

Machine learning in sports and exercise analytics: applications to training load, technique and injury risk using performance and biomechanical data

Sally T. West^{1*}, Cathie A. Freeman¹

¹ Department of Physiotherapy, The University of Melbourne, Melbourne, Australia

ABSTRACT

Introduction and Objective. The rapid expansion of wearable sensing, biomechanical measurement and automated movement analysis has created athlete datasets that are high-frequency, multimodal and strongly individualized. This narrative review examines how machine-learning (ML) methods can be integrated with force-plate, motion-capture and wearable-sensor data to support training-load management, movement-technique analysis and injury-risk assessment. **Material and method.** The review was organized around three data modalities commonly used in applied sport science and the ML approaches paired with them, including classification, regression, ensemble learning and temporal models. Particular attention was given to the relationship between internal and external load, model validation, interpretability and translation to coaching or clinical decisions. **Results.** Across the reviewed applications, ML is most useful when it combines complementary physiological, mechanical and contextual features rather than relying on a single summary metric. Wearable data support longitudinal load profiling; force-time and kinematic data support technique and neuromuscular assessment; and multivariable models can identify patterns that warrant closer review. The main constraints are small athlete cohorts, heterogeneous acquisition procedures, overfitting, limited external validation and insufficiently transparent outputs. **Conclusions.** ML can extend conventional sports analytics by modelling complex interactions and temporal patterns, but predictive performance alone is insufficient. Robust applied use requires standardized data pipelines, appropriate validation, interpretable reporting and integration of model output with practitioner expertise.

Keywords: machine learning, sports analytics, training load, biomechanics, wearable sensors, injury risk

INTRODUCTION

Contemporary sport science increasingly depends on simultaneous measurements of external work, internal physiological response and movement mechanics. External load can be described using distance, speed, acceleration and other movement-derived variables, whereas internal load reflects the physiological and perceptual stress imposed by that work. In team sports, the same external task can produce different internal responses among athletes, making individualized monitoring essential for interpreting adaptation and fatigue.

Biomechanical systems add another layer of information. Force plates quantify ground-reaction forces and centre-of-pressure behaviour, three-dimensional motion capture describes joint and segment kinematics, and wearable devices continuously record heart-rate, inertial and positioning signals across laboratory, training and competition settings [1-4]. These technologies provide richer information than a single session summary, but they also create data that are noisy, high-dimensional and temporally dependent.

Conventional monitoring methods remain valuable. Heart-rate-based training impulse (TRIMP), session rating of perceived exertion and external-load summaries have established practical roles in football and other sports [5-10]. However, the relationship between workload, performance and injury is rarely determined by one variable alone. The training-injury prevention paradox and the workload-injury aetiology model both emphasize that load must be interpreted in relation to exposure, adaptation and athlete-specific characteristics [5,6]. Injury has likewise been conceptualized as a complex systems problem in which multiple interacting factors can produce different risk patterns [8].

Machine learning offers methods for modelling such nonlinear and multivariable relationships. In human movement biomechanics, supervised and unsupervised approaches have been used for classification,

regression, clustering and pattern recognition, while team-sport studies have applied artificial intelligence to performance prediction and injury-risk assessment [9,10]. The purpose of this review is therefore to synthesize how ML can be paired with three major sport-science data modalities - force plates, motion capture and wearable sensors - and to identify the methodological conditions required for useful translation to coaching and clinical practice.

MATERIAL AND METHOD

Review design. A narrative review format was used to integrate methodological and applied literature spanning sports biomechanics, athlete monitoring and machine learning. The manuscript was structured to mirror an applied sport-science workflow: data acquisition, feature construction, model selection, validation, interpretation and practitioner action. The evidence base comprised the references listed in this article, with emphasis on peer-reviewed work describing training-load monitoring, human movement biomechanics, injury modelling and AI applications in sport.

Data domains. Three data domains were considered. Force-plate data provide high-frequency kinetic signals such as peak force, impulse and rate of force development. Motion-capture systems provide joint angles, segment velocities and coordination patterns. Wearable sensors, including inertial units, heart-rate monitors and GPS systems, provide repeated measurements of internal and external load in ecologically valid environments [2-4,12-15].

Machine-learning approaches. The principal model families considered were supervised classification, regression, ensemble learning and temporal/sequence modelling. Classification is appropriate when the target is categorical, such as movement quality or a risk flag; regression is appropriate for continuous outcomes; tree-based ensembles are useful for heterogeneous tabular features; and temporal models can exploit ordered sequences in repeated physiological or movement data [9,10]. Unsupervised approaches can additionally identify latent athlete or movement subgroups when labelled outcomes are unavailable.

Evaluation principles. For applied use, model performance should be interpreted together with data provenance, cohort size, cross-validation strategy, calibration, external validation and model explainability. Particular caution is required when the number of candidate predictors is large relative to the number of athletes or injury events, because such settings can produce optimistic in-sample performance and unstable feature importance [9,10].

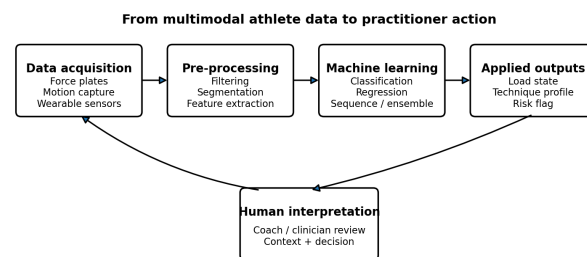


Figure 1. Conceptual pipeline from multimodal athlete data to practitioner action. The figure summarizes the review framework rather than presenting empirical results.

Table 1. Primary data modalities used in ML-supported sports analytics

Data source	Representative variables	Primary applied use
Force plates	Peak force; impulse; RFD; asymmetry; centre of pressure	Neuromuscular status; landing/jump profiling; technique assessment
Motion capture	Joint angles; segment velocity; coordination; pose trajectories	Technique classification; gait/skill analysis; return-to-sport comparison
Wearable sensors	Heart rate; HR zones; accelerations; distance; inertial features	Training-load monitoring; fatigue/recovery profiling; longitudinal risk modelling

Validation and reporting. A publishable sports-analytics model should report more than a single accuracy value. Classification studies benefit from sensitivity, specificity, area under the receiver-operating characteristic curve, precision-recall measures and calibration, while regression studies should report absolute and squared-error metrics alongside explained variance. Repeated or nested cross-validation can reduce optimism during model selection. The intended unit of generalization—new session, new athlete, new team or new competition period—should be explicitly stated because it determines how validation folds are constructed. Confidence intervals, resampling distributions or repeated-fold variability are useful for showing the stability of reported performance [9,10].

Model-development strategy. Model complexity should match the structure and size of the dataset. Penalized regression and generalized linear models provide useful baselines when the sample is small and the number of predictors is moderate. Tree-based ensembles can capture nonlinear interactions in heterogeneous tabular data and tolerate different feature scales. Temporal models may be justified when sequential information itself is a central predictor, such as when fatigue develops across repeated sessions. Deep learning becomes more defensible when there are sufficiently large signal, image or sequence datasets; it is not automatically superior in the small cohorts that are common in elite sport [9,10].

Preprocessing and leakage control. Preprocessing decisions can materially alter model performance. Scaling, normalization, missing-value imputation, outlier handling and feature selection should be performed within the training folds rather than on the full dataset. This is especially important in athlete-monitoring studies because repeated observations from the same athlete are correlated. If observations from one athlete appear simultaneously in the training and test partitions, the model can partially identify the athlete rather than learn a generalizable performance pattern. Grouped or athlete-level data splitting is therefore preferable when the intended application is prediction in new athletes, whereas temporally ordered splitting is more appropriate when the intended use is forecasting a future state within the same athlete [9,10].

Data preparation and feature construction. Raw sport data are rarely suitable for direct modelling. Force-platform signals may require filtering, event detection and temporal normalization; motion-capture data may require reconstruction of missing markers, coordinate transformation and calculation of joint- or segment-level variables; and wearable data require synchronization across heart-rate, accelerometry and GPS streams [1-4,9]. For tabular modelling, continuous signals are often summarized as domain-informed features such as peak force, impulse, rate of force development, contact time, acceleration counts, time in heart-rate zones or session-level load indices. Feature construction should remain physiologically interpretable whenever possible, because this facilitates error checking and later practitioner interpretation.

Search and selection framework. The source base was organized around three linked literatures: athlete-load monitoring, human movement biomechanics, and machine-learning applications in sport. Priority was given to papers that defined measurable input variables, described how the variables were processed, and linked model output to a concrete sport-science question. The review did not attempt to calculate a pooled effect because the included studies differ substantially in sport, sample structure, sensing technology and target outcome. Instead, evidence was synthesized by identifying recurring analytical tasks and by examining where machine learning adds information beyond established load-monitoring procedures. The training-load literature in the template article was used as an applied reference point because it demonstrates how internal-load metrics can be compared against physiological outcomes and interpreted within a sport-specific testing context [12-21].

RESULTS

Training-load optimization. Wearable-derived load metrics are especially suited to longitudinal modelling because they can combine repeated internal- and external-load information. Internal load reflects the athlete's response to the work performed, while external load describes the work itself [12-14]. Studies in football have shown meaningful associations among heart-rate-based methods, perceived exertion and fitness-related responses [16-21]. ML does not replace these constructs; rather, it can combine them with contextual features and individual histories to identify patterns that conventional univariate summaries may miss. In

practice, the most defensible output is an individualized probability, state classification or expected response that supplements - rather than replaces - established monitoring metrics.

Technique and biomechanical analysis. Force-time curves and kinematic trajectories contain information across an entire movement rather than at a single time point. Feature engineering can reduce these signals to interpretable descriptors, whereas sequence models can retain more of their temporal structure [2,9]. This supports applications such as classifying landing strategies, detecting departures from an athlete's baseline, comparing movement execution against reference patterns and quantifying changes during rehabilitation or skill acquisition. The advantage is objective, repeatable comparison; the limitation is that model output remains dependent on acquisition quality and the relevance of the chosen reference condition.

Injury-risk assessment. Predictive injury models combine workload, biomechanical and physiological variables to identify patterns associated with short-term risk [7,10,11]. Because injury is multifactorial, a risk flag should be interpreted as a triage signal rather than a diagnosis [8]. This distinction is particularly important when event counts are low or when model development occurs within one team. High discrimination in a development dataset does not necessarily imply transfer to another sport, club, competitive level or sensor ecosystem.

Model selection and interpretability. Different algorithms answer different questions. Linear or regularized models can provide transparent effect directions when the feature set is controlled. Tree-based ensembles can represent nonlinear interactions in heterogeneous tabular data and provide feature-importance summaries. Sequence models can use temporally ordered information, but their complexity increases data requirements and may reduce interpretability. Accordingly, model complexity should be justified by the structure of the data and the decision problem rather than by predictive performance alone [9,10].

Table 2. Alignment of common ML tasks with applied sport-science questions

ML task	Typical question	Key methodological risk
Classification	Which movement pattern, readiness state or risk category is most likely?	Class imbalance; optimistic validation; unstable thresholds
Regression	What continuous load, fatigue or performance value is expected?	Nonlinearity; calibration error; out-of-range extrapolation
Ensemble models	Which combination of heterogeneous features best separates outcomes?	Overfitting; misleading feature importance; site-specific patterns
Temporal models	How does the ordered sequence of physiological or movement data relate to outcome?	Large data requirement; temporal leakage; limited interpretability
Clustering	Are there natural athlete or movement subgroups without predefined labels?	Unstable clusters; uncertain practical meaning; sensitivity to scaling

Practical deployment. For field use, the analytical pipeline must operate on a timescale that matches coaching decisions. A model used to adjust tomorrow's training should produce stable outputs soon after the current session, while a return-to-sport model may tolerate slower processing if it integrates laboratory measurements. Automation is valuable when it reduces manual data handling, but automated pipelines require quality-control checks for missing sensor data, implausible values and device synchronization. The practical value of ML therefore depends as much on data engineering and workflow design as on the selected algorithm.

Injury-risk modelling as decision support. The literature supports a cautious interpretation of injury forecasting. Training and injury relationships are complex and cannot be reduced to a universal threshold [5,6,8]. ML can nevertheless organize multiple weak signals into a probabilistic estimate, particularly when workload history, prior injury, biomechanical asymmetry and physiological state are available together [10,11]. The most defensible use is as a decision-support layer that prompts additional assessment. A high-risk classification should lead to review of training exposure, clinical status and movement findings rather than automatic restriction of the athlete. This distinction is central to translating predictive performance into responsible sport practice.

Technique classification and movement quality. Biomechanical ML applications can be organized into three levels. The first level classifies a discrete movement or technical state, such as acceptable versus altered landing mechanics. The second predicts continuous biomechanical quantities, such as joint angles or loading variables. The third analyzes temporal structure across the entire movement, preserving information that would otherwise be lost by reducing the signal to a few peaks. For teaching and applied practice, these levels clarify that the “best” algorithm depends on the question: a transparent classifier may be sufficient for screening, whereas a sequence model may be required when the timing and coordination of movement are themselves the outcome.

Feature importance and actionable output. Model interpretation is most useful when it identifies modifiable factors. In load monitoring, an output that indicates elevated risk without showing which variables contributed is difficult to use. In contrast, a model that shows that recent high-speed running exposure, reduced recovery, altered jump impulse and an unusually high heart-rate response jointly contributed to the classification can support targeted follow-up. Feature-importance methods, partial dependence, local explanations and standardized coefficient estimates can all contribute, although none should be treated as causal proof. The interpretation should remain tied to the measurement process and to known physiological or biomechanical mechanisms [8-10].

Longitudinal athlete modelling. Repeated monitoring creates an opportunity to model within-athlete change rather than only between-athlete differences. A longitudinal system can use recent workload, baseline fitness, prior responses and current physiological markers to estimate whether a new observation deviates from the athlete's own normal range. This is often more useful in elite sport than comparison with a broad population norm. However, longitudinal models must account for non-stationarity: training phases, competition congestion, rehabilitation, sensor replacement and changes in tactical role can shift the distribution of predictors over time. Updating rules and drift monitoring should therefore be considered part of the applied model rather than an afterthought.

Relationship with established training-load metrics. ML should not be interpreted as making established metrics obsolete. The TRIMP framework illustrates why conventional indices remain valuable: heart-rate-derived measures provide a compact, physiologically interpretable representation of internal load, and their relationship with oxygen consumption can be evaluated directly [12,18-20]. In the source article used as the writing template, Edwards' TRIMP showed stronger correlations with oxygen consumption than several alternative TRIMP formulations under the reported treadmill conditions. This type of domain-grounded variable can be incorporated into an ML model as an input feature, a comparator, or a benchmark against which added predictive value is tested [18-21].

Multimodal data fusion. The reviewed literature indicates that the main analytical advantage of ML is the ability to combine variables that are normally interpreted separately. Internal-load measures describe the athlete response to work, external-load measures describe the work performed, and biomechanical variables characterize how the work is mechanically executed. These dimensions can be integrated at the session, drill, repetition or movement-cycle level. In practical terms, a model can distinguish two sessions with similar external distance but different cardiovascular strain, or two jumps with similar height but different force-time strategies. Such integration is particularly relevant when the goal is individualized monitoring rather than group-average description [3,4,12-14].

Integrating model output with practitioner judgment. The final decision should remain a synthesis of measured data, model output and domain expertise. A model can identify an unusual combination of signals, but practitioners can contribute contextual information that was not recorded in the dataset, such as travel, tactical changes, recent illness or a modified rehabilitation plan. This interaction is not a weakness of the model; it is the normal structure of applied decision making. The most useful systems therefore make it easy for practitioners to inspect the evidence behind an alert and to record the action taken, creating a feedback loop that can later be evaluated and refined.

Reproducibility and documentation. Reproducibility requires more than publishing the name of an algorithm. The analysis should specify software, package versions, random seeds, preprocessing rules, feature definitions, model-tuning ranges and validation procedures. Where raw athlete data cannot be shared,

authors can still provide pseudocode, synthetic examples, analysis scripts or a detailed data dictionary. Version control is especially important in long-term monitoring because model updates can otherwise make results from different periods incomparable. A documented pipeline also reduces dependence on a single analyst and facilitates integration with the routine workflow of coaches, physiotherapists and sport scientists.

Visualization and communication. The final analytical product should communicate uncertainty and context. Dashboards can show current values, recent trends, individual baselines and model estimates, but visual simplicity is important. Red-amber-green traffic-light systems are easy to interpret yet may conceal uncertainty if thresholds are treated as absolute. A better approach is to pair concise status indicators with the underlying change-from-baseline data and an uncertainty interval. In research publications, figures should likewise show distributions, individual trajectories or calibration where possible rather than relying only on summary accuracy. Visualization is part of model interpretation, not merely presentation.

External validation and transportability. A model that performs well within one squad may fail when applied to another because of differences in athlete level, playing style, training methodology, device technology or data-processing rules. External validation is therefore the strongest test of transportability. When an independent cohort is unavailable, temporal validation in a later season or leave-one-athlete-out approaches can provide more realistic evidence than a random row-level split. The intended deployment environment should be described explicitly so that readers can distinguish a model designed for internal team monitoring from one proposed as a general tool across sports or populations.

Calibration and decision thresholds. Discrimination and calibration answer different questions. A model may correctly rank athletes from lower to higher estimated risk yet systematically overestimate absolute probabilities. Calibration plots, observed-to-predicted comparisons and reliability statistics therefore add information beyond area-under-the-curve measures. Decision thresholds should also reflect the cost of errors. In an injury-screening context, a lower threshold may increase sensitivity but generate more false alarms; in a resource-limited laboratory assessment, a higher threshold may be chosen to restrict follow-up testing. The threshold is a practitioner decision informed by model performance, not a universal property of the algorithm.

Model comparison and baseline performance. Complex models should always be compared with simple baselines. If logistic regression, a decision tree and a gradient-boosted model produce similar out-of-sample performance, the simpler model may be preferable because it is easier to interpret, update and audit. Conversely, a more complex model is justified when it produces a consistent and practically meaningful improvement. Reporting only the best-performing algorithm creates selection bias and can exaggerate the apparent value of ML. A transparent comparison should show the same validation folds, the same outcome definitions and the same evaluation metrics across candidate models [9,10].

Readiness and fatigue classification. Readiness models may integrate subjective wellness, heart-rate response, neuromuscular tests and recent external load. Because the target state itself can be difficult to define, labels such as “fatigued” should not be used casually. A stronger design links the model to an observable outcome, for example a meaningful decrement in jump performance, a change in standardized submaximal heart-rate response, or a clinician-defined status. When subjective ratings are included, they provide valuable athlete information but should be recognized as measurements with their own variability. The model should therefore combine rather than automatically privilege one source of information.

Injury forecasting and class imbalance. Injury prediction creates a severe statistical challenge because injury events are usually much less frequent than non-injury observations. A model can achieve apparently high accuracy simply by predicting that no injury will occur. Evaluation should therefore emphasize sensitivity, precision, recall, calibration and the practical consequences of false-positive and false-negative alerts rather than accuracy alone. Resampling or class-weighting can be useful, but they should be applied only within the training data. More importantly, the prediction target must be clearly defined: injury type, severity, time horizon and exposure denominator can substantially change the problem being modelled [5-8,10,11].

Force-plate analytics. Force plates are particularly suitable for teaching the transition from conventional biomechanics to ML because the raw signal is easy to visualize and the derived variables have clear mechanical meaning. A countermovement jump, for instance, can be described by peak force, impulse, rate of force development, eccentric and concentric phase durations, jump height and asymmetry indices. Conventional analysis may compare selected variables across time; an ML model can evaluate whether the multivariable force-time profile jointly distinguishes fatigue states or predicts a later performance outcome. Nevertheless, the model should not obscure the underlying mechanics. Inspection of force-time curves remains essential for confirming that the extracted features represent genuine movement changes rather than signal-processing artefacts.

Capturing temporal exposure. A single session rarely explains adaptation, fatigue or injury risk. Models often need exposure windows that summarize recent training history. These can include rolling sums, exponentially weighted averages, monotony, strain, changes from baseline or trends in selected variables. However, the choice of window should be justified by the decision horizon rather than selected solely because it maximizes model accuracy. Short windows may capture acute fatigue, while longer windows may represent accumulated fitness or chronic exposure. Because the physiological meaning of a window can differ across sports and training phases, sensitivity analyses using several plausible windows are preferable to selecting one arbitrary period.

Representing internal and external load. Training load is a multidimensional construct. External load may be represented by distance, high-speed running, accelerations, decelerations, mechanical work or sport-specific actions, whereas internal load may be represented by heart rate, perceived exertion, oxygen consumption or derived indices such as TRIMP [12-20]. The source paper used as the structural template demonstrates that alternative TRIMP formulations can produce different relationships with oxygen consumption under the same exercise protocol. For ML applications, this implies that feature selection should not treat all load variables as interchangeable. Multiple representations may be compared, or physiologically justified variables may be retained together if they capture complementary aspects of the athlete response.

Establishing measurement quality. Model quality cannot exceed measurement quality. Wearable sensors may lose satellite reception, chest straps may produce transient heart-rate artefacts, motion-capture systems may lose markers, and force-platform trials may contain incorrect foot contacts or event-detection errors. Quality control should therefore precede modelling and should include plausible physiological ranges, sensor-completeness checks, synchronization checks and protocol-compliance rules. When measurements are removed, the exclusion criteria should be documented. In applied settings, a simple quality flag attached to each session can also be retained as a predictor or used to suppress model output when confidence in the underlying data is insufficient.

Defining the observation unit. Sports datasets can be organized at multiple levels: athlete, match, training session, drill, repetition, gait cycle or individual sensor window. The observation unit must correspond to the intended decision. Session-level workload models may summarize several thousand raw sensor points into one row per athlete-session, whereas movement-quality models may preserve hundreds of time-normalized samples within each repetition. Repeated observations from the same athlete are not statistically independent, and the hierarchical structure should be reflected in the validation plan. Clear definition of the observation unit also helps prevent accidental duplication, pseudo-replication and leakage between training and test sets.

Formulating the sport-science question. The analytical process should begin with a decision question rather than with an algorithm. Examples include whether an athlete is tolerating the planned training dose, whether a movement pattern has changed after fatigue, whether a return-to-sport progression is restoring mechanical symmetry, or whether a combination of workload and physiological markers indicates unusual short-term risk. Defining the question determines the outcome variable, the time horizon of prediction and the appropriate unit of analysis. A session-level readiness model, for example, is fundamentally different from a repetition-level technique classifier. This step also prevents the common mistake of collecting many variables first and only later deciding what the model should predict.

DISCUSSION

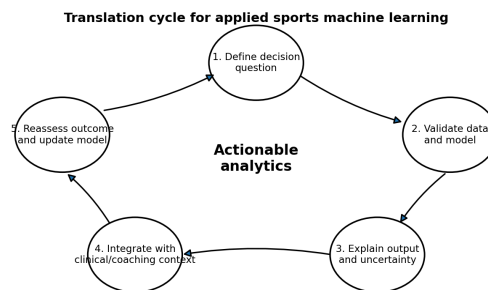
The central finding of this review is that ML adds the greatest value when it is positioned as an extension of established sport-science reasoning rather than as a substitute for it. Load monitoring already distinguishes the work performed from the athlete's response to that work [12-15], and football research has demonstrated the practical value of HR-based and perceived-exertion methods [16-21]. ML becomes useful when these repeated measures are combined with biomechanical, contextual and historical information in a reproducible pipeline.

The strongest methodological challenge is the mismatch between model dimensionality and athlete sample size. Applied sport settings often generate thousands of observations but only a small number of independent athletes, teams or injury events. Treating repeated observations as if they were independent can exaggerate apparent model performance. Validation should therefore respect the data hierarchy, prevent information from the same athlete leaking across training and validation sets, and - when possible - test the final model on a genuinely external cohort [9,10].

Data heterogeneity is a second limitation. Sensor placement, device manufacturer, sampling frequency, filtering procedures, feature definitions and missing-data handling may differ across studies and teams. These differences can alter the statistical distribution of features even when the underlying movement or physiological state is similar. A model that performs well within one acquisition ecosystem may therefore fail when transferred elsewhere. Standardized preprocessing and transparent reporting of signal-processing decisions are necessary for reproducibility.

Interpretability also determines whether predictive output becomes actionable. Coaches and clinicians need to know not only that a model produced a high-risk classification, but which data contributed to that output, how uncertain the estimate is, and what practical response is justified. Feature importance, partial dependence, local explanations or simpler surrogate models can help, but none can compensate for poor study design. The output should remain embedded in a decision pathway in which practitioner judgement, athlete history and current context are retained.

Finally, predictive modelling should be judged against practical utility. A statistically accurate model may have little value if it produces unstable alerts, cannot be updated with new data, or does not change training or rehabilitation decisions. Conversely, a moderately accurate but well-calibrated and interpretable model may be useful as a screening layer. This reinforces the distinction between prediction and decision support.



Recommended reporting structure. For classroom and research use, a reproducible ML sports-science paper should state the sport and competition level, participant number, number of repeated observations, sensing systems and sampling rates, preprocessing steps, feature definitions, model hyperparameters, data-splitting strategy, validation metrics and the intended decision context. Results should include both discrimination or prediction performance and an assessment of calibration or error distribution. The discussion should then separate statistical performance from practical utility. This reporting structure makes it easier for readers to judge whether the model could be reproduced or transferred to another environment.

Comparison with conventional statistical analysis. Machine learning and classical statistics should be viewed as complementary. Traditional inferential models are strong when the objective is estimation, hypothesis testing and interpretable effect quantification. ML is advantageous when prediction, nonlinear relationships, high-dimensional feature spaces or complex interactions are central. A rigorous paper can use both: conventional models can describe group differences and associations, while ML can test whether

multivariable patterns improve out-of-sample prediction. Using a simple baseline model alongside more complex algorithms also helps determine whether additional complexity is justified [9,10].

From group models to individualized models. Population models are useful when data are limited, but individualized modelling is often the long-term objective in high-performance sport. One practical strategy is a hierarchical approach: begin with a group-level model, then update expectations as athlete-specific observations accumulate. This balances statistical stability with personalization. The same logic is already familiar in conventional monitoring, where a coach interprets a session load relative to both team expectations and the individual athlete's normal response. ML provides a formal way to extend that reasoning across a larger number of interacting variables.

Monitoring model drift. Sport environments change continuously. A model trained during preseason may encounter different load distributions during congested competition periods, and a model trained with one GPS generation may behave differently after hardware or firmware changes. Model drift can also arise when coaching philosophy changes or when an athlete returns from injury with a new training profile. Periodic recalibration, performance monitoring and version control are therefore necessary. In a mature workflow, model maintenance is treated as part of the scientific process and all changes to preprocessing, feature definitions and decision thresholds are documented.

Interpretability for coaching practice. Practitioner-facing interpretation should be designed around decisions rather than around algorithmic terminology. Coaches rarely need to know that a model uses a particular tree depth or regularization parameter; they need to know which measured factors are outside the athlete's normal range, how uncertain the estimate is, and what follow-up assessment is appropriate. Useful systems therefore translate model outputs into concise categories, trend plots, or change-from-baseline summaries while retaining access to the underlying measurements. This allows domain experts to challenge an implausible prediction instead of accepting a black-box result.

Female athlete datasets. Female athletes remain underrepresented in many areas of performance analytics, and this has implications for model transportability. Training-load research in women's soccer shows that internal and external monitoring methods are feasible and informative, but the evidence base is smaller than that for male athletes [17,20]. A model developed exclusively in male cohorts should not be assumed to transfer without evaluation. At minimum, sex, competition level, training context and measurement protocol should be clearly reported, and external validation should include the population in which the model is intended to be used.

The role of physiological anchors. Physiological anchors can strengthen model interpretability. Variables such as maximal oxygen uptake, heart-rate response, ventilatory thresholds and established internal-load indices have clear links to exercise physiology and can help place data-driven patterns in context. The source treadmill study demonstrates a classical example: several heart-rate-based TRIMP methods were compared with oxygen-consumption measurements to determine which index most closely reflected physiological load under intermittent and incremental conditions [18-21]. In an ML workflow, similar anchors can be used to test whether a complex model is learning a plausible training-response signal or merely exploiting dataset-specific correlations.

Small samples and repeated measurements. Elite sport frequently presents the apparently contradictory situation of “big data” and small samples. Thousands of rows may be available because each athlete contributes many sessions or repetitions, yet the number of independent athletes may remain small. Models can therefore achieve high apparent accuracy while generalizing poorly outside the original team. Reporting the number of athletes, the number of observations per athlete and the exact validation unit is essential. When possible, external validation in a separate team or season provides stronger evidence than randomly splitting all rows from one squad [9,10].

Prediction versus explanation. A recurring conceptual issue is the difference between predicting an outcome and explaining why that outcome occurs. Machine learning is optimized primarily for prediction and pattern recognition. A model may predict fatigue accurately by combining many correlated variables without identifying a causal physiological pathway. For sport scientists, this means that high predictive performance should not be described as proof that an individual predictor causes the outcome. Mechanistic

interpretation still requires experimental design, temporal reasoning and domain-specific evidence. This distinction is particularly important in injury research, where risk emerges from interacting and changing constraints rather than one isolated exposure [6,8].

Teaching implications. For classroom teaching, ML-based sport analytics provides a useful framework for integrating exercise physiology, biomechanics, statistics and coaching practice. Students can begin with a familiar metric such as heart rate or TRIMP, examine how it relates to an external workload, and then progressively add biomechanical or contextual variables. This progression makes clear that machine learning is not a separate discipline detached from sport science; it is a modelling toolbox whose usefulness depends on the quality of the measurements and the clarity of the sport-science question. Teaching should therefore emphasize data provenance, validation and interpretation before algorithmic complexity.

Robustness under real-world conditions. Laboratory data are usually cleaner than field data, yet many intended applications operate in uncontrolled environments. Competitive schedules, travel, weather, tactical changes and incomplete sensor recordings all introduce variability that may be absent from model development datasets. Robust models should therefore be stress-tested under plausible missingness, sensor noise and changes in workload distribution. Sensitivity analyses can show whether a prediction is stable when one sensor stream is unavailable or when feature values shift moderately. Such testing is particularly important for systems intended to operate automatically, because silent technical failures can otherwise be mistaken for meaningful athlete responses.

Interpreting model improvement. An increase in predictive performance is meaningful only when it is large enough to alter a decision. A small numerical improvement in area under the curve or prediction error may be statistically reproducible but practically irrelevant if it does not change athlete classification, session planning or follow-up assessment. Researchers should therefore report absolute performance differences between the baseline and the ML model and, where possible, describe how many decisions would differ at the selected threshold. This connects model development directly to applied value and discourages the pursuit of complexity for its own sake.

CONCLUSIONS

Machine learning can extend sports and exercise analytics by integrating internal load, external load and biomechanical information across repeated observations. Its principal contribution is not the replacement of established monitoring methods, but the ability to combine them, model nonlinear relationships and generate individualized predictions. Force-plate, motion-capture and wearable data can support training-load management, technique assessment and injury-risk decision support when preprocessing, validation and interpretation are rigorously controlled. Established physiological measures such as heart-rate-based training load remain important as interpretable anchors and benchmarks. The most credible applications therefore combine domain knowledge, transparent modelling, athlete-level or temporal validation, and practitioner-led interpretation. In applied sport, model performance should ultimately be judged by whether it improves the consistency and usefulness of decisions about training, recovery, technique and athlete health.

REFERENCES

1. Bartlett R, Wheat J, Robins M. Is movement variability important for sports biomechanists? *Sports Biomech.* 2007;6(2):224-243. <https://doi.org/10.1080/14763140701322994>
2. Alahaidib R, et al. Biomechanical assessment tools for injury risk prediction and return-to-sport evaluation in athletes: a systematic review. *Cureus.* 2025;17(9):e93210. <https://doi.org/10.7759/cureus.93210>
3. Seshadri DR, Li RT, Voos JE, et al. Wearable sensors for monitoring the physiological and biochemical profile of the athlete. *NPJ Digit Med.* 2019;2:72. <https://doi.org/10.1038/s41746-019-0150-9>
4. Halson SL. Monitoring training load to understand fatigue in athletes. *Sports Med.* 2014;44(Suppl 2):S139-S147. <https://doi.org/10.1007/s40279-014-0253-z>
5. Gabbett TJ. The training-injury prevention paradox: should athletes be training smarter and harder? *Br J Sports Med.* 2016;50(5):273-280. <https://doi.org/10.1136/bjsports-2015-095788>
6. Windt J, Gabbett TJ. How do training and competition workloads relate to injury? The workload-injury aetiology model. *Br J Sports Med.* 2017;51(5):428-435. <https://doi.org/10.1136/bjsports-2016-096040>

7. Bowen L, Gross AS, Gimpel M, Bruce-Low S, Li FX. Spikes in acute:chronic workload ratio associated with a 5-7 times greater injury rate in English Premier League football players: a comprehensive 3-year study. *Br J Sports Med.* 2020;54:731-738. <https://doi.org/10.1136/bjsports-2018-099422>
8. Bittencourt NFN, Meeuwisse WH, Mendonca LD, et al. Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition. *Br J Sports Med.* 2016;50(21):1309-1314. <https://doi.org/10.1136/bjsports-2015-095850>
9. Halilaj E, Rajagopal A, Fiterau M, Hicks JL, Hastie TJ, Delp SL. Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. *J Biomech.* 2018;81:1-11. <https://doi.org/10.1016/j.jbiomech.2018.09.009>
10. Claudino JG, Capanema DO, de Souza TV, Serrao JC, Machado Pereira AC, Nassis GP. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. *Sports Med Open.* 2019;5:28. <https://doi.org/10.1186/s40798-019-0202-3>
11. Rossi A, Pappalardo L, Cintia P, Iaia FM, Fernandez J, Medina D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS One.* 2018;13(7):e0201264. <https://doi.org/10.1371/journal.pone.0201264>
12. Impellizzeri FM, Marcora SM, Coutts AJ. Internal and external training load: 15 years on. *Int J Sports Physiol Perform.* 2019;14(2):270-273. <https://doi.org/10.1123/ijsp.2018-0935>
13. Bourdon PC, Cardinale M, Murray A, et al. Monitoring athlete training loads: consensus statement. *Int J Sports Physiol Perform.* 2017;12(Suppl 2):S2-161-S2-170. <https://doi.org/10.1123/IJSP.2017-0208>
14. Djaoui L, Haddad M, Chamari K, et al. Monitoring training load and fatigue in soccer players with physiological markers. *Physiol Behav.* 2017;181:86-94. <https://doi.org/10.1016/j.physbeh.2017.09.004>
15. Sarmiento H, Clemente FM, Araujo D, et al. What performance analysts need to know about research trends in association football (2012-2016): a systematic review. *Sports Med.* 2018;48(4):799-836. <https://doi.org/10.1007/s40279-017-0836-6>
16. Rabbani A, Kargarfard M, Castagna C, et al. Associations between selected training-stress measures and fitness changes in male soccer players. *Int J Sports Physiol Perform.* 2019;14(8):1050-1057. <https://doi.org/10.1123/ijsp.2018-0462>
17. Costa JA, Rago V, Brito P, et al. Training in women soccer players: a systematic review on training load monitoring. *Front Psychol.* 2022;13:943857. <https://doi.org/10.3389/fpsyg.2022.943857>
18. Silva P, Santos ED, Grishin M, et al. Validity of heart rate-based indices to measure training load and intensity in elite football players. *J Strength Cond Res.* 2018;32(8):2340-2347. <https://doi.org/10.1519/JSC.0000000000002057>
19. Ellis M, Penny R, Wright B, et al. The dose-response relationship between training-load measures and aerobic fitness in elite academy soccer players. *Sci Med Footb.* 2021;5(2):128-136. <https://doi.org/10.1080/24733938.2020.1817536>
20. Alexiou H, Coutts AJ. A comparison of methods used for quantifying internal training load in women soccer players. *Int J Sports Physiol Perform.* 2008;3(3):320-330. <https://doi.org/10.1123/ijsp.3.3.320>
21. Casamichana D, Castellano J, Calleja-Gonzalez J, et al. Relationship between indicators of training load in soccer players. *J Strength Cond Res.* 2013;27:369-374.
22. Iturricastillo A, Yanci J, Granados C, et al. Quantifying wheelchair basketball match load: a comparison of heart-rate and perceived-exertion methods. *Int J Sports Physiol Perform.* 2016;11(4):508-514. <https://doi.org/10.1123/ijsp.2015-0257>
23. Smith NDW, Peiffer JJ, Girard O, et al. Self-paced cycling at the highest sustainable intensity with blood flow restriction reduces external but not internal training loads. *Int J Sports Physiol Perform.* 2022;17(8):1272-1279. <https://doi.org/10.1123/ijsp.2022-0021>
24. Aoki M, Ronda L, Marcelino P, et al. Monitoring training loads in professional basketball players engaged in a periodized training program. *J Strength Cond Res.* 2017;31(2):348-358. <https://doi.org/10.1519/JSC.0000000000001507>
25. Conte D, Kolb N, Scanlan AT, et al. Monitoring training load and well-being during the in-season phase in national collegiate athletic association division I men's basketball. *Int J Sports Physiol Perform.* 2018;13(8):1067-1074. <https://doi.org/10.1123/ijsp.2017-0689>