,474 movement samples from 14 participants who performed running, sidestepping, and crossover activities. Artificial neural networks were utilized to estimate three-dimensional ground reaction forces based on two-dimensional pose estimates. AlphaPose and OpenPose outputs were partially interchangeable, although BlazePose produced inferior keypoint detection and transfer performance. The findings showed that downstream biomechanical accuracy was dependent not just on the prediction model, but also on the pose-estimation system that generated the input.

Ishida *et al.* (2024) tested pose-estimation AI in 18 healthy young men who performed single-leg landings. Two-dimensional video estimations were compared with three-dimensional motion capture and force-plate measurements. The study demonstrated the estimating of predicting peak vertical ground reaction force with readily available video input, however the small homogeneous sample and controlled laboratory environment limited generalizability. Lahkar *et al.* (2022) contrasted markerless and marker-based motion capture in three elite boxers who performed standardised movememnts. Shoulder and wrist joint-centre errors were often less than 2.5 cm, whereas elbow errors exceeded 3 cm. The findings revealed that markerless accuracy varied by anatomical location, and that apparently acceptable overall tracking performance could mask clinically or technically significant joint-specific inaccuracies. Taken together, the biomechanical studies showed a transition from proof-of-concept analysis to criterion validation and authentic field testing. Nonetheless, most studies used small samples and did not demonstrate cross-device repeatability, independent-site performance, or effects on coaching or clinical decisions.

3.3.5 Athlete Monitoring and Wearable Analytics

Wearable and athlete-monitoring applications included workload monitoring, physiological monitoring, injury risk estimation, heat illness prevention, movement recognition, and return-to-play assistance. The evidence came mainly from reviews, rather than prospective deployment studies. Wearable data consisted of physiological signals, cardiovascular measurements, workload indicators, mobility data, inertial signals, and ambient variables. According to the reviewed studies, continuous sensing can capture patterns that intermittent laboratory testing cannot. However, the evidence identified recurring problems with sensor location, signal loss, device drift, proprietary processing, reference-standard selection, and inconsistency in validation. Dolson *et al.* (2022) demonstrated that wearable core body temperature estimation is technically promising but methodologically diverse. Sensor combinations, environmental conditions, reference measurements, and model-validation systems all differed, making direct comparison between systems impossible. Similarly, reviews of wearable injury and workload monitoring revealed that many systems were evaluated retrospectively, with little evidence that real-time outputs influenced athlete management decisions.

3.3.6 Physical Education and Coaching

AI applications in physical education included movement assessment, automatic feedback, instructional visualization, personalized training, and interactive learning. Wang *et al.* (2024) and Zhong *et al.* (2025) reviewed a new instructional cycle in which performance data are collected, evaluated, and translated into focused feedback. However, teacher readiness, infrastructure, privacy, and unequal technological access remained persistent implementation challenges. Makaruk *et al.* (2025) provided the strongest convincing validation evidence in this arena. The study included 236 participants: 126 schoolchildren and 110 university students. A jumping-rope exercise was evaluated by an AI model using five process-oriented criteria. The agreement with expert judgement was excellent, with an intraclass correlation coefficient of 0.96 and weighted kappa of 0.87. Criterion-specific kappa values varied between 0.83 and 0.87, and repeated AI assessments revealed perfect intra-rater consistency. The system thus demonstrated strong reference validity for a defined motor-skill task, despite not being externally verified at an independent institution. He *et al.* (2024) tested interactive visual AI in ten university students. The intervention was associated with improved 400-meter running performance and positive satisfaction responses, with 70% of participants reporting a better understanding of physical-education content. The study demonstrated field feasibility but was restricted by its small sample size, pre-post design, and lack of a strong independent comparator. Gao (2025) also discussed AI-enabled teaching, evaluation, and personalized training in higher education. The study provided a policy and practice viewpoint rather than original outcome evidence, and was therefore used only for contextual interpretation.

3.3.7 Generative AI and Foundation Models

The evidence for foundation models and generative AI was minimal, consisting primarily of expert evaluations, reviews, and perspective pieces. The applications comprised workout prescription, athlete sleep and jet-lag guidance, conversational suport, educational assistance, automated reporting, and interdisciplinary decision support. Dergaa *et al.* (2024) examined GPT-4-generated exercise prescriptions and discovered that the results could be plausible and useful as initial drafts. However, recommendations were not consistently tailored to individual needs or sufficiently contextualized for unsupervised implementation. Vitale et al. (2026) also

tested large language models for athlete sleep and jet-lag guidance. The systems generated potentially helpful responses, but experienced reviewers found inconsistencies and varying reliability. No included study demonstrated a fully validated sport-specific foundation model, a prospective conversational-coaching experiment, a reliable automated clinical-reporting system, or an athlete-outcome improvement attributable to generative AI. Retrieval augmentation, fine-tuning procedures, prompt reproducibility, model-version stability, and thorough hallucination testing were all inconsistently reported. As a result, the evidence remained at the feasibility or controlled expert evaluation level.

3.3.8 Athlete Digital Twins

Surveys, conceptual architectures, technical prototypes, individualised physiological simulations, and performance-modeling studies were represented in the digital twin evidence. The systems' actual functionality differed significantly. Gámez Díaz *et al.* (2020) examined digital-twin coaching structures for physical activity, emphasizing multimodal interaction, privacy, security, and standardization. This work was classified as conceptual or prototype evidence, not as an operational athlete twin. Fister Jr. *et al.* (2021) demonstrated an artificial sports trainer on embedded hardware. The implementation demonstrated technical capability, but it did not establish continuous bidirectional updating or athlete-outcome impact. Boillet *et al.* (2024) provided the most advanced digital-twin example in the included corpus. An individualised physiology based cyclist model was tested against field performance and physiological data. The system met the criteria for a predictive athlete twin by including athlete-specific parameters and producing performance simulations. However, it did not show continuous updating, recurring interaction with training decisions, or closed-loop adaptation. As a result, none of the featured systems met the full requirements for an interactive or closed-loop athlete twin. Several systems labelled or described as digital twins were more accurately classified as digital models, technical prototypes, or predictive athlete simulations.

3.3.9 Ethics and Governance

Ethics and governance were addressed directly in the systematic scoping review by Kim *et al.* (2025) and indirectly across educational, wearable, foundation-model, and digital-twin studies. Kim *et al.* (2025) analyzed 25 research from 397 screened records to identify four primary ethical areas. Privacy and data ethics were covered in 22 studies, fairness and bias in 15, transparency and explainability in 13, and accountability in eight. The governance evidence revealed that technical performance was frequently reviewed more rigorously than consent, data ownership, secondary use, contestability, or responsibility for incorrect recommendations. Athletes' representation was likewise uneven. According to Claudino *et al.* (2019), male athletes accounted for approximately 97% of the team-sport AI literature. This disparity hampered model transferability to female athletes and highlighted the significance of subgroup-specific validation. When the evidence was classified by stakeholder and decision type, the same computational methods were found to support distinct functions across athletes, coaches, physicians, teachers, sports scientists, and organizations, as shown in Figure 3.

3.4 Data Modalities and Computational Methods

Structured tabular data were used primarily to predict injuries, classify performance, and model athlete risk. These datasets included variables such as training load, sleep, stress, recovery, previous injury, demographics, physical fitness tests, and anthropometric data. This group commonly used conventional machine-learning approaches such as support vector machines, random forests, decision trees, and related classifiers. Ensemble approaches were employed to improve classification performance or to compare multiple algorithms, whereas SHAP was utilized in a few studies to interpret key variables. Video and biomechanical data were the primary inputs for computer vision applications. Pose-estimation models extracted two-dimensional keypoints, which were then mapped to ground reaction forces or other biomechanical outcomes using neural networks. Markerless motion-capture systems evaluated joint locations, whereas self-supervised detection aided athlete tracking during competition-related occlusion. As a result, deep neural networks were most commonly used in movement analysis, joint load calculation, and vision-based tracking.

Physiological and wearable data included temperature signals, workload indices, mobility signals, and other continuously recorded variables. These data were analyzed using regression, conventional machine learning, deep learning, and multimodal integration.

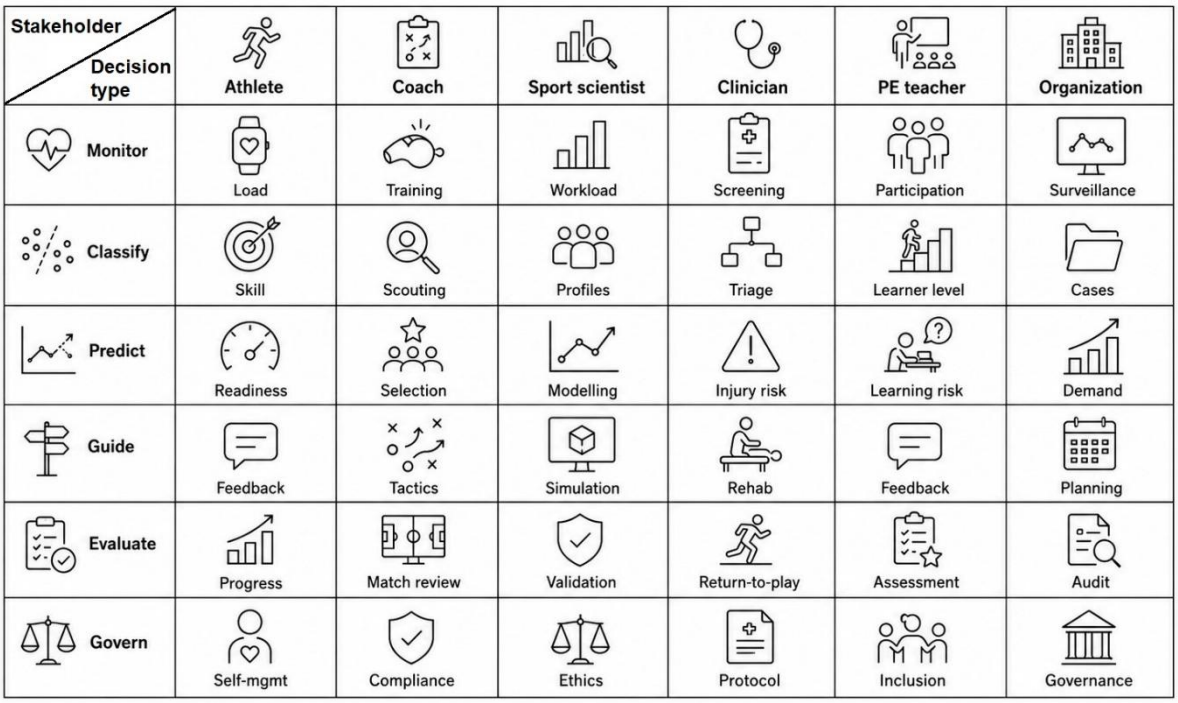

Figure 3. Taxonomy of AI applications by stakeholder and decision type.

However, the details of sampling frequency, synchronization, missing-data management, and signal drift were described inconsistently throughout the studied literature. Inertial sensing was used in movement detection and wearable monitoring research. The review evidence revealed the widespread use of inertial measurement devices for sport-action classification, but methodological variation in placement, preprocessing, window length, and validation hampered comparison. GPS or GNSS data appeared mostly in the larger athlete-monitoring and workload literature, rather than as the predominant modality in the verified primary empirical sample.

Language data were employed in generative-AI research. Large language models generated workout prescriptions, sleep suggestions, instructional explanations, and decision-support text. Their outputs were evaluated using expert opinion rather than traditional predictive calibration. No included study demonstrated consistent, independently validated natural-language performance across model versions or sporting institutions. Simulation outputs were central to the digital-twin investigations. These models used athlete-specific physiological characteristics to predict pacing, energy system contribution, and future performance. Unlike purely data-driven models, they provided mechanistic interpretability; yet, their practical value was dependent on parameter quality, updating frequency, and field validation. True multimodal designs were less common than the broader rhetoric of multimodal sport AI implied. Many studies used multiple variables from the same data source or merged sensors at the feature level, but relatively few integrated video, physiological, workload, language, and contextual information into a single architecture. No included study demonstrated a mature foundation model capable of robust transfer across numerous sports, tasks, and modalities.

3.5 Validation and Methodological Quality

Validation procedures varied considerably and Verified primary empirical studies are listed in table 3. Internal train-test splits and cross-validation were the most commonly used approaches in predictive studies. Jauhiainen *et al.* (2022) employed repeated cross-validation, whereas Ma *et al.* (2025) tested ten algorithms on a single public dataset. Such processes evaluated internal consistency but did not establish transportability to other teams, seasons, institutions, or populations. Criterion validation was more prevalent in biomechanics and physical education assessments. Mundt *et al.* (2022), Johnson *et al.* (2019), Ishida *et al.* (2024), and Lahkar *et al.* (2022) compared AI-derived outputs to force platforms, marker-based motion capture, or reference biomechanical models, while Makaruk *et al.* (2025) compared automated motor-skill appraisal to expert assessment.

Table 3. Verified primary empirical studies

Study	Sample	Method	Validation category	Primary finding	Translation Maturity
Jauhiainen *et al.* (2022)	791 analysed; 60 ACL injuries	Four classifiers; linear SVM best	Repeated internal cross-validation	ROC-AUC about 0.63	Level 2: retrospective internal validation
Makaruk et al. (2025)	236 participants	AI scoring of five movement criteria	Concurrent expert-reference validity; no external site	ICC 0.96; weighted kappa 0.87	Level 3a: criterion validation
He *et al.* (2024)	10 university students	Interactive visual AI	Small uncontrolled pre-post field application	Performance and satisfaction outcomes	Level 4: field feasibility/evaluation
Ma *et al.* (2025)	800 public-dataset records	Ten ML models; SVM; SHAP	Single-dataset internal validation	Accuracy 95.6%; F1 95.7%; AUC 99.2%	Level 2: retrospective internal validation
Mundt *et al.* (2022)	14 participants; 1,474 movements	ANN for 3D GRF estimation	Laboratory criterion validation; no independent cohort	GRF waveform and peak/load metrics	Level 3a: criterion validation
Johnson *et al.* (2019)	30 participants	Pretrained CaffeNet CNN plus multivariate regression	Laboratory criterion comparison	Strongest mean correlation 0.8895	Level 3a: criterion validation
Ishida *et al.* (2024)	18 healthy young men	2D pose-estimation AI for VGRF	Concurrent force-plate criterion validation	Peak VGRF agreement/error	Level 3a: criterion validation
Lahkar *et al.* (2022)	3 elite boxers	Commercial markerless motion capture	Concurrent optoelectronic criterion validation	Joint-centre errors	Level 3a: criterion validation
Schortgen *et al.* (2026)	22 fighters; 18 rounds	Self-supervised monocular tracking	Authentic competition field evaluation	Position/trajectory tracking	Level 4: field feasibility/evaluation
Raju *et al.* (2026)	200 collegiate athletes	Random forest, XGBoost, and ANN; SHAP	Stratified five-fold internal cross-validation	RF accuracy 0.98; ROC-AUC 0.97	Level 2: retrospective internal validation

These investigations demonstrated that the systems measured the target concept, however criterion validity was not treated as external transportability. The samples remained small or context-specific, and no independent-site replication was demonstrated. Independent external validation was uncommon. Boillet *et al.* (2024) used field and physiological observations to anchor an individualized simulation, whereas Schortgen *et al.* (2026) tested tracking in authentic competition situations. However, neither represented complete independent replication by a different development team or sporting system. No injury prediction model had been prospectively evaluated across numerous independent organisations. Calibration

received little attention. Most classification studies reported accuracy, F1 score, area under the receiver operating characteristic curve, correlation, or agreement but did not assess whether predicted probabilities corresponded to observed event rates. This omission was particularly problematic for injury-risk models, as poorly calibrated probabilities could misdirect preventive interventions even when rank-order discrimination appeared high.

Explainability reporting was also inconsistent. Ma *et al.* (2025) and Raju *et al.* (2026) employed SHAP to identify influential variables. Mechanistic simulation studies were interpretable since their parameters reflected physiological processes. Biomechanical models also generated results that could be compared to known movement parameters. In contrast, several neural-network, computer-vision, and generative-AI experiments did not provide an adequate explanation for how specific outputs were formed. Few studies investigated subgroup performance or fairness. Some studies reported sex, age, or competition level, but few compared model performance across subgroups.

The significant male dominance found in earlier team-sport findings, the use of healthy young male samples in several biomechanical investigations, and the use of single-country educational and football datasets all limited representativeness. Unavailable code, proprietary systems, limited participant data, incomplete prompt reporting, and inconsistent preprocessing descriptions all hindered reproducibility. For instance, because the original consent did not allow publication, Mundt *et al.* (2022) were unable to make participant-level source data publicly available. Despite transparent validation against a reference standard, Lahkar *et al.* (2022) used a proprietary markerless method that limited algorithmic reproducibility.

3.6 Evidence Maturity and Implementation Status

The literature remained concentrated at the evidence-synthesis, conceptual, prototype, and retrospective-validation levels, according to the maturity assessment, (table 4). Instead of direct empirical validation, fourteen studies mainly provided contextual or governance synthesis. The translation maturity criteria are listed in supplementary table S5. Level 0 was assigned to conceptual and perspective contributions, and Levels 0–1 were assigned to surveys and early digital twin structures. The majority of technical prototypes and feasibility systems were classified at Level 1. Simulation studies and expert-evaluated generative-AI applications generally occupied Levels 1–2 since they demonstrated feasibility or internal review without independent field validation. Level 2 included retrospective predictive investigations. Injury prediction and single-dataset machine learning experiments fell within this category. Laboratory models validated against force platforms or marker-based systems occupied Level 3a since they provided criterion validity but lacked independent multi-site testing. Level 3a evidence was presented by Makaruk *et al.* (2025), and Level 3b evidence was presented by Boillet *et al.* (2024) using either field-anchored physiological validation or expert-reference validation. He *et al.* (2024) and Schortgen *et al.* (2026) reached Level 4 since their systems were tested in educational or competition-related field situations. However, neither demonstrated sustained decision improvement or downstream athlete results. No included primary study met Level 5, defined as demonstrated improvement in human decision-making, or Level 6, defined as measurable improvement in athlete, educational, clinical, or organisational outcomes attributable to the AI system. As a result, demonstrated translational impact was not included in the highest level of technical maturity.

3.7 Foundation Models and Generative AI

Compared to conventional machine learning and computer vision, evidence for foundation models and generative AI remained scarce. Exercise prescription, sleep and jet lag advice, educational assistance, interdisciplinary decision support, and the future structure of sports-science teams were all covered in the included studies. The primary benefit of generative AI was its ability to produce rapid, conversational, and apparently personalised output. Nevertheless, rather than examining long-term athlete use, the available studies evaluated response quality under controlled circumstances. Expert evaluation revealed reasonable recommendations combined with inconsistent safety, inadequate contextualization, and the requirement for professional verification. Robust evidence for sport-specific conversational coaching, automated longitudinal athlete reporting, validated synthetic data generation, retrieval-augmented sports medicine support, or autonomous training adaptation was not found in any of the eligible studies. There was inconsistent disclosure of prompt design, model version, retrieval sources, and hallucination testing.

Table 4. Methodological and translational appraisal

Evidence area	Dominant evaluation	Main strength	Main limitation	Highest maturity
Injury prediction	Internal cross-validation within single datasets	Supports nonlinear risk modelling and feature interpretation	Rare events, leakage risk, weak calibration, and little external validation	Level 2
Biomechanics and computer vision	Criterion comparison with force plates or marker-based systems	Clear reference standards and measurable error	Small samples, task specificity, and limited transportability	Level 3a
Physical education	Expert-reference assessment and small field applications	Direct educational use and interpretable task scoring	Single-site samples and limited long-term evaluation	Level 4 field feasibility
Wearable and multimodal analytics	Mostly review-level and retrospective evidence	Continuous field sensing and real-time potential	Drift, missingness, placement effects, proprietary processing, and weak interoperability	Level 1–2
Generative AI	Controlled expert assessment of model outputs	Rapid drafting and conversational support	Hallucination, version instability, privacy, and need for professional verification	Level 1–2
Athlete digital twins	Prototype, simulation, and field-anchored modelling	Individualised simulation and mechanistic interpretation	No recurrent bidirectional updating or closed-loop outcome evidence	Level 3b
Ethics and governance	Review and conceptual evidence	Identifies privacy, fairness, transparency, and accountability requirements	Few empirical studies operationalised athlete rights	Level 0 contextual

Therefore, rather than being considered established sports-science systems, foundation-model applications were viewed as new decision-support tools. According to Naughton *et al.* (2024), generative AI should be implemented as a sociotechnical shift and should be integrated into sports-science and sports-medicine teams. Their investigation reaffirmed the necessity of managing unintended consequences, human oversight, and clearly defined professional roles.

3.8 Wearable and Multimodal Athlete Analytics

Continuous, low-burden monitoring outside of the laboratory was supported by wearable analytics. Workload evaluation, movement recognition, core-temperature estimation, injury-risk monitoring, and return-to-play assistance were among the applications that were examined. Only a small percentage of the evidence, meanwhile, included real-time deployment in authentic training or competition settings. Sensor fusion can integrate physiological, inertial, positional, and environmental inputs through sensor fusion. However, whether fusion occurred at the raw-data, feature, or decision level varied among the reviewed experiments. Missing-data handling and synchronization processes were frequently not fully documented. Potential sources of error included motion artifacts, inconsistent positioning, device drift, and changes in environmental conditions. In some prototypes, such as the embedded artificial sport trainer reported by Fister Jr. *et al.* (2021), real-time inference and edge processing were technically possible. The results did not, however, demonstrate whether latency was low enough for time-

sensitive sports judgments or whether edge implementation maintained model accuracy across devices. Because sensors, software platforms, sampling frequencies, and data formats were rarely standardised, interoperability remained restricted. Reproducibility and transfer between teams or organizations were further limited by the use of proprietary hardware and algorithms. As a result, wearable and multimodal systems remained methodologically fragmented despite showing strong practical promise.

3.9 Athlete Digital Twins

The term "digital twin" was used more widely than the evidence justified. Conceptual repositories and coaching structures without automatic updating were classified as digital records or digital models. Digital shadows were defined as systems receiving one-way sensor updates without recurrent decision feedback. Fister Jr. *et al.* (2021) reported an operational computational prototype of an artificial sport trainer, but it did not meet the criteria for an interactive athlete twin. Because the physiology-based cyclist model developed by Boillet *et al.* (2024) represented an individual athlete and simulated future performance responses, it was classified as a predictive athlete twin. The related rowing simulation was likewise treated as a predictive digital model rather than a continuously updated twin. No included study demonstrated an interactive athlete twin with recurrent updating embedded in real training or recovery decisions, and no system met the closed-loop requirement of continuous bidirectional data exchange, prospective adaptation, and measurable feedback effects. Accordingly, athlete digital-twin functionality ranged from digital records to predictive twins but not to fully developed closed-loop systems. Figure 4 presents the six DT Functional Stages; these stages describe system functionality and are distinct from the separate Evidence Maturity Levels used to judge translational evidence. Table 5. Shows Emerging systems in sports AI.

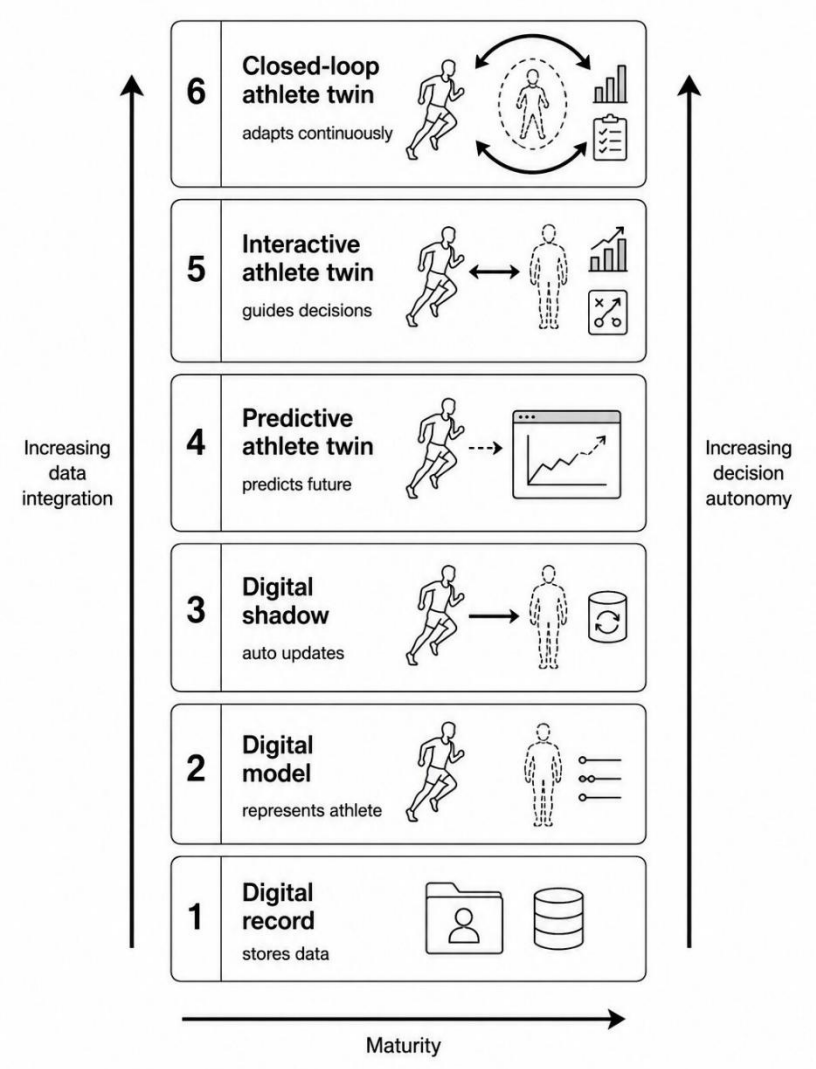

Figure 4. Athlete digital-twin functional-stage ladder (DT Functional Stages 1–6).

Table 5. Emerging systems in sports AI

System class	Typical inputs	Demonstrated function	Main unresolved issue	Current evidence stage
Wearable and multimodal analytics	Physiological, inertial, positional, environmental, and video data	Continuous monitoring, workload assessment, movement recognition, and heat-risk estimation	Cross-device validity, missing data, drift, synchronisation, and interoperability	Review evidence and Levels 1–3
Generative AI	Prompts, athlete questions, guidelines, and text records	Exercise, sleep, travel, education, and reporting support	Hallucination, source verification, model-version instability, privacy, and overreliance	Controlled expert evaluation at Levels 1–2
Athlete digital twins	Longitudinal athlete data, physiology, simulation parameters, and field observations	Individualised modelling and prediction of future performance states	Continuous updating, bidirectional decision integration, and outcome validation	Digital models to predictive twins; maximum Level 3
Governance systems	Consent, data-access, workflow, and accountability information	Policy guidance, role definition, and athlete-rights safeguards	Operational testing, contestability, and enforcement in real organisations	Conceptual and review evidence

3.10 SPORT-AI Translational Evidence Map

The translation pipeline is evaluated across the seven SPORT-AI dimensions depicted in Figure 5, progressing from technical feasibility and internal validation to field evaluation, decision impact, and measurable athlete or organizational outcomes. The seven dimensions of the SPORT-AI map showed uneven development. Studies that used specific reference standards such as force platforms, marker-based motion capture, expert evaluation, or observed field performance, showed the strongest scientific validity. Scientific validity was weaker in conceptually defined digital-twin systems, short uncontrolled interventions, and opaque public datasets. Reports of predictive performance were common, especially in the areas of motor-skill evaluation, movement analysis, and injury prediction. High internal performance did not consistently correspond to strong out-of-sample evidence. Prospective independent testing was uncommon, and there was little external validation. The competition-based boxing tracker and the physical education field research provided the most direct support for real-world utility. Even in these situations, the evidence focused more on feasibility than on improved decisions or results. The majority of other systems were assessed either in a controlled laboratory setting or in retrospect. Transparency differed depending on the modeling method. Both criterion-based biomechanical outputs and mechanistic physiological models were reasonably interpretable. In several prediction experiments, SHAP enhanced feature-level interpretation; nonetheless, code, preprocessing information, prompts, and proprietary system internals were frequently inaccessible. The SPORT-AI translational assessment at the study level is shown in Supplementary Table S4.

Only studies focused on governance addressed athlete rights in detail. Seldom were privacy, permission, fairness, data ownership, and contestability operationalized in empirical assessments. This dimension was further undermined by representation gaps, specifically the underrepresentation of women and the use of small, homogeneous samples. Narrowly defined tracking or assessment systems with readily available inputs and reference validation had the best implementation readiness. Due to a lack of external validation, workflow evaluation, interoperability, and governance protections, it was lower for injury prediction models, generative AI systems, multimodal wearables, and athlete digital twins. No aggregate SPORT-AI score was calculated.

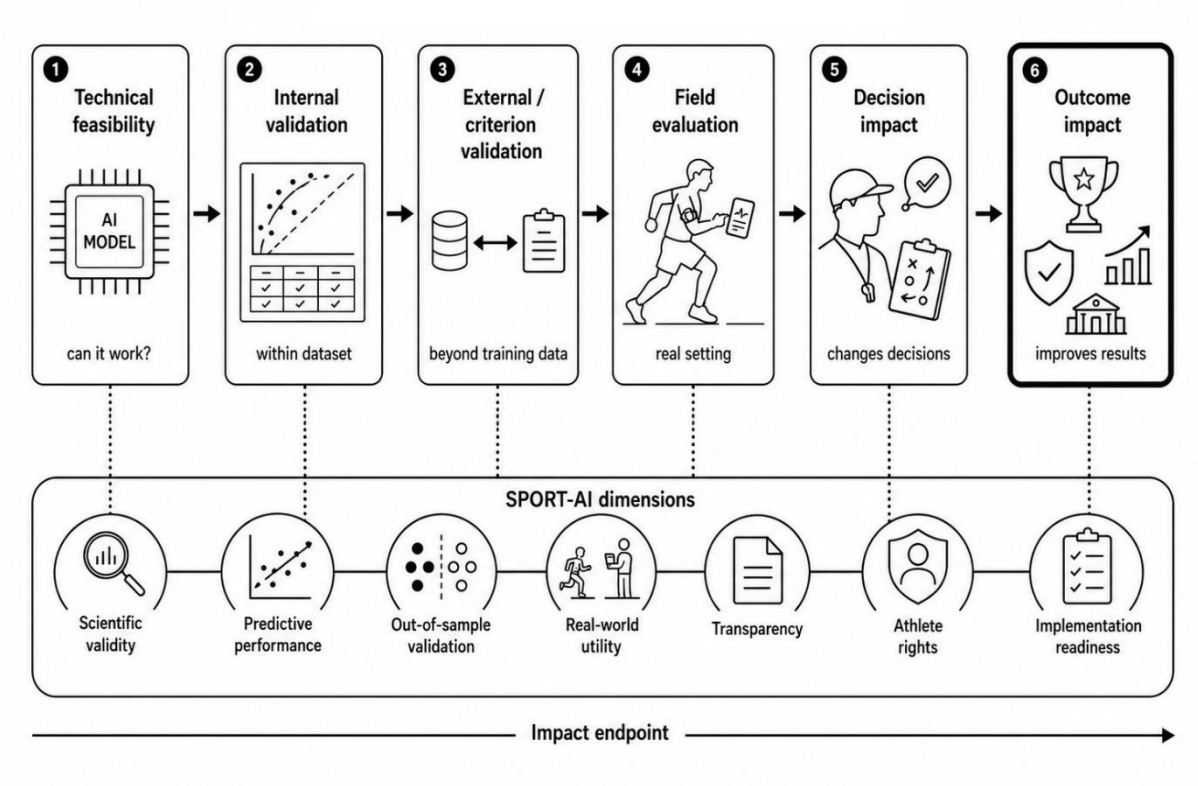

Figure 5. SPORT-AI translation pathway.

The distinct dimensions demonstrated that a system could perform well technically but fall short in terms of athlete rights, external validity, practical utility, or implementation readiness.

4. Discussion

4.1 Principal Findings

Performance analysis, injury-risk estimation, biomechanics, wearable monitoring, physical education, generative decision support, and athlete-specific simulation are among the sports-science areas in which artificial intelligence is now applied. The evidence base is still inconsistent, though. Technical development has advanced faster than governance, real-world application, external validation, and demonstrated outcome impact. Narrowly specified tasks with quantifiable reference standards yielded the strongest empirical evidence among the 32 included studies. These included athlete tracking under authentic competition conditions, biomechanical estimation versus force platforms or marker-based motion capture, and motor-skill evaluation against expert appraisals. In these applications, an established physical or professional reference could be used to interpret the error, agreement, or correlation that resulted from evaluating the AI system against a clearly defined target (Johnson *et al.,* 2019; Mundt *et al.,* 2022; Lahkar *et al.,* 2022; Ishida *et al.,* 2024; Makaruk *et al.,* 2025).

In contrast, more general applications like multimodal decision assistance, athlete digital twins, generative workout recommendations, injury prediction, and conversational coaching were less developed. These systems frequently demonstrated conceptual feasibility or internal predictive performance, but they lacked prospective multi-site testing, comparison with routine professional practice, or proof that their application enhanced athlete outcomes. As a result, the maturity assessment showed a distinct discrepancy between translational readiness and computational competence. Another conclusion was that many publications described AI in ambitious terms while reporting limited operational functionality. This was especially noticeable in the literature on digital twins, where retrospective simulations, one-way data representations, and static athlete models were occasionally presented as full digital twins. Similarly, despite the lack of a thorough longitudinal or prospective evaluation, foundation-model and generative-AI applications were frequently mentioned as possible coaching or decision-support aids. Overall, the evidence indicates that sports AI has gained technical credibility in a few measurement and classification tasks, but it has not yet reached the same

level of trust in workflow integration, athlete protection, decision assistance, or outcome enhancement.

4.2 From Prediction Accuracy to Decision-Relevant Evidence

Predictive metrics are necessary, but they do not independently establish practical value. Accuracy, sensitivity, specificity, F1 score, area under the receiver operating characteristic curve, correlation, and mean error all reflect a specific element of model behaviour. None of these criteria alone can predict whether a system will improve a coaching, sports medicine, educational, or organizational decision. When results are unbalanced, accuracy might be deceptive. For instance, non-injury instances typically outweigh injury cases in injury prediction. Therefore, if a model successfully classifies the majority group but fails to identify athletes who will actually sustain an injury, it may achieve high overall accuracy. Although sensitivity and specificity offer more specific information, their practical interpretation is contingent upon the consequences of false-positive and false-negative judgments. While a high false-negative rate could lead to false reassurance, a high false-positive rate could result in needless training modifications or athlete restrictions.

This issue is demonstrated by the difference between Jauhiainen *et al.* (2022) and Ma *et al.* (2025). While Ma *et al.* demonstrated very high internal performance using a single public football dataset, Jauhiainen *et al.* reported limited discrimination for anterior cruciate ligament injury despite an intensive screening battery. Because the outcome may reflect dataset composition, class structure, predictor selection, or internal validation design, the higher metrics in the former study may not necessarily imply more practical utility. Such models are therefore vulnerable to performance inflation in the absence of independent testing, calibration, and temporal validation.

In biomechanics, mean error and correlation also need to be interpreted carefully. Systematic bias or clinically significant local errors may coexist with a robust correlation. For instance, Lahkar *et al.* (2022) discovered that certain joint centers had more accurate markerless estimates than others. Therefore, anatomically specific limits that are important for system assessment or injury-risk interpretation could be hidden by a global summary statistic. Predictive performance, calibration, uncertainty, transportability, workflow compatibility, comparison with expert judgment, and proof that the system changes decisions in the right direction must all be included in the decision relevant evidence. Few studies in this study went beyond technical validation to show decision impact, and none clearly demonstrated effects on Level 6 athlete, educational, clinical, or organizational outcomes.

4.3 Performance Analysis, Biomechanics, and Tactical Decision Support

AI has achieved its most credible technical performance in tasks where the target can be directly measured and the decision context is relatively narrow. Computer vision, pose estimation, neural networks, and markerless motion capture have all been utilized to predict joint locations, ground reaction forces, knee joint moments, and athlete trajectories. These systems have practical advantages since they can lessen dependency on laboratory-based equipment and allow analysis using standard video or portable sensors. Mundt *et al.* (2022) showed that two-dimensional pose information could support three-dimensional ground-reaction-force estimation, although performance was dependent on the pose-estimation system used. Johnson *et al.* (2019) demonstrated that deep learning can estimate knee-joint moments with a strong correlation to reference outputs. Ishida *et al.* (2024) also demonstrated that accessible two-dimensional video could support estimation of vertical ground reaction force during single-leg landing. These findings support the technological feasibility of AI-assisted biomechanical assessment.

However, the evidence remains dataset and task specific. Most research used small sample sizes, controlled movement procedures, limited camera configurations, or particular athletic activity. Model performance in the laboratory does not guaranteethe same level of accuracy when subjected to fatigue, contact, fast direction changes, equipment variations, or environmental interference. The joint-specific inaccuracy discovered by Lahkar *et al.* (2022) further demonstrates that system validity cannot be assumed uniformly across all body segments. The field evaluation was more limited, but infomrative. Schortgen *et al.* (2026) demonstrated that markerless tracking could function during real-world boxing competition despite frequent occlusion. This is a significant step forward from controlled laboratory testing to realistic sporting circumstances. Nonetheless, the study showed monitoring feasibility rather than tactical decision impact. It did not determine whether the outputs enhanced corner guidance, athlete placement, technical

preparation, or competitive results. Tactical decision support was less developed than measurement and tracking. The available research produced descriptive or predictive results, but few investigated how coaches interpreted these results, how quickly they might be employed, or whether they enhanced tactical decisions over existing video analysis. As a result, AI has stronger evidence as a tool for measuring and recognizing patterns than as an independent tactical decision system.

4.4 Injury Prediction, Prevention, and Rehabilitation

Injury prediction is an attractive but methodologically challenging application of sports AI. Injury episodes are relatively rare, multifactorial, and temporally dependent. Workload, previous injury, recovery, sleep, stress, competition exposure, surface, contact events, system, and unmeasured environmental factors all have the potential to influence them. These factors increase the risk of unstable modelling. The sample size is a significant concern. The relevant quantities are the total number of athletes and the number of injury events. Models built from a small number of events may appear accurate during internal testing but are highly sensitive to resampling or feature selection. Jauhiainen *et al.* (2022) included 60 anterior cruciate ligament injuries among 791 studied athletes, although the best model produced only modest discrimination. This finding highlights the difficulties of predicting injury from baseline screening data alone. Rare-event prediction also raises the possibility of class imbalance. If imbalance is not addressed properly, a model may favor the majority of non-injury classes. Oversampling, class weighting, and synthetic data generation can improve internal metrics, but they may produce artifacts if used prior to data partitioning. Leakage occurs when information from the test set influences feature selection, preprocessing, imputation, or resampling. The included research did not always provide enough information to exclude this possibility.

Temporal validation is especially crucial. Randomly dividing observations from the same season or dataset does not simulate the real deployment setting in which a model is applied to future athletes, teams, or seasons. A more robust design would train the model using earlier data and evaluate it prospectively on later observations. External validation should also assess performance in a different club, institution, league, or demographic group. Calibration was seldom investigated. This omission is significant since injury prevention decisions are sometimes based on estimated probabilities rather than ranking alone. A model that generates exaggerated risk estimates may cause coaches to restrict training needlessly, whereas underestimation may fail to detect vulnerable athletes.

Causal interpretation remains problematic. A variable identified as important by a prediction model is not necessarily a modifiable factor. Stress, sleep length, workload, or balance may be associated with injury risk without independently causing the incident. SHAP and other feature-importance approaches describe how a model uses a variable, but they do not prove that modifying that variable reduces injury. A good prediction score alone is insufficient for clinical and coaching applications. Systems should establish how risk outputs are communicated, how uncertainty is represented, who has the authority to override the recommendation, and what actions are taken in response to a high-risk categorization. None of the included research found that an injury-prediction model lowered actual injury incidence via a prospective decision pathway. The rehabilitation evidence was similarly limited. AI-supported monitoring and exercise prescription may assist return-to-play assessment, but the included literature did not establish that automated recommendations improved recovery time, reinjury rates, functional outcomes, or adherence. These systems should remain adjunctive to clinical and coaching judgement.

4.5 Wearable and Multimodal Systems

Wearable and multimodal systems support continuous and context-sensitive athlete monitoring. Physiological signals, workload data, inertial measurements, video, location information, and environmental variables may be combined to produce a more complete representation of athlete state than any single modality. Their main value is in real-time monitoring. Wearable systems can monitor changes in movement, fatigue, workload, cardiovascular response, temperature, and recovery outside the laboratory oratory. When paired with edge processing, some outputs can be produced locally with lower latency and less reliance on cloud connectivity. This could help with training adjustments, heat risk management, system feedback, and return-to-play monitoring. Practical reliability depends on data quality. Sensor placement can affect readings, particularly for inertial devices and physiological monitors. Small variations in attachment, orientation, skin contact, or garment movement can result in systematic error. Device drift can potentially

affect consistency over multiple sessions or competitive seasons. Missing data are common in field situations. Sensors can disconnect, lose synchronization, fail during contact, or generate distorted signals. The examined literature did not consistently describe how missing data were detected, imputed, or excluded. Real-time systems require explicit failure management since silent missingness might result in deceptive outputs.

Synchronisation is another issue with multimodal systems. Accurate temporal alignment is required when combining video, inertial, physiological, and location data. Even small timing differences might distort the relationship between movement and physiological reaction. The fusion method should thus be fully documented, including whether data were fused at the raw-signal, feature, or decision level. Proprietary platforms introduce additional limits. Commercial devices may employ undisclosed filtering, calibration, or feature extraction systems. This reduces reproducibility and makes cross-platform comparisons difficult. Data export may be limited by vendor-specific formats or licensing agreements. Interoperability remains weak. Systems with various sampling frequencies, coordinate systems, signal definitions, and file structures are difficult to integrate. Cross-device validation is thus required before outputs from different vendors are considered equal. The evidence examined supports the feasibility of wearable and multimodal monitoring, although there is currently no seamless transfer among devices, teams, or organisations.

4.6 Foundation Models and Generative AI in Sport

The ability of foundation models and generative AI to generate natural-language explanations, workout recommendations, summaries, reports, instructional materials, and conversational support has attracted interest. They may be helpful to athletes, coaches, medical professionals, physical education instructors, and sports organizations due to their versatility. Drafting training session reports, summarizing monitoring data, creating instructional explanations, creating travel-recovery recommendations, assisting with exercise planning, and facilitating communication among interdisciplinary personnel are some examples of potential uses. Retrieval augmentation may theoretically link a language model to athlete records, clinical recommendations, validated team protocols, or sport-specific knowledge bases. The available evidence did not support reliable autonomous use. Large language models could generate apparently useful outputs, as demonstrated by Dergaa *et al.* (2024) and Vitale *et al.* (2026), but expert evaluation was still required. Hallucination is the main risk: a system may produce a fluent recommendation that is hazardous, inadequate, or unjustified. Evaluation is further complicated by model-version changes. After an update, a prompt that was tested on one model version can yield different results. If the model, version, date, settings, and prompt are not reported, this restricts reproducibility. This degree of detail was not consistently provided by the included research.

In sports medicine and athlete health, unverifiable recommendations are particularly problematic. A language model might mix correct general information with unsuitable personal recommendations. The user might not be able to discern between generated inference and established instruction in the absence of source attribution or retrieval from verified content. Privacy also requires careful consideration. Injury history, physiological measurements, sleep habits, workload, mental-state markers, and performance flaws are examples of athlete data. There may be hazards associated with data retention, secondary use, or unauthorized access when such data is uploaded to a third-party generative-AI system. An further risk is over-reliance. Without independent confirmation, athletes or coaches with less expertise can accept confident results. Thus, generative AI ought to be positioned as supervised assistance. Interpretation, contextualization, and final decision-making should be left to human experts. Instead of autonomous coaching, therapeutic prescription, or closed-loop training adaption, the available evidence supports experimentation with controlled, low-risk applications.

4.7 Athlete Digital Twins: Current Reality Versus Conceptual Ambition

One of the most ambitious directions in sports AI is athlete digital twins. An individualized computational representation that is updated using athlete data, simulates potential future states, and participates in training or rehabilitation decisions through recurrent interaction would be necessary for a mature twin. The reviewed evidence did not show that this level of functionality has been achieved, according to the evaluated evidence. The majority of systems were better categorized as retrospective predictive systems, digital models, or digital shadows. A digital record stores athlete data, but it cannot predict future reactions. Athlete physiology or performance is

represented by a digital model, which may be updated manually or just during study analysis. A digital shadow does not use a bidirectional feedback process to influence decisions; instead, it receives one-way data from the athlete. A large portion of the present literature falls within these categories.

Fister Jr. *et al.* (2021) described an artificial sport trainer that showed technical feasibility on embedded hardware but failed to create a bidirectional, continuously updated athlete twin. Because Boillet *et al.* (2024) employed individual athlete parameters and replicated features of field and physiological performance, their physiology-based cyclist model was more sophisticated. As a result, it satisfied the requirements for a predictive athlete twin. Even so, this system failed to show closed-loop intervention, recurrent decision integration, continuous sensor updating, or adaptation across a training cycle. Without demonstrating operational twin behaviour in daily athlete management, the rowing simulations addressed performance scenarios in a comparable manner. The barriers to mature athlete twins are not solely computational. Longitudinal data, individual parameter estimation, multimodal synchronization, handling of missing data, uncertainty modeling, ongoing recalibration, and integration with coaching or clinical workflows are all necessary for reliable implementation. Additionally, the system must recognize when an injury, illness, change in equipment, altered training, or competition setting renders its depiction invalid. Validation is especially difficult. A twin should not be considered valid merely because it reproduces past performance. It should demonstrate prospective prediction under evolving circumstances and show that its recommendations improve decisions or results. This requirement was not met by any of the included studies. Therefore, the term should be used cautiously the term "digital twin." Digital modeling and predictive simulation are supported by current data, but fully interactive or closed-loop athlete twins are still a research goal rather than a established sports-science technology.

4.8 Explainability, Fairness, and Athlete Representation

The included studies' approaches to explainability varied. While biomechanical systems and mechanistic physiological models provided direct relationships between input parameters and outputs, several prediction studies employed SHAP to identify influential variables. However, stakeholder understanding and technical explainability are not the same thing. A variable's impact on a model prediction in relation to a reference distribution is shown by a SHAP value. It does not prove causation, clinical significance, or the appropriate course of action. In a similar vein, a data scientist might be able to comprehend a visual heatmap or feature ranking, but an athlete, coach, teacher, or physician would not. Outputs must be provided in a format suitable for the decision in order to provide stakeholders with a meaningful explanation. The forecast, uncertainty, relevant inputs, constraints, and safe range of interpretation should all be explained. Additionally, it should be clear to the user whether the system is making a prescriptive, predictive, or descriptive assertion.

There was little evaluation of fairness. Whether model performance varied by sex, age, disability, ethnicity, sport level, playing position, or competitive situation was not consistently assessed by the evidence. According to Claudino *et al.* (2019), team-sport AI research has a disproportionately high male representation. Small, homogeneous samples of young men or athletes in good health from a single nation or organization were employed in a number of empirical studies. These limitations may lead to systematic variations in the model's performance. Because female athletes have different injury patterns and physiological traits, an injury model that was primarily trained on male athletes might not transfer to female athletes. Athletes with disabilities may not respond similarly to a motor-assessment technique that has been validated in school children and university students. When evaluated in controlled shadow boxing, a markerless system can lose accuracy due to fatigue or contact. Competition level and playing position are also important. There are differences in training load, movement demands, and injury processes among developmental, amateur, and elite athletes as well as between positions. When these circumstances are disregarded, models may generate average forecasts that are unsuitable for particular subgroups. Future research should evaluate subgroup performance when sample size allows, properly state demographic and sport-specific composition, and refrain from claiming a model is universally valid when it has only been evaluated in a narrow population.

4.9 Privacy, Athlete Autonomy, and Governance

AI-powered sport systems have the potential to improve athlete monitoring. Movement, location, sleep, physiology, workload, injury history, psychological signs, and behavioral patterns can all be continuously

monitored. While this type of data can help with safety and performance, it also poses concerns when it is required or extends beyond the purpose originally agreed by the athlete. Consent needs to be specific, informed, and revisable. It is important for athletes to know what information is gathered, why it is gathered, who may access it, how long it is kept, and whether it will be utilized for contract negotiations, commercial development, research, or selection. In situations where there is a significant power disparity, especially in scholarship or elite environments, consent may not be entirely voluntary. One of the main governance issues is secondary use. Data initially collected for the purpose of preventing injuries may subsequently be utilized for commercial product development, contract decisions, performance ranking, or insurance evaluation.

Separate governance and, if necessary, fresh consent should be required for such uses. Intellectual property rights and ownership are still ambiguous. Different parties may have control over raw athlete data, processed features, model outputs, training recommendations, and proprietary algorithms. When it is both legally and practically feasible, athletes should have clear rights regarding access, correction, portability, and deletion. It is necessary to define human oversight precisely. Instead of automatically deferring to the system, a coach, therapist, or sports scientist should be in charge of analyzing model output. Errors, detrimental advice, data breaches, and unfair decisions should all be Accountability should be established in advance. Additionally, athletes need to be able to contest automated recommendations. An explanation and a suitable review procedure should be available to a player who has been categorized as high risk or low potential. Without contestability, AI might turn ambiguous forecasts into fixed organisational judgements. Data minimization, access control, audit trails, model documentation, role-based responsibilities, appeal mechanisms, and periodic reviews of the system's continued necessity and proportionality should all be included in governance arrangements.

4.10 Practical Implications

AI outputs should not be treated as objective truth, but rather as information that helps athletes make decisions. Athletes should retain the ability to question recommendations, understand how their data is utilized, and receive clear explanations. AI is most helpful to coaches when it is used in conjunction with domain expertise and observation. A system's testing in a similar sport, competitive level, age range, and operating environment should be considered by coaches. Training modification shouldn't be determined solely by high internal accuracy. Data quality, temporal validation, calibration, uncertainty, and reproducibility should be the top priorities for sports scientists. Models should be assessed using suitable benchmarks, such as standard techniques and professional opinion. AI-based risk or exercise recommendations necessitate independent clinical evaluation for medical experts and rehabilitation specialists. Without the proper professional supervision, systems shouldn't be used to diagnose injuries, approve return to play, or recommend rehabilitation. Automated evaluation may facilitate feedback and lessen effort for physical education teachers, but it should not replace teacher judgment. Systems ought to be transparent, age-appropriate, inclusive, and workable with the infrastructure at hand.

Model architecture, preprocessing, data partitioning, missing-data handling, calibration, subgroup performance, and version control should all be included in reports for developers. Sufficient validation documentation should be provided by proprietary systems so that users can evaluate their limitations. Instead of focusing solely on vendor-reported performance, procurement decisions for sports organizations should take governance, data ownership, interoperability, auditability, and long-term maintenance into account. Minimum requirements for athlete data protection, validation, accountability, and transparency must be met by regulating organizations. High-stakes AI applications must to be reviewed independently and reevaluated on a regular basis.

4.11 Research Agenda for 2026–2035

4.11.1 Methodological Standards

Research agenda for sports AI, 2026–2035 is listed in Table 6. Sports-AI research should use standardized reporting for population characteristics, outcome definition, data preprocessing, feature selection, train-test separation, missing data, class imbalance, calibration, uncertainty, and subgroup performance. Reporting should distinguish between internal and external validation, as well as temporal validation. Generative-AI investigations should include information on the model version, prompt, retrieval source, sampling parameters, and expert-review mechanism.

Table 6. Research agenda for sports AI, 2026–2035

Priority area	Current evidence gap	Required methodological advance	Responsible stakeholders	Expected translational outcome
External and temporal validation	Predominance of internal or same-dataset validation	Independent team, season, site, and prospective temporal validation	Researchers; clubs; federations; journals	More reliable transportability across sporting contexts
Calibration and uncertainty	Discrimination metrics often reported without calibration or uncertainty	Calibration curves, decision-curve analysis, prediction intervals, and uncertainty-aware reporting	Methodologists; developers; sports scientists	Safer decision support and clearer risk communication
Multimodal benchmarks	Heterogeneous sensors, labels, sampling rates, and fusion pipelines	Shared multimodal datasets, harmonised labels, reference protocols, and interoperability standards	Consortia; device manufacturers; governing bodies	Comparable and reproducible multimodal athlete analytics
Wearable field validation	Laboratory performance may not generalise to training and competition	Prospective field testing under motion, fatigue, missingness, drift, and environmental variation	Teams; clinicians; engineers; athletes	Robust real-time monitoring in authentic environments
Foundation-model evaluation	Limited sport-specific validation and risk of hallucinated recommendations	Model/version disclosure, prompt transparency, expert comparison, hallucination testing, and retrieval verification	Developers; coaches; clinicians; educators	Accountable and domain-grounded generative-AI use
Athlete digital-twin standards	Frequent use of the term digital twin for static records or one-way models	Functional classification, recurrent updating, individualised simulation, prospective testing, and closed-loop criteria	Researchers; federations; standards bodies	Clear distinction between conceptual models and operational athlete twins
Fairness and representation	Insufficient subgroup analysis across sex, age, disability, competition level, and geography	Representative sampling, subgroup calibration, fairness analysis, and transparent limitation reporting	Researchers; ethics boards; sport organisations	More equitable model performance and reduced exclusion
Privacy and athlete rights	Unclear ownership, consent, secondary use, and contestability of athlete data	Granular consent, data minimisation, governance protocols, audit trails, and appeal mechanisms	Athletes; unions; clubs; regulators	Rights-preserving and trustworthy AI deployment
Human–AI decision integration	Technical predictions often lack evidence of changed or improved decisions	Human-factors studies, workflow evaluation, prospective decision-impact trials, and role definition	Coaches; clinicians; analysts; athletes	Evidence that AI improves rather than merely automates decisions
Outcome-impact evaluation	Few studies demonstrate improved athlete, clinical, educational, or organisational outcomes	Pragmatic trials, implementation studies, cost-effectiveness analysis, and longitudinal follow-up	Teams; health systems; schools; funders	Demonstrated real-world value beyond model accuracy

4.11.2 Multimodal Benchmarks and Interoperability

The field requires common benchmark datasets that include video, inertial signals, physiological parameters, workload, contextual variables, and athlete outcomes. Datasets should contain standardised labels, metadata, sampling data, and reference standards. Interoperability standards are required for device formats, time synchronization, coordinate systems, and data transmission.

4.11.3 Prospective and External Validation

Models should be tested across multiple independent teams, universities, seasons, devices, and athlete populations. Temporal validation should come before deployment. Prospective research should compare AI-assisted decisions to standard practice to see if performance remains consistent under fatigue, missingness, environmental change, and changing competition conditions.

4.11.4 Athlete-Centred Governance

Future systems should incorporate privacy, consent, fairness, transparency, and contestability into the design process. Athlete representatives should take part in governance and system evaluation. Studies should report on subgroup performance and specifically address data ownership, secondary use, commercialization, and human oversight.

4.11.5 Implementation and Outcome Evaluation

The field should move beyond technical feasibility towards decision-impact and outcome studies. Research should determine whether AI changes coaching, clinical, or educational decisions and whether those changes improve performance, injury prevention, rehabilitation, learning, safety, or organisational efficiency. Cost-effectiveness, staff workload, user acceptance, and sustainability should also be evaluated.

4.12 Strengths and Limitations of the Review

This review integrates areas that are frequently examined separately, including performance analysis, injury prediction, biomechanics, wearables, physical education, generative AI, athlete digital twins, and governance. Its principal strength is the explicit separation of technical performance, validation, field feasibility, decision impact, outcome impact, and athlete-rights considerations. The review also distinguishes primary empirical evidence from contextual review and governance literature and applies a common translation framework across heterogeneous technologies. Several limitations should nevertheless be considered. First, the evidence base was highly heterogeneous in study design, population, sport, outcome, data modality, and AI method, which precluded meaningful meta-analysis. Second, the review included only English-language peer-reviewed publications and may therefore underrepresent relevant work reported in other languages or non-journal technical sources. Third, the protocol was not prospectively registered and no public protocol was released; although the review procedures were defined before final screening and synthesis, prospective registration would have strengthened procedural transparency. Fourth, because the corpus contained heterogeneous designs, the review used structured design-specific appraisal domains rather than one formal risk-of-bias instrument across all studies. These appraisals should therefore be interpreted as methodological assessments, not as interchangeable validated risk-of-bias scores. Fifth, some included publications lacked sufficient information for direct study-level quantitative comparison, necessitating the fully verified empirical subset. Finally, rapidly evolving AI systems, especially large language models and generative systems, may change after publication, so conclusions regarding specific model capabilities should be interpreted in relation to the versions and evidence available during the review period.

5. Conclusions

Artificial intelligence is now an important component of sports science, but the available evidence does not yet justify broad claims of mature, routine, or autonomous implementation. The strongest evidence concerns narrowly defined measurement, tracking, and biomechanical tasks with explicit reference standards. In contrast, injury prediction, multimodal athlete monitoring, generative AI, and athlete digital twins remain limited by insufficient external and temporal validation, weak calibration, heterogeneous reporting, restricted reproducibility, and limited evidence of decision or outcome impact. Translation should therefore proceed from technical feasibility to independent validation, authentic field evaluation, demonstrated decision benefit, and ultimately measurable athlete, educational, clinical, or

organisational outcomes. Future work should also incorporate representative datasets, transparent uncertainty reporting, interoperable systems, subgroup and fairness evaluation, athlete-centred consent and data governance, and meaningful human oversight. The SPORT-AI Translation Framework and the separate Evidence Maturity Levels proposed in this review are intended to help researchers and practitioners distinguish promising technical systems from those that are genuinely ready for responsible implementation.

Athlete digital twins and generative AI require the same prudence. Although they remain susceptible to hallucinations, context loss, unreliable suggestions, and overreliance, large language models can generate reasonable exercise, educational, or recovery-related advice. Instead of being fully interactive or closed-loop athlete twins, the majority of systems referred to as digital twins were better categorized as digital records, digital models, digital shadows, or predictive simulations. The next phase of sports AI should focus on translation rather than technical novelty alone. External validation and prospective testing of models, calibrated and uncertainty-aware outputs, inclusive and representative datasets, transparent reporting, multimodal interoperability, and comparison with standard practice and expert judgment are among the priority topics. Human oversight must remain explicit, particularly in high-stakes decisions regarding injury, rehabilitation, selection, and athlete welfare. Athlete rights should be considered fundamental design criteria. These rights include privacy, consent, data ownership, fairness, explanation, and the ability to content automated recommendations. Reproducible gains in coaching, sports medicine, education, athlete, and organizational outcomes should define progress rather than higher nominal model performance.

References

Adlou, B., Wilburn, C., Weimar, W. (2025). Motion capture technologies for athletic performance enhancement and injury risk assessment: A review for multi-sport organizations. *Sensors*, 25(14), 4384. [DOI] [PubMed]

Boillet, A., Messonnier, L.A., Cohen, C. (2024). Individualized physiology-based digital twin model for sports performance prediction: a reinterpretation of the Margaria–Morton model. *Scientific Reports*, 14(1), 5470. [DOI]

Boillet, A., Messonnier, L.A., Cohen, C. (2025). Energetic parameters of rowing performance: will the distance change in the 2028 Los Angeles Olympic Games matter?. *Medicine & Science in Sports & Exercise*, 57(10), 2294-2306. [DOI]

Chidambaram, S., Maheswaran, Y., Patel, K., Sounderajah, V., Hashimoto, D.A., Seastedt, K.P., McGregor, A.H., Markar, S.R., Darzi, A. (2022). Using artificial intelligence-enhanced sensing and wearable technology in sports medicine and performance optimisation. *Sensors*, 22(18), 6920. [DOI]

Claudino, J.G., Capanema, D.D.O., de Souza, T.V., Serrão, J.C., Machado Pereira, A.C., Nassis, G.P. (2019). Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Medicine - Open*, 5(1), 28. [DOI] [PubMed]

Dergaa, I., Saad, H.B., El Omri, A., Glenn, J., Clark, C., Washif, J., Guelmami, N., Hammouda, O., Al-Horani, R., Reynoso-Sánchez, L., Romdhani, M., Paineiras-Domingos, L.L., Vancini, R.L., Taheri, M., Mataruna-Dos-Santos, L.J., Trabelsi, K., Chtourou, H., Zghibi M., Eken, Ö., Swed, S., Aissa, M.B., Shawki, H.H., El-Seedi, H.R., Mujika, I., Seiler, S., Zmijewski, P., Pyne, D.B., Knechtle, B., Asif, I.M., Drezner, J.A., Sandbakk, Ø, Chamari, K. (2024). Using artificial intelligence for exercise prescription in personalised health promotion: A critical evaluation of OpenAI's GPT-4 model. *Biology of Sport*, 41(2), 221–241. [DOI] [PubMed]

Dolson, C.M., Harlow, E.R., Phelan, D.M., Gabbett, T.J., Gaal, B., McMellen, C., Geletka, B.J., Calcei, J.G., Voos, J.E., Seshadri, D.R. (2022). Wearable sensor-based estimation of core body temperature during exercise: A systematic review. *Sensors*, 22(19), 7639. [DOI] [PubMed]

Fister Jr, I., Salcedo-Sanz, S., Iglesias, A., Fister, D., Gálvez, A., Fister, I. (2021). New perspectives in the development of the artificial sport trainer. *Applied Sciences*, 11(23), 11452. [DOI]

Gámez Díaz, R., Yu, Q., Ding, Y., Laamarti, F., El Saddik, A. (2020). Digital twin coaching for physical activities: A survey. *Sensors*, 20(20), 5936. [DOI] [PubMed]

Gao, Y. (2025). The role of artificial intelligence in enhancing sports education and public health in higher education: Innovations in teaching

models, evaluation systems, and personalized training. *Frontiers in Public Health,* 13, 1554911. [DOI] [PubMed]

He, Q., Chen, H., Mo, X. (2024). Practical application of interactive AI technology based on visual analysis in the professional system of physical education in universities. *Heliyon*, 10(3), e24627. [DOI] [PubMed]

Ishida, T., Ino, T., Yamakawa, Y., Wada, N., Koshino, Y., Samukawa, M., Kasahara, S. Tohyama, H., (2024). Estimation of vertical ground reaction force during single-leg landing using two-dimensional video images and pose estimation artificial intelligence. *Physical Therapy Research*, 27(1), 35-41. [DOI] [PubMed]

Jauhiainen, S., Kauppi, J.-P., Krosshaug, T., Bahr, R., Bartsch, J., Äyrämö, S. (2022). Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes. *American Journal of Sports Medicine*, 50(11), 2917–2924. [DOI] [PubMed]

Johnson, W. R., Mian, A., Robinson, M. A., Verheul, J., Lloyd, D. G., & Alderson, J. A. (2020). Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning. *IEEE Transactions on Biomedical Engineering*, *68*(1), 289-297. [DOI] [PubMed]

Kim, J.H., Kim, J., Kang, H., Youn, B.Y. (2025). Ethical implications of artificial intelligence in sport: A systematic scoping review. *Journal of Sport and Health Science*, 14, 101047. [DOI] [PubMed]

Lahkar, B. K., Muller, A., Dumas, R., Reveret, L., Robert, T. (2022). Accuracy of a markerless motion capture system in estimating upper extremity kinematics during boxing. *Frontiers in Sports and Active Living*, 4, 939980. [DOI] [PubMed]

Ma, J., Liu, S., Pei, Y. (2025). SHAP-based interpretable machine learning for injury risk prediction in university football players: a multi-dimensional data analysis approach: J. Ma et al. *Scientific Reports*, 15(1), 40252. [DOI] [PubMed]

Makaruk, H., Porter, J.M., Webster, E.K., Makaruk, B., Tomaszewski, P., Nogal, M., Gawłowski, D., Sobański, Ł., Molik, B., Sadowski, J. (2025). Artificial intelligence-enhanced assessment of fundamental motor skills: validity and reliability of the FUS test for jumping rope performance. *Frontiers in Artificial Intelligence*, 8, 1611534. [DOI] [PubMed]

Mundt, M., Born, Z., Goldacre, M., & Alderson, J. (2022). Estimating ground reaction forces from two-dimensional pose data: a biomechanics-based comparison of alphapose, blazepose, and openpose. *Sensors*, *23*(1), 78. [DOI] [PubMed]

Musat, C.L., Mereuta, C., Nechita, A., Tutunaru, D., Voipan, A.E., Voipan, D., Mereuta, E., Gurau, T.V., Gurău, G., Nechita, L.C. (2024). Diagnostic applications of AI in sports: a comprehensive review of injury risk prediction methods. *Diagnostics*, 14(22), 2516. [DOI]

Naughton, M., Salmon, P.M., Compton, H.R., McLean, S. (2024). Challenges and opportunities of artificial intelligence implementation within sports science and sports medicine teams. *Frontiers in Sports and Active Living*, 6, 1332427. [DOI] [PubMed]

Peters, M.D.J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A.C., Khalil, H. (2020). Chapter 11: Scoping reviews. *In JBI Manual for Evidence Synthesis*. [DOI]

Raju, S., Singamaneni, K.K., Hooi, L.B., Rani, K.U., Chandrika, B. (2026). Machine learning framework for predicting athletic injuries and optimising performance. *BMC Sports Science, Medicine and Rehabilitation*, 18(1), 107. [DOI] [PubMed]

Reis, F.J.J., Alaiti, R.K., Vallio, C.S., Hespanhol, L. (2024). Artificial intelligence and machine learning approaches in sports: Concepts, applications, challenges, and future perspectives. *Brazilian Journal of Physical Therapy*, 28(3), 101083. [DOI] [PubMed]

Schortgen, A., Goyallon, T., Saulière, G., Muller, A., Revéret, L. (2026). Monocular markerless position tracking of elite amateur boxing fighters in real combat situation. *Journal of Sports Sciences*, 44(10), 1319–1333. [DOI] [PubMed]

Seshadri, D.R., Thom, M.L., Harlow, E.R., Gabbett, T.J., Geletka, B.J., Hsu, J.J., Drummond, C.K., Phelan, D.M., Voos, J.E. (2021). Wearable technology and analytics as a complementary toolkit to optimize workload and to reduce injury burden. *Frontiers in Sports and Active Living*, 2, 630576. [DOI] [PubMed]

Souaifi, M., Dhahbi, W., Jebabli, N., Ceylan, H.I., Boujabli, M., Muntean, R.I., Dergaa, I. (2025). Artificial intelligence in sports biomechanics: A scoping review on wearable technology, motion analysis, and injury prevention. *Bioengineering*, 12(8), 887. [DOI] [PubMed]

Van Eetvelde, H., Mendonça, L.D., Ley, C., Seil, R., Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. *Journal of Experimental Orthopaedics*, 8(1), 27. [DOI] [PubMed]

Vitale, J., McCall, A., Cina, A., Athlete Travel, Sleep Interest Group, van Rensburg, D. C.J., Halson, S. (2026). "Can we trust them?" An expert evaluation of large language models to provide sleep and jet lag recommendations for athletes. *Sports Medicine*, 56(1), 257–270. [DOI] [PubMed]

Wang, X., Wang, X., (2024). Artificial intelligence applications and teacher preparation in physical education. *Frontiers in Public Health*, 12, 1484848. [DOI] [PubMed]

Zhong, Q., Jiang, J., Bai, W., Yin, Z., Liao, Z., Zhong, X. (2025). Application of digital-intelligent technologies in physical education: a systematic review. *Frontiers in public health*, 13, 1626603. [DOI] [PubMed]

Funding Source

This study received no external funding.

Conflicts of Interest

The author declares no conflict of interest regarding the publication of this article.

Does this article pass screening for similarity?

Yes

About the License