AI

MDPI

Review

From Prediction to Decision Support: A Critical Review and Six-Layer Framework for Responsible Artificial Intelligence in Sport Science

Stefan Alecu * and Gheorghe Adrian Onea

Department of Physical Education and Special Motricity, Faculty of Physical Education and Mountain Sports, Transilvania University of Brașov, 500068 Brașov, Romania; onea.gheorghe@unitbv.ro
* Correspondence: alecu.stefan@unitbv.ro

Abstract

Artificial intelligence is increasingly used in sport science to classify movement, analyse tactics, estimate readiness, forecast performance and injury risk, and support rehabilitation. Predictive performance alone, however, does not establish practical usefulness, safety, or responsible use. Through purposive source selection and critical synthesis, this review integrates sport-specific evidence with guidance on prediction modelling, human–AI interaction, and AI risk management. It highlights five recurring gaps: limited data and labels; leakage-prone or temporally inappropriate validation; weak reporting of calibration and uncertainty; unclear translation from prediction to action; and insufficient attention to athlete rights and post-deployment monitoring. We propose a six-layer framework covering data and context, prediction, decision translation, individualisation, human oversight, and deployment monitoring. For each applicable layer, the framework specifies evidence to document, appraisal questions, and common failure modes. It distinguishes model development from operational decision support and links technical evaluation with workflow utility, fairness, privacy, accountability, contestability, and lifecycle governance. The framework is conceptual, not a validated scoring or certification instrument. Sport-science AI should be evaluated as a socio-technical intervention rather than an isolated algorithm. Future studies should prioritise athlete-level and temporal validation, calibrated probabilities, external and prospective evaluation, prespecified action pathways, subgroup performance, documented override procedures, and ongoing monitoring.

Keywords: artificial intelligence; machine learning; sport science; athlete monitoring; injury prediction; decision support; model validation; explainable AI; responsible AI; human oversight

Academic Editor: Olufemi A. Omitaomu

Received: 2 August 2026
Revised: 27 August 2026
Accepted: 31 August 2026
Published: 2 September 2026

1. Introduction

Sport science has become a high-frequency data environment. Wearable sensors, global and local positioning systems, inertial measurement units, video, force platforms, physiological monitoring, medical records, training logs, and athlete-reported outcomes can describe external load, internal load, movement quality, tactical behaviour, recovery, symptoms, readiness, and exposure. Artificial intelligence (AI) can integrate these sources and perform classification, prediction, representation learning, optimisation, and interactive decision support [1–3].

The application landscape is broad. Machine-learning models are used to estimate injury risk, predict performance, identify tactical patterns, classify technique, recognise

activities, quantify biomechanics, and personalise training. Recent reviews show rapid growth across biomechanics, sports medicine, health monitoring, coaching, talent identification, and team-sport analytics [1–4]. TacticAI provides a prominent example of a system that combines prediction and generative recommendations with expert evaluation in professional football [5].

Growth in model development has not been matched by evidence of deployment readiness. A 2022 review of 30 sport injury-prediction studies found no external validation, limited calibration reporting, and predominantly high or unclear risk of bias [4]. Later reviews still describe marked heterogeneity, class imbalance, small samples, and insufficient external validation [6,7]. These limitations matter because a model can report high accuracy or discrimination while producing poorly calibrated probabilities, unstable recommendations, or unsafe performance when the team, season, device, or athlete population changes.

Sport data create specific methodological problems. Repeated observations from the same athlete are dependent. Randomly assigning observations from one athlete to both training and test sets can leak identity and inflate performance. Elite datasets may contain thousands of time windows but only a small number of athletes. Injury labels are rare, delayed, and sensitive to definitions. Tactical and biomechanical outcomes depend on game state, role, opponent, surface, fatigue, and measurement systems. These features require grouped, temporally ordered, and externally evaluated designs rather than generic random splits [4,8–10].

Sport-science AI also affects rights and relationships. Athlete data may reveal health, location, sleep, mood, behaviour, recovery, and employment-related information. Consent can be constrained when teams, federations, sponsors, universities, or technology providers control access to participation or selection. A 2025 scoping review grouped the main ethical concerns into fairness and bias, transparency and explainability, privacy and data ethics, and accountability [11]. These issues are not separate from model validity. An unrepresentative dataset or unequal error distribution is both a technical weakness and an ethical risk.

This review takes an AI-centred perspective. It does not rank algorithms by headline accuracy. It asks what evidence is required before a model can influence a real training, performance, medical, rehabilitation, selection, or return-to-sport decision. The main contribution is a six-layer framework connecting data quality, predictive validity, decision translation, individualisation, human oversight, and deployment monitoring. The framework is intended for researchers, coaches, sport scientists, clinicians, athletes, clubs, federations, reviewers, and technology providers.

Objectives

The review has four objectives:

1. Describe the main AI tasks, data modalities, and decision contexts in sport science.
2. Explain why discrimination or prediction error alone is insufficient for decision support.
3. Define a six-layer framework with proposed minimum evidence requirements and structured appraisal questions.
4. Provide a practical research, reporting, procurement, and governance agenda for responsible deployment.

2. Materials and Methods: Critical Review Approach

2.1. Review Design and Scope

This article is a critical review with a conceptual framework. It is not presented as a systematic review, scoping review, or meta-analysis. The purpose is to integrate

sport-specific research with cross-domain methodological and governance guidance and to develop an operational framework for responsible AI decision support. This design supports critical comparison of evidence standards across heterogeneous applications without making unsupported claims of exhaustive study identification.

The scope covers athlete- and team-facing uses of AI in training, performance, readiness, injury risk, rehabilitation, recovery, biomechanics, tactics, technique, talent identification, selection, and return-to-sport decisions. Fan engagement, betting, marketing, broadcasting, venue management, and purely administrative applications are outside the primary scope unless they provide a directly transferable methodological lesson.

2.2. Source Selection and Critical Synthesis

Sources were selected purposively across eight complementary evidence streams, detailed in Supplementary File S1. These streams covered the sport-science AI landscape; prediction-model limitations in sport; representative decision-support applications; ethics and responsible AI in sport; reporting and appraisal guidance; explainability and human–AI interaction; documentation and governance; and deployment, dataset shift, calibration, and decision utility. Priority was given to peer-reviewed sport-science reviews, representative empirical systems, authoritative methodological guidance, and sources that clarified a failure point in the evidence chain from data to action.

Targeted searches were conducted in Web of Science, PubMed, and Scopus and were supplemented by backward and forward citation searching. The searches were last updated on 6 August 2026. The concept blocks, prioritisation criteria, and exclusion principles are reported in Supplementary File S1. Because the aim was critical integration rather than exhaustive retrieval, no PRISMA flow, pooled estimates, or prevalence claims are presented.

Sources were retained when they contributed directly to athlete-facing use, methodological authority, recent guidance, practical transferability, or clarification of a six-layer failure point. Duplicate or superseded guidance, sources without sufficient methodological detail, and opinion claims without a clear evidential or conceptual contribution were not retained. The synthesis was narrative and framework-oriented; no formal study-level risk-of-bias assessment or certainty-of-evidence grading was undertaken. The thematic source map in Supplementary File S1 records the primary role of each reference.

The retained evidence base comprised 32 cited sources: 10 sport-specific reviews or empirical examples [1–8,11,12], six reporting, appraisal, or AI trial-guidance sources [9,10,13–16], eight explainability, human–AI, documentation, and governance sources [17–24], seven methodological sources on deployment, dataset shift, calibration, and decision utility [25–31], and one sport-medicine decision-centred framework [32]. To reduce selection bias inherent in purposive selection, the synthesis triangulated targeted searches across three databases with backward and forward citation searching, predefined evidence streams, and explicit prioritisation and exclusion principles; Supplementary File S1 records each reference's primary role. These safeguards improve transparency but do not substitute for a formal risk-of-bias assessment or make the corpus exhaustive.

Sport-specific claims were anchored in reviews of AI applications, injury prediction, ethical implications, and recent decision-support examples [1–10]. Methodological recommendations were informed by Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus AI (TRIPOD+AI) and Prediction model Risk Of Bias ASsessment Tool plus AI (PROBAST+AI), AI trial-reporting guidance, calibration and decision-utility literature, and deployment frameworks [9,10,13–16,25–31]. Explainability, documentation, and human-oversight recommendations drew on established work on SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic

Explanations (LIME), interpretable modelling, model cards, datasheets, and human–AI interaction [17–24].

2.3. Framework Development

The framework was developed through a staged conceptual synthesis. First, the intended endpoint was defined as accountable decision support rather than prediction alone. Second, the evidence chain was traced from the data-generating context through model evaluation, action selection, longitudinal adaptation, human review, and post-deployment learning. Third, recurring requirements from sport-specific reviews and cross-domain guidance were mapped to the point in the chain at which they could materially weaken the validity, safety, or accountability of an action.

A requirement was assigned to a layer when it addressed a distinct operational question and when failure at that point could compromise the action even if earlier model metrics were strong. Overlapping requirements were then consolidated into six layers, each expressed as a primary question, evidence to document, and common failure modes. Supplementary File S1 records the operational definitions and evidence-to-framework mapping.

In the evidence-to-framework mapping, sport-specific evidence most directly informed Data and Context and Prediction; decision-utility guidance informed Decision Translation; longitudinal monitoring and adaptation requirements informed Individualisation; human–AI guidance informed Human Oversight; and governance and deployment guidance informed Deployment Monitoring. Requirements were consolidated only when they answered the same operational question, while cross-cutting issues were retained at the earliest layer where failure could alter the downstream action.

The resulting framework is sequential but iterative: monitoring outcomes can trigger changes to data governance, model calibration, thresholds, interfaces, or workflow. The layers are not a validated scoring scale, and the checklist is intended for structured appraisal and documentation rather than summation into a total score. The level of evidence expected should be proportionate to intended use, reversibility, athlete choice, and potential harm.

2.4. Relationship to Existing Guidance

TRIPOD+AI and PROBAST+AI address reporting and appraisal of prediction-model studies [9,10], whereas Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI), Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (SPIRIT-AI), and Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) extend evaluation toward workflow use and intervention trials [14–16]. Model cards and datasheets structure documentation [22,23], and the National Institute of Standards and Technology (NIST) AI Risk Management Framework addresses lifecycle risk governance [24]. A recent sports medicine review similarly frames AI as physician-led decision support [32]. The proposed framework integrates these complementary requirements across athlete- and team-facing contexts, adding explicit attention to repeated athlete data, team and season transfer, action thresholds, longitudinal individualisation, athlete contestability, and operational monitoring. It does not replace the source instruments; Supplementary File S1 provides a crosswalk.

Direct transfer from clinical AI guidance has limits in sport. Decisions may occur within hours rather than along clinical care pathways, may be led by coaches or analysts without medical authority, and may balance performance, availability, and reversible risk rather than patient outcomes alone. Sport implementations should therefore adapt clinical governance to role-specific accountability, rapid escalation, and proportionate risk tolerance

while retaining core requirements for validation, documentation, privacy, human review, and monitoring.

2.5. Generative AI Assistance and Human Verification

OpenAI ChatGPT was used to support conceptual organisation, language refinement, literature triage, and the preparation of initial table and figure wording. It was not used to create or alter empirical data. Final source retention, interpretation, and wording remained under author review; the authors checked the cited sources against the original publications, revised the interpretations, verified the final text, and accept full responsibility for the manuscript.

3. Application Landscape of AI in Sport Science

AI applications in sport science can be organised by the decision they are intended to support, as can be seen in Table 1. This distinction is more useful than grouping studies only by algorithm because the same model family can serve very different purposes and risk levels. A classifier used to label movement segments has a different consequence profile from a model used to restrict training or delay return to competition.

Table 1. Athlete-facing AI applications, typical decision targets, and dominant failure risks.

Application Area	Typical Data	Common AI Task	Primary User and Decision	Main Risk
Monitoring and readiness	GPS/LPS, IMU, heart rate, sleep, wellness, training logs	Forecasting, anomaly detection, state classification	Head coach, strength and conditioning coach, or sport scientist adjusts load, recovery, or session content	Identity leakage, unstable baselines, missing context, alert fatigue
Performance and tactics	Tracking data, event data, video, opposition analysis	Prediction, pattern discovery, retrieval, generative recommendation	Head coach or performance analyst selects tactics, line-ups, or technical priorities	Proxy outcomes, context shift, overconfidence, weak human evaluation
Injury risk and health	Exposure, workload, history, symptoms, screening, clinical data	Risk estimation, classification, time-to-event modelling	Team physician or physiotherapist changes screening or referral; performance staff may adjust exposure within agreed medical constraints	Rare outcomes, label ambiguity, miscalibration, harmful thresholds
Rehabilitation and return to sport	Strength, range of motion, biomechanics, pain, function, imaging	Trajectory modelling, classification, recommendation	Team physician or physiotherapist leads rehabilitation and return-to-sport decisions; coach or performance staff manage graded performance exposure	Causal overreach, small samples, unequal errors, automation bias
Biomechanics and technique	Video, pose, force, IMU, motion capture	Pose estimation, segmentation, classification, feedback	Coach, analyst, or athlete changes technique	Measurement error, occlusion, domain shift, misleading explanations

Table 1. *Cont.*

Application Area	Typical Data	Common AI Task	Primary User and Decision	Main Risk
Talent and selection	Testing, match data, anthropometrics, scouting reports	Ranking, clustering, potential prediction	Coach, academy, or federation selects or develops athletes	Historical bias, self-fulfilling labels, opacity, limited contestability
Interactive and generative AI	Reports, text, video summaries, knowledge bases	Question answering, summarisation, scenario generation	Practitioner explores evidence or options	Hallucination, provenance gaps, confidential-data exposure

Across these application areas, multimodal AI may combine synchronised or asynchronous streams such as global or local positioning data, inertial measurement units, video, physiological signals, medical records, training logs, and athlete reports. Integration can occur at the input, representation, or decision level, but it adds risks from time misalignment, missing modalities, unequal sampling rates, modality-specific bias, and device or context shift. Within the six-layer framework, Layer 1 should document provenance, synchronisation, missing-modality handling, and permissions; Layer 2 should test incremental value and performance when a modality is absent or degraded; and Layers 3–6 should define which modalities are required for an action and monitor modality-specific failures after deployment.

3.1. Athlete Monitoring, Readiness, and Training Adaptation

Wearables and daily monitoring systems generate longitudinal data that may support individualisation. Models may classify readiness, detect unusual responses, forecast fatigue, or estimate performance. The central challenge is that readiness is not directly observed. Labels may derive from coach judgement, self-report, subsequent performance, or physiological thresholds, each representing a different construct and time horizon [1,2,12]. A model can therefore learn the measurement process rather than the intended underlying state.

Personalisation also requires more than including athlete identity as a predictor. A credible adaptive system should define an initial baseline, update rules, minimum data requirements, uncertainty during cold-start periods, and safeguards against feedback loops. When coaches change training in response to a model, future data are affected by the intervention. Evaluation must therefore distinguish forecasting accuracy from the effect of using the forecast [21,25–28].

3.2. Performance, Tactics, and Coaching Support

Tactical AI combines event, tracking, and video data to retrieve patterns, predict outcomes, and generate alternatives. TacticAI illustrates movement toward decision support because it connects a defined football situation to recommendations and includes expert evaluation [8]. Even so, qualitative preference for a recommendation does not replace prospective evidence that using the system improves decisions or outcomes across teams, competitions, and coaching styles.

In performance prediction, outcome selection is decisive. Models that forecast match results or aggregate ratings may be technically interesting but weakly connected to controllable actions. More useful targets are often intermediate and decision-linked, such as the probability of a successful tactical pattern under specified constraints, the expected response to a training option, or the uncertainty around an athlete-specific performance range.

3.3. Injury Risk, Rehabilitation, and Return to Sport

Injury prediction is among the most studied and consequential applications. Reviews report promising discrimination but recurring weaknesses in sample size, class imbalance, calibration, validation, and reporting [3,4,6–8]. The 2022 review by Bullock et al. found no external validation among 30 included studies and calibration in only a small minority [4]. Two 2026 reviews continued to call for larger studies, external validation, appropriate calibration, and careful handling of imbalance [6,7].

The term injury prevention can also overstate what a prediction model does. A risk estimate may support prevention, but prevention requires an effective action, correct timing, adherence, and acceptable trade-offs. Restricting load may reduce one risk while reducing fitness or availability. Referral may consume resources and produce false alarms. Return-to-sport decisions can affect health, employment, selection, and athlete autonomy. These applications need explicit thresholds, action protocols, multidisciplinary review, and prospective impact evaluation.

3.4. Computer Vision, Biomechanics, and Technique Feedback

Computer vision can scale pose estimation, event detection, and technique analysis beyond laboratory motion capture. Yet output quality depends on camera placement, lighting, occlusion, clothing, body morphology, sport equipment, and movement speed. Error should be reported for the derived biomechanical variable and intended decision, not only for pixel or key-point accuracy. A small joint-centre error can propagate into materially larger error in an angle, moment, or technique classification [1,12,13,29].

Real-time feedback creates an additional risk. Immediate advice may appear useful, but the model must distinguish natural variability from correctable error and avoid reinforcing a narrow technique template. Evaluation should include learning, transfer, safety, and practitioner interpretation, not merely agreement with a reference label [13,21,29].

3.5. Generative and Conversational Interfaces

Generative AI can summarise monitoring reports, answer questions, draft session options, and provide an interface to complex data. These functions may improve access, but they can introduce provenance, hallucination, confidentiality, and version-control risks. A language model should not silently convert uncertain model outputs into definitive prescriptions. For high-consequence uses, recommended safeguards include source attribution, uncertainty communication, constrained retrieval, audit logs, and named human approval [20–24,28].

Generative interfaces create failure modes beyond ordinary predictive error. They can fabricate tactical or rehabilitation rationales, expose proprietary athlete metrics when prompts or logs are sent to external services, or merge stale and current guidance without clear provenance. A retrieval-augmented generation (RAG) workflow can reduce, but not eliminate, these risks by retrieving only from approved, version-controlled sources, returning source evidence with the generated answer, refusing unsupported recommendations, and requiring named human verification for high-consequence advice. The retrieval layer should itself be evaluated for coverage, source freshness, citation faithfulness, access control, and behaviour when relevant evidence is absent [21–24,28].

Figure 1 shows the generic evidence chain, while Figure 2 operationalises the same process as the six-layer appraisal framework. Data maps to Layer 1; Model plus Evaluation to Layer 2; Decision to Layers 3 and 4; Human Use to Layer 5; and Lifecycle to Layer 6. Figure 1 therefore explains the flow from evidence to action, whereas Figure 2 specifies the appraisal questions applied at each stage.

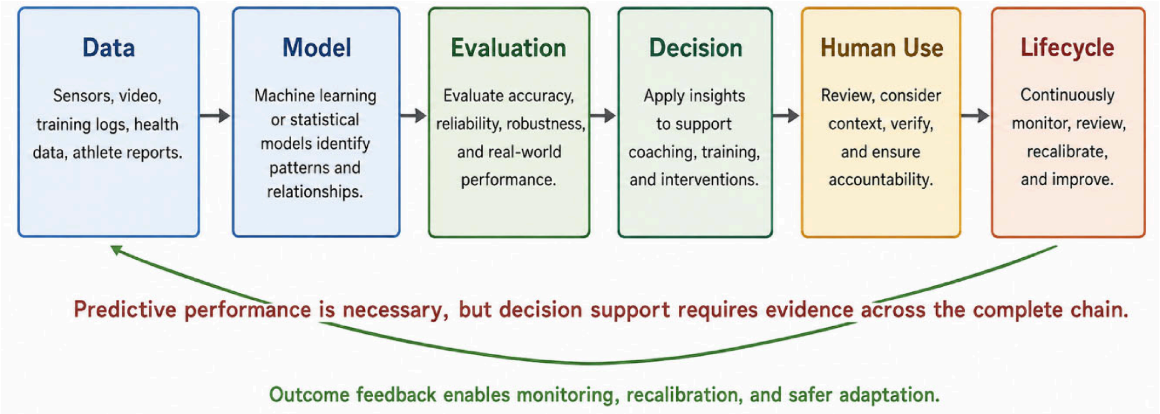

Figure 1. Evidence chain from sport data to responsible decision support. Each transition introduces requirements that cannot be inferred from predictive performance alone.

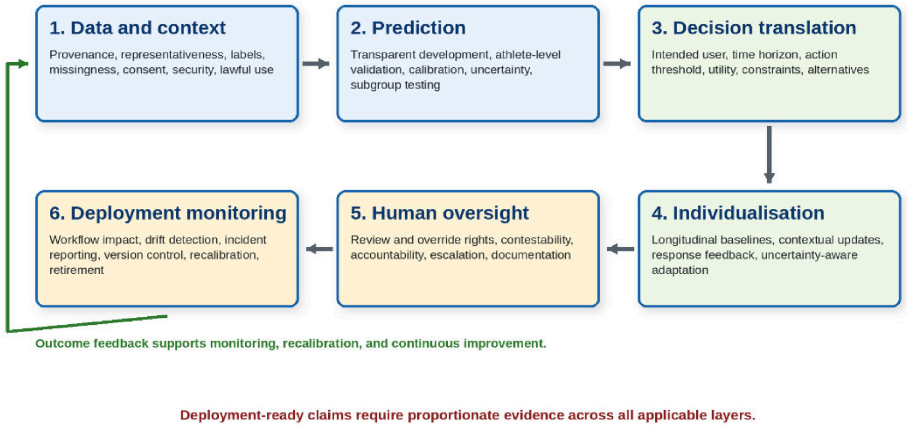

Figure 2. Six-layer framework for responsible AI decision support in sport science.

4. Why Predictive Performance Is Not Decision Value

4.1. Data Provenance, Labels, and Target Definition

The target variable defines what the system can learn. Injury, readiness, fatigue, technique quality, and performance can each be operationalised in multiple ways. A label based on missed participation differs from a medical diagnosis. A wellness score differs from physiological recovery. A coach rating may encode useful expertise and existing bias. Authors should report who created the label, when it was created, which information was available, how disagreement was resolved, and whether the label matches the intended decision.

Dataset documentation should cover the purpose, composition, collection process, exclusions, missingness, measurement changes, recommended use, and known limitations [23]. In sport, documentation should also record season, competition level, position or event, sex, age group, device and firmware, sampling frequency, pitch or surface, and relevant rule changes. Without this information, transportability cannot be judged.

Sport-specific outcome labels can be grouped as event labels, such as injury or movement events; state labels, such as readiness, fatigue, or recovery; performance labels, such as time, power, or tactical success; and decision or proxy labels, such as availability, selection, or coach rating. Injury labels in particular should distinguish time-loss, medical-attention, training-restriction, and match-unavailability definitions. These definitions are not interchangeable: they can change event frequency, prediction horizon, class balance, calibration, and the action triggered by a risk estimate. Layer 1 should therefore document the opera-

tional label and adjudication rule, Layer 2 should validate the model under that definition, and Layer 3 should link thresholds to the same decision-relevant outcome.

4.2. Leakage, Dependence, and Validation Design

Leakage occurs when information from the evaluation set influences model training or preprocessing. Sport datasets are especially vulnerable because the same athlete can contribute many correlated observations. Random row-level cross-validation may allow the model to recognise the athlete, team, or season rather than learn a generalisable relationship. Feature selection, imputation, scaling, and oversampling performed before data splitting can create additional leakage.

The validation design should mirror intended use. Generalisation to unseen athletes requires grouped splits by athlete; future prediction requires temporal evaluation; and deployment in a new team or site requires independent external evaluation. Hyperparameter tuning should be separated from final evaluation, often through nested validation. A single random train-test split rarely supports a strong claim in a small, dependent dataset [8–10,29].

4.3. Discrimination, Calibration, and Uncertainty

Discrimination measures ranking. Calibration measures whether predicted probabilities correspond to observed frequencies. A model can rank athletes well while systematically over- or underestimating risk. Because decisions often use thresholds, miscalibration can change who receives an intervention. Calibration plots, calibration-in-the-large, calibration slope, and observed-to-expected ratios should accompany discrimination metrics; proper scoring rules such as the Brier score, which measures mean squared probability error, can provide complementary information [9,30].

For operational purposes, uncertainty can be considered in at least three forms. Aleatoric uncertainty reflects irreducible variability. Epistemic uncertainty reflects limited knowledge or data. Distributional uncertainty arises when the deployment context differs from development data. A responsible interface should communicate when the model is uncertain, outside the evaluated distribution, or affected by missing or low-quality data. Confidence should not be represented through a single probability without documenting its calibration and evaluation context.

4.4. Class Imbalance and Metric Selection

Rare outcomes such as injury make accuracy misleading. A model can achieve high accuracy by predicting no injury. Sensitivity, specificity, positive and negative predictive values, precision-recall curves, and class-specific errors should be interpreted in relation to outcome prevalence and the intended action. Resampling and synthetic data techniques must be performed within the training process and evaluated for their effect on calibration. Prevalence changes across teams or seasons can alter predictive values even when sensitivity and specificity remain stable [3–7].

4.5. Decision Thresholds, Utility, and Consequences

A threshold converts a continuous prediction into an action. The threshold should reflect the relative consequences of false positives and false negatives, available resources, athlete preferences, and alternative actions. It should not be chosen only because it maximises a statistical index. Decision-curve analysis and related utility-based approaches can compare net benefit across plausible thresholds under explicit assumptions about the consequences of false-positive and false-negative decisions [31].

Sport decisions often involve multiple outcomes. Reducing injury risk may conflict with training stimulus, selection opportunity, or performance preparation. Tactical rec-

ommendations may improve expected scoring probability while increasing fatigue or defensive risk. The decision target should therefore specify benefits, harms, constraints, and time horizon. A prediction with no defined action pathway remains an analytic output, not a prescription.

4.6. Association, Causation, and Intervention

Predictive features are not automatically intervention targets. A variable may predict injury because it is a marker of prior exposure, role, or underlying condition, according to Table 2. Changing that variable may not change the outcome and may cause harm. Feature importance and SHAP values describe model behaviour; they do not establish causal effects. Claims that a model prevents injury or optimises training require evidence that the linked intervention changes outcomes under appropriate study designs.

Table 2. Proposed minimum technical evidence for sport-science AI evaluation.

Domain	Core Analysis	Stronger Evidence	Common Weak Practice
Data split	Grouped and temporally appropriate partitioning	External team, season, device, or site evaluation	Random row split with repeated athlete observations
Preprocessing	Fit only within training folds	Reproducible pipeline and locked evaluation set	Scaling, feature selection, or oversampling before splitting
Performance	Task-appropriate discrimination and error metrics	Confidence intervals and comparison with simple baselines	Accuracy alone for rare outcomes
Calibration	Calibration plot, intercept/slope, observed-to-expected ratio	External calibration assessment, recalibration, and threshold evaluation	Assuming a high AUC implies reliable probabilities
Uncertainty	Document missingness and out-of-distribution behaviour	Prediction intervals or uncertainty-aware abstention	Displaying one precise score without limits
Subgroups	Performance by relevant sex, age, level, role, and site	Prespecified fairness and error analysis	Reporting only pooled performance
Utility	Action threshold, alternatives, benefits, and harms	Decision analysis or prospective workflow evaluation	Selecting a threshold only by maximal statistical accuracy
Reproducibility	Code, hyperparameters, model version, data description	Independent replication or model card	Opaque commercial pipeline with no audit information

4.7. Review Findings Informing the Framework

Across the retained sport-specific literature, five recurring findings informed the framework: limited or weakly defined data and labels; leakage-prone or temporally inappropriate validation; limited calibration, uncertainty, and external validation; weak specification of how predictions change actions; and incomplete treatment of athlete rights, accountability, and post-deployment monitoring [1–10]. Sport-specific evidence was strongest for data, prediction, and applied sport examples, whereas the later decision, oversight, and lifecycle layers drew more heavily on cross-domain methodological and governance guidance because sport-specific deployment standards remain limited [11–31].

These statements summarise the evidence synthesis. The six-layer framework that follows is a normative synthesis that translates those findings, together with cross-domain guidance, into proposed evidence and appraisal requirements. Unless a statement is explicitly attributed to an empirical source, recommendations using terms such as should or require in Sections 5–9 should be read as proposed practice rather than prevalence estimates or validated standards.

5. Six-Layer Framework for Responsible AI Decision Support

The framework separates six questions that are often collapsed into one claim of AI performance. The layers are sequential but iterative. A material failure in any applicable layer can undermine the validity, safety, or accountability of the resulting decision. Monitoring outcomes feed back into data governance, model updating, threshold review, and workflow redesign.

The framework is intended to calibrate claims to evidence stage. A research prototype need not demonstrate every deployment layer, but claims of deployment-ready responsible decision support require adequate evidence across all applicable layers, proportionate to the intended use and potential consequences.

5.1. Layer 1: Data and Context

The first layer asks whether the data are fit for the intended sport context. Evidence to document includes population and setting definitions, provenance, measurement protocols, label quality, missingness, consent, security, retention, and lawful use. Representativeness should be judged against the intended users, not against an abstract population. A model developed in adult male professional football cannot be assumed to generalise to women athletes, youth athletes, para athletes, amateur participants, or athletes in other sports. Device and workflow changes should be treated as potential distribution shifts.

5.2. Layer 2: Prediction

The second layer asks whether the model estimates the target accurately and reliably. Evaluation should include transparent preprocessing, appropriate comparators, grouped and temporal validation, calibration, uncertainty, and subgroup performance. Independent external evaluation is needed to support claims beyond the development population, season, team, device, or setting. The model should be compared with simple baselines and existing practice. Complex models require evidence that their incremental value justifies the added opacity, maintenance, and data requirements [9,10,19].

5.3. Layer 3: Decision Translation

The third layer defines how the output could change an action. Developers and evaluators should specify the intended user, decision time, prediction horizon, threshold, candidate action, alternatives, contraindications, benefits, harms, and resource constraints. The output may support discussion, trigger additional assessment, prioritise review, or recommend a bounded action. The system should state what it does not decide.

5.4. Layer 4: Individualisation

The fourth layer applies when a system claims or performs athlete-level adaptation. It asks whether the system adapts to the athlete rather than merely attaching a personal label to a population model. Evidence should cover longitudinal baselines, minimum personal data, update frequency, response feedback, context changes, and uncertainty during sparse-data periods. Adaptive models require safeguards against self-reinforcing feedback and undetected performance drift. When no individualisation is intended, the layer may be marked not applicable with a written justification.

For cold-start or shifted-baseline cases, such as new recruits, youth entering elite squads, or return after a long injury layoff, a system should begin from a transparent population- or role-matched prior, update athlete-specific estimates conservatively, widen uncertainty, and abstain from individualised recommendations until prespecified minimum data are available. Bayesian or other adaptive updating rules and uncertainty-aware abstention thresholds should be specified in advance and validated for sparse-data periods.

5.5. Layer 5: Human Oversight

The fifth layer defines who reviews, accepts, rejects, overrides, or escalates an output and who remains accountable for the final action. Roles should be named, and users should receive information appropriate to their expertise. Athletes should have a route to ask questions, correct relevant data, and contest consequential outputs. Overrides should be possible, documented, and analysed. Human involvement is not sufficient by itself; oversight must be designed to reduce automation bias and responsibility gaps [21].

Human input should be distinguished across development and deployment. During development, domain experts may contribute label adjudication, error review, preference feedback, and evaluation criteria; during deployment, named practitioners review context and approve, override, or escalate outputs. Feedback used for later model updating should be logged, quality-checked, and separated from immediate decisions rather than treated as automatically correct training data. Accountability for consequential actions remains with defined human and organisational roles [21,24].

5.6. Layer 6: Deployment Monitoring

The sixth layer, as in Table 3, asks whether the system remains safe and useful after implementation. Monitoring should cover data quality, calibration, subgroup performance, workflow burden, overrides, incidents, drift, model and software versions, and unintended consequences. Teams should define triggers for review, recalibration, rollback, suspension, or retirement. Model updates, including automatically retrained versions, should not be deployed without change control, renewed evaluation, and documented approval [24–28].

Table 3. Operational appraisal structure for the six-layer framework.

Layer	Primary Appraisal Question	Proposed Evidence to Document	Typical Failure
1. Data and context	Are data representative, lawful, secure, and fit for use?	Provenance, labels, missingness, consent, governance, security	Selection bias, leakage, weak labels, privacy breach
2. Prediction	Does the model estimate the target reliably?	Transparent development, calibration, uncertainty, subgroup and external evaluation	Overfitting, miscalibration, hidden subgroup harm
3. Decision translation	How does the output change an action?	User, timing, threshold, utility, constraints, alternatives, contraindications	Prediction presented as prescription
4. Individualisation	Does the system adapt safely over time?	Baseline, update rules, feedback, cold-start safeguards	Static personalisation, unstable adaptation
5. Human oversight	Who is responsible for review and escalation?	Named roles, override, contestability, records, escalation	Automation bias, responsibility gap
6. Deployment monitoring	Does the system remain safe and useful?	Drift and incident monitoring, version control, recalibration, retirement	Silent decay, unmonitored harm, uncontrolled updates

To apply the framework, define the intended use and evidence stage, identify applicable layers, document supporting evidence and gaps, and set claim or deployment limits.

The layers are non-compensatory: evidence in one cannot offset a critical failure in another, so no composite score should be calculated.

For operational appraisal, practitioners should record each applicable layer as adequately supported, insufficiently supported, or not assessable, with a short written rationale and evidence reference rather than a numerical score. A critical unresolved gap should limit the claim or trigger escalation, additional evaluation, or non-deployment; stronger evidence in another layer does not compensate for it. This qualitative use remains proposed practice pending formal validation.

5.7. Illustrative Application: TacticAI

An illustrative application of the framework to TacticAI is provided in Supplementary File S1. Based on the published report, the system addresses a defined corner-kick task, evaluates predictive and generative components, and includes qualitative expert assessment [8]. The appraisal separates these strengths from evidence not established by that study, including independent organisational transfer, prospective workflow impact, longitudinal athlete adaptation, and lifecycle monitoring. The example shows how the framework can constrain claims without treating every absent layer as an equivalent deficiency in a research prototype.

6. Explainability and Human–AI Collaboration

6.1. Explain the Decision, Not Only the Model

Explainability should answer the question the user needs to act safely. A coach may need to know which recent data contributed to an alert, how uncertain the estimate is, whether the athlete is outside the training distribution, and which alternative actions are available. A clinician may need to see contraindications, missing inputs, and evidence supporting the threshold. A global feature-importance chart alone does not provide this information.

Post-hoc methods such as LIME and SHAP can support local or global inspection [17,18]. They should not be treated as proof that a complex model is correct, causal, or fair. Explanations can be unstable, sensitive to background data, and difficult to validate. For high-stakes decisions, interpretable models should be considered when they provide comparable performance and clearer operational behaviour [19,20].

6.2. Interface Design and Appropriate Reliance

Human–AI interaction should support appropriate reliance rather than maximum acceptance. The interface should communicate what the system can do, how well it performs, what inputs are missing, and when it is likely to fail. It should allow correction, dismissal, and efficient escalation. Feedback should be collected without converting every user response into unverified training data. The 18 human–AI interaction guidelines proposed by Amershi et al. provide a useful foundation for timing, explanation, correction, and error recovery [21].

For many high-consequence uses, a staged workflow may be appropriate. A model can prioritise cases, a practitioner can review contextual information, and a multidisciplinary team can make the final decision. The system should record the model output, relevant evidence, human judgement, final action, and outcome. These records support learning, accountability, and incident analysis.

6.3. Documentation for Researchers, Deployers, and Purchasers

Model cards can document intended uses, excluded uses, evaluation conditions, subgroup performance, limitations, and ethical considerations [22]. Datasheets can document

the training and evaluation datasets [23]. Where relevant to the intended use and consequence level, sport organisations should request both, together with version history, security information, data-flow diagrams, validation reports, and a change-management process. Procurement should not rely on vendor accuracy claims that cannot be mapped to the local population, outcome definition, and workflow.

7. Fairness, Privacy, Security, and Athlete Rights

7.1. Fairness as Performance and Process

Fairness begins with defining which groups and decisions matter. Relevant groups may include sex, age, disability, competition level, role, ethnicity, socioeconomic access, club, country, and sport. Pooled performance can hide substantial differences in false-positive rates, false-negative rates, calibration, or missingness. Subgroup analyses should be interpreted in relation to sample size and uncertainty. The appropriate fairness analysis depends on the decision and harm; equal error rates, equal calibration, and equal selection rates cannot always be achieved simultaneously.

Process fairness is equally important. Athletes should know when AI influences a consequential decision, which data were used, who can access the result, and how to challenge an error. Systems used for talent selection or contract-related decisions need stronger safeguards because historical labels may encode prior opportunity and bias. Fairness should be evaluated throughout the lifecycle, not only once during development [11,22–24].

7.2. Privacy, Consent, and Data Governance

Sport data can be deeply personal even when they are not medical records. Location, sleep, mood, menstrual-cycle information, injury history, genetic data, and behavioural patterns can affect selection, employment, and reputation. Data minimisation should ask whether each variable is necessary for the defined purpose. Consent should be specific enough to distinguish care, performance, research, commercial development, and secondary model training [11,22–24].

Governance should define ownership or control, permitted users, retention periods, export, deletion, de-identification, cross-border transfer, vendor access, and reuse after an athlete leaves the organisation. De-identification should not be assumed to eliminate privacy risk in rich longitudinal datasets. Federated or other privacy-preserving methods can reduce central data movement, but they do not remove the need for governance, security, representativeness, and local validation [11,23–25].

7.3. Security and System Integrity

AI systems depend on data pipelines, devices, software libraries, cloud services, and user permissions. Threats include unauthorised access, manipulated inputs, model theft, insecure integration, and silent vendor updates. Security testing should cover the complete system, not only the model. Organisations should maintain access logs, least-privilege permissions, backups, incident procedures, and a software bill of materials where appropriate [24,28].

7.4. Accountability and Contestability

Accountability should be allocated across the lifecycle and remain identifiable. Developers are responsible for design, testing, and documentation within their control; deploying organisations for local validation, integration, governance, and monitoring; practitioners for professional judgement within defined roles; and governance bodies for policy, audit, and escalation. Responsibilities can overlap and do not transfer simply because a human reviews the output. Athletes need an accessible route to correct relevant data and contest

consequential decisions. A statement that a human remains in the loop is insufficient unless authority, information, timing, and documentation are specified.

8. From Development to Deployment: An Evidence Ladder

The evidence ladder, which can be seen in Table 4, is a hierarchy of claims rather than a universal linear development pathway. In general, higher-stage claims require evidence from earlier stages and add new questions about transfer, action, workflow, impact, and lifecycle performance. Not every research prototype must reach deployment; the essential requirement is to label the evidence stage accurately and avoid implying operational readiness from internal model performance.

Table 4. Evidence ladder for claims about sport-science AI.

Stage	Core Question	Evidence Expected	Illustrative Claim Boundary
1. Technical feasibility	Can a model learn a useful signal?	Transparent data split, baseline comparison, internal validation	The method shows technical promise in the development dataset.
2. External performance	Does performance transfer?	Independent team, season, site, device, or population evaluation; calibration	The locked model demonstrated performance under the specified external conditions.
3. Decision utility	Would using the output improve decisions?	Defined thresholds, benefits and harms, decision analysis, user testing	The model showed estimated decision value at the evaluated thresholds and under the stated assumptions.
4. Prospective workflow evaluation	Can the system be integrated into practice with acceptable workflow and safety performance?	Prospective silent or assisted deployment, usability, overrides, workload, and incidents	The system was operationally feasible in the evaluated workflow; safety observations are limited to that setting and duration.
5. Impact evaluation	Does use improve athlete or organisational outcomes?	Controlled prospective evaluation with prespecified outcomes and harms	The AI-supported intervention improved the specified outcomes in the evaluated setting.
6. Lifecycle assurance	Does performance remain acceptable over time?	Drift, incidents, subgroup monitoring, change control, recalibration and retirement	The deployed system met the prespecified monitoring criteria during the documented period.

8.1. Prospective and Impact Evaluation

Prospective evaluation can begin with silent deployment, in which predictions are generated but not shown to decision makers. This can assess data availability, timing, calibration, and early signs of shift without changing care or training. Silent deployment still requires appropriate data governance, security, and incident procedures because operational athlete data are being processed. The next stage can evaluate usability, reliance, and decision changes with human review. High-consequence systems should ultimately be evaluated for impact on athlete outcomes, decision quality, workload, equity, and unintended harms. DECIDE-AI, SPIRIT-AI, and CONSORT-AI provide useful principles for early clinical evaluation and intervention trials, even when the sport context requires adaptation [14–16].

8.2. Monitoring and Change Control

Monitoring should be planned before deployment. Key indicators include input missingness, data-range violations, outcome prevalence, discrimination or error, calibration, subgroup performance, alert frequency, overrides, response time, incidents, and user complaints. A new device, software version, team, coaching policy, competition format, or outcome definition can require revalidation. Every released update should be versioned, tested against predefined acceptance criteria, approved, and reversible.

8.3. Procurement Checklist

- What exact decision and user does the system support, and which evidence stage has been reached?
- Which population, sport, competition level, devices, seasons, and outcomes were used for development and evaluation?
- Were athletes separated across training and evaluation data, and was temporal or external validation performed?
- Are probabilities calibrated, and are uncertainty and out-of-distribution cases communicated?
- How does performance vary across relevant subgroups and local conditions?
- What action follows an alert, and what evidence supports the threshold?
- Can users override, annotate, and contest outputs?
- Who owns and can reuse athlete data, and what happens when a contract ends?
- How are software and model changes documented and validated?
- Which monitoring, incident, recalibration, and retirement procedures are contractually supported?

9. Research and Reporting Agenda

Progress depends less on additional isolated accuracy demonstrations than on cumulative, decision-linked evidence. The priorities from Table 5 can improve scientific value and practical safety.

Table 5. Priority research and reporting advances for responsible sport-science AI.

Priority	Required Advance	Practical Example
Athlete-level and temporal validation	Use grouped, nested, and temporally ordered evaluation	Hold out complete athletes and future seasons
External and multi-site studies	Test across clubs, devices, competition levels, and underrepresented populations	Evaluate one locked model in several teams
Calibration and uncertainty	Report probability reliability and abstention behaviour	Show calibration curves and flag out-of-distribution cases
Decision-linked outcomes	Define the action, threshold, comparator, benefits, and harms	Compare model-assisted review with current practice
Prospective impact	Measure athlete, practitioner, workflow, and equity outcomes	Evaluate whether use changes injuries, availability, or decision time
Fairness and inclusion	Recruit and report sex, age, disability, level, and context	Prespecify subgroup evaluation and missing-data analysis
Reproducibility	Share code, model cards, dataset documentation, and full pipelines	Publish preprocessing and hyperparameters

Table 5. *Cont.*

Priority	Required Advance	Practical Example
Human factors	Study trust, reliance, override, usability, and accountability	Observe decisions with and without AI support
Lifecycle governance	Monitor drift, incidents, updates, and retirement	Maintain a local model registry and audit log

Proposed Minimum Reporting Set for Sport-Science AI

A proposed sport-specific reporting set can extend general AI guidance with context that directly affects validity. Reports should state:

- the intended user, decision, prediction horizon, and action pathway;
- sport, event or position, competition level, age, sex, disability or para-sport status where relevant, and recruitment setting;
- number of athletes, teams, seasons, events, and repeated observations, not only the total number of rows;
- device, firmware, sampling rate, preprocessing, missingness, and label definition;
- athlete, team, and time separation across development, tuning, and evaluation;
- baseline models, discrimination or error metrics, calibration, uncertainty, and confidence intervals;
- subgroup performance and the handling of class imbalance;
- intended and excluded uses, action thresholds, benefits, harms, and alternatives;
- explanation method, user interface, override and escalation process;
- code, model version, prespecified protocol or analysis plan where appropriate, data documentation, monitoring plan, and update policy.

10. Discussion

AI in sport science has progressed from isolated classification tasks toward systems that combine multimodal data, generate tactical alternatives, and support repeated decisions. The central challenge is now the evidence gap between model development and safe use, rather than a lack of algorithms. Sport-specific reviews repeatedly identify small samples, limited validation, weak calibration, and poor reporting [2–7]. Ethical reviews identify privacy, fairness, transparency, and accountability concerns [10]. A recent decision-centred sports medicine framework reaches a convergent conclusion: predictive outputs require structured integration into contextual human judgement [32]. Together, these findings point to an integrated problem: sport-science AI is often evaluated as a model when it should be evaluated as a socio-technical intervention.

The six-layer framework makes this distinction operational. Data and context determine what the system can learn. Prediction establishes whether the output is statistically credible. Decision translation defines what changes because of the output. Individualisation governs adaptation over time. Human oversight assigns responsibility and supports correction. Deployment monitoring tests whether performance and utility persist. The framework prevents a common category error in which a high-performing prediction is described as personalised, prescriptive, or preventive without evidence for the linked action.

The framework also clarifies the role of explainability. Explanations should not serve as decorative evidence of trustworthiness. They should help users understand the basis, uncertainty, limitations, and consequences of a decision. In many settings, a simpler interpretable model, a clear threshold policy, and a well-designed review process may be more useful than a complex model with post-hoc explanations [17–21].

Responsible AI practices can support cumulative evaluation and safer adoption by making evidence and limitations visible. Clear intended-use statements, data documentation, model cards, external evaluation, and monitoring make systems easier to compare, procure, improve, and retire [22–24]. They also reduce the risk of overstated claims and help practitioners identify the contexts in which a model is genuinely useful.

Sport organisations can apply the framework proportionately. A lower-consequence technique-classification tool may warrant proportionate governance and clear feedback testing. A system that influences medical referral, return to sport, selection, or employment requires stronger validation, fairness analysis, consent, contestability, and prospective impact evidence. The level of assurance should follow the potential harm, reversibility, and degree of athlete choice.

10.1. Implications for Researchers

Researchers should design evaluation around the intended deployment split and decision. They should report athletes, teams, seasons, and repeated observations separately; fit preprocessing within training data; evaluate calibration and uncertainty; compare against simple baselines; prespecify and justify action thresholds when claims extend to decision support; and avoid causal or preventive claims from predictive associations. Reporting should follow TRIPOD+AI where relevant, with PROBAST+AI informing quality and risk-of-bias assessment [9,10].

10.2. Implications for Practitioners and Organisations

Practitioners should ask what action the model supports, who remains accountable, and what happens when the output is wrong. Organisations should require local validation, data and model documentation, clear contracts for data reuse, user training, audit logs, and monitoring of drift, subgroup performance, incidents, and updates. The procurement questions in Section 8.3 provide a practical starting point.

10.3. Implications for Journals and Reviewers

Reviewers should challenge studies that use row-level random splits for repeated athlete data, report accuracy alone for rare outcomes, omit calibration, or describe association-based feature importance as causal evidence. Claims should match the evidence ladder. A feasibility study should not be required to demonstrate deployment monitoring when its claims remain appropriately bounded, but it should state which evidence stages remain untested. Journals can improve cumulative knowledge by requiring detailed data-split diagrams, model and dataset documentation, code where feasible, and explicit intended-use and excluded-use statements.

11. Limitations

This is a critical review based on purposive source selection rather than an exhaustive systematic search. It therefore does not estimate the prevalence of methods or reporting practices across all sport-science AI studies. Although the source-selection framework and thematic map improve transparency, the review did not use a closed study cohort, pooled synthesis, formal study-level risk-of-bias assessment, or certainty grading. The interpretive synthesis may therefore reflect author judgement, and alternative mappings are possible.

The framework was developed through author-led conceptual synthesis rather than a formal consensus process or stakeholder co-design. It integrates sport-specific and cross-domain guidance, and some recommendations originate from health AI because sport-specific deployment standards remain limited. Normative, privacy, and governance requirements may also require adaptation to applicable law, contracts, and institutional responsibilities across jurisdictions. The framework has not yet been prospectively tested

as an appraisal or procurement instrument, and the illustrative application is not a formal validation study. Future work should assess content validity with athletes, coaches, clinicians, data scientists, and governance stakeholders, then evaluate reliability, usability, inter-rater agreement, and the ability to predict implementation success across sports and organisational settings.

A first validation step should use a structured Delphi study or expert-panel process with athletes, coaches, clinicians, sport scientists, data scientists, and governance stakeholders to test content coverage, terminology, and decision rules. Subsequent studies should assess inter-rater reliability and usability in simulated appraisal or procurement cases before prospective testing in operational sport settings.

12. Conclusions

Artificial intelligence can support sport-science decisions, but model performance is only one part of the evidence required. Responsible decision support needs representative and governed data, appropriate validation, calibrated and uncertainty-aware predictions, a defined action pathway, safe individualisation where applicable, accountable human oversight, and ongoing deployment monitoring. The proposed six-layer framework connects these requirements and provides a practical structure for research, review, procurement, and governance. The field should move from asking whether an algorithm predicts well to asking whether the complete system improves decisions for athletes under real conditions, with measurable benefits, acceptable harms, and clear responsibility.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ai7090343/s1: Supplementary File S1: Source-selection framework, operational definitions, evidence-to-framework mapping, six-layer structured appraisal checklist, and illustrative application. The supplementary file cites and maps the sources included in the main manuscript [1–32].

Author Contributions: Conceptualization, S.A. and G.A.O.; methodology, S.A.; investigation, S.A. and G.A.O.; resources, S.A.; writing—original draft preparation, S.A.; writing—review and editing, S.A. and G.A.O.; visualization, S.A.; supervision, G.A.O.; project administration, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This article reviews published literature and presents a conceptual framework.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new empirical data were created or analysed.

Acknowledgments: OpenAI ChatGPT, GPT-5.6 Thinking, accessed 6 August 2026, was used as described in Section 2.5. The authors verified the sources, claims, interpretations, and final manuscript and take full responsibility for its content.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, D.; Keogh, J.W.L.; Ma, Y.; Tong, R.K.Y.; Khan, A.R.; Jennings, N.R. Artificial intelligence in sport: A narrative review of applications, challenges and future trends. *J. Sports Sci.* **2025**, 1–16. [CrossRef] [PubMed]
2. Claudino, J.G.; Capanema, D.O.; de Souza, T.V.; Serrão, J.C.; Pereira, A.C.M.; Nassis, G.P. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: A systematic review. *Sports Med. Open* **2019**, *5*, 28. [CrossRef] [PubMed]
3. Van Eetvelde, H.; Mendonça, L.D.; Ley, C.; Seil, R.; Tischer, T. Machine learning methods in sport injury prediction and prevention: A systematic review. *J. Exp. Orthop.* **2021**, *8*, 27. [CrossRef] [PubMed]

4. Bullock, G.S.; Mylott, J.; Hughes, T.; Nicholson, K.F.; Riley, R.D.; Collins, G.S. Just how confident can we be in predicting sports injuries? A systematic review of the methodological conduct and performance of existing musculoskeletal injury prediction models in sport. *Sports Med.* **2022**, *52*, 2469–2482. [CrossRef] [PubMed]
5. Wang, Z.; Veličković, P.; Hennes, D.; Tomašev, N.; Prince, L.; Kaisers, M.; Bachrach, Y.; Elie, R.; Wenliang, L.K.; Piccinini, F.; et al. TacticAI: An AI assistant for football tactics. *Nat. Commun.* **2024**, *15*, 1906. [CrossRef] [PubMed]
6. Lou, S.; Zhu, H.; He, Z.; Ye, B.; Sun, G.; Shen, C. The application of machine learning in the prediction of sports injuries: Systematic review and meta-analysis. *Sci. Rep.* **2026**. *Advance online publication*. [CrossRef] [PubMed]
7. Aslani, M.; Zarei, M.; Yaghoubitajani, Z.; Malekahmadi, A. Machine learning models for injury risk prediction in football players: A systematic review of predictors, performance, practical applications, and limitations. *J. Int. Med. Res.* **2026**, *54*, 3000605261469442. [CrossRef] [PubMed]
8. Rossi, A.; Pappalardo, L.; Cintia, P. A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer. *Sports* **2022**, *10*, 5. [CrossRef] [PubMed]
9. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
10. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Navarro, C.A.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef] [PubMed]
11. Kim, J.-H.; Kim, J.; Kang, H.; Youn, B.-Y. Ethical implications of artificial intelligence in sport: A systematic scoping review. *J. Sport Health Sci.* **2025**, *14*, 101047. [CrossRef] [PubMed]
12. Hammes, F.; Hagg, A.; Asteroth, A.; Link, D. Artificial intelligence in elite sports: A narrative review of success stories and challenges. *Front. Sports Act. Living* **2022**, *4*, 861466. [CrossRef] [PubMed]
13. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef] [PubMed]
14. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [CrossRef] [PubMed]
15. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef] [PubMed]
16. Rivera, S.C.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat. Med.* **2020**, *26*, 1351–1363. [CrossRef] [PubMed]
17. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2017; pp. 4765–4774.
18. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
19. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
20. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef] [PubMed]
21. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for human–AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2019; Paper 3; pp. 1–13. [CrossRef]
22. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 220–229. [CrossRef]
23. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Iii, H.D.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
24. Tabassi, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. [CrossRef]
25. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef] [PubMed]
26. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [CrossRef] [PubMed]

27. Finlayson, S.G.; Subbaswamy, A.; Singh, K.; Bowers, J.; Kupke, A.; Zittrain, J.; Kohane, I.S.; Saria, S. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **2021**, *385*, 283–286. [CrossRef] [PubMed]
28. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; Dennison, D. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*; Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2015; pp. 2503–2511.
29. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [CrossRef]
30. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [CrossRef] [PubMed]
31. Vickers, A.J.; Elkin, E.B. Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* **2006**, *26*, 565–574. [CrossRef] [PubMed]
32. Palermi, S.; Pucciatti, R.; Regnard, N.-E.; Guermazi, A.; Araujo, F.; Demeco, A.; Mekki, Y.; D'Antona, G.; Guarnera, A.; Cerciello, S.; et al. Artificial intelligence in sports medicine: A decision-centered framework for the future sports physician. *Diagnostics* **2026**, *16*, 1448. [CrossRef] [PubMed]