



AI-Enabled Wearable Sensor Technologies for Athlete Performance Monitoring: A Critical Review of Signal Processing, Machine Learning, and Real-Time Embedded Systems

Dr. Vikas Kundu

Principal, Geeta College of Education, Butana Kundu, District Sonapat, Haryana- 131302 (India)

(Affiliated to Deenbandhu Chhotu Ram University of Science & Technology- DCRUST, Murthal)

Email ID: dr.vikaskundu@gmail.com

ORCID: <https://orcid.org/0009-0000-0148-317X>

Abstract

Wearable sensing and artificial intelligence (AI) are increasingly being used to quantify athlete movement, physiological response, training load, and recovery in field and laboratory settings. However, the practical value of these systems depends not only on the predictive capability of an algorithm but also on sensor validity, signal quality, hardware constraints, real-time processing, and the transferability of models across athletes and sports. This review examines the technical pipeline linking wearable electronic sensors to athlete-monitoring decisions, with particular attention to inertial measurement units, global navigation satellite systems, photoplethysmography, electrocardiography, electromyography, pressure sensing, and multimodal systems. It synthesizes evidence on signal conditioning, segmentation, feature engineering, classical machine learning, deep learning, sensor fusion, edge inference, and embedded implementation. Existing reviews indicate that wearable devices can support useful monitoring, but accuracy varies by metric, device, movement pattern, and validation setting; performance under controlled conditions should not be assumed to generalize to competition. Key challenges include motion artifacts, inconsistent protocols, small and imbalanced datasets, leakage between training and test sets, limited external validation, battery and communication constraints, privacy, and unclear decision thresholds. The review proposes a layered reference architecture and a validation framework that links engineering performance to sport-specific usefulness. Future research should prioritize synchronized multimodal datasets, athlete-independent testing, interpretable and uncertainty-aware models, energy-efficient edge AI, secure data governance, and transparent reporting. The article's contribution is an engineering-oriented synthesis that connects sensor physics and embedded constraints with the needs of physical education and sports science.

Keywords: Inertial Measurement Units; Physiological Sensing; Sensor Fusion; Edge Computing; Sports Biomechanics

1. Introduction

Athlete monitoring has moved from periodic observation and isolated laboratory tests toward continuous or repeated measurement during training, competition, and recovery. Wearable technologies make it possible to collect movement and physiological data outside specialized laboratories, while machine learning (ML) can identify patterns that are difficult to capture through simple summary statistics. In sports practice, these systems may support workload characterization, technique analysis, fatigue monitoring, rehabilitation, and



individualized feedback. Their relevance to electrical and electronics engineering lies in the complete measurement chain: transducer design, analog front-end and signal conditioning, sampling, synchronization, communication, embedded computation, and reliable interpretation of sensor outputs.

A wearable device is not a single technology. It may combine accelerometers and gyroscopes in an inertial measurement unit (IMU), a GNSS receiver for outdoor position and speed, optical photoplethysmography (PPG) for pulse-related signals, ECG electrodes for cardiac electrical activity, surface electromyography (sEMG) for muscle activation, pressure or force sensors, and temperature or respiratory sensors. Each modality has different measurement physics, sampling requirements, placement sensitivities, noise sources, and validation standards. The value of a device therefore cannot be judged by a single generic label such as “accurate.” Accuracy must be assessed for a specific metric, activity, population, reference method, and environment.

Reviews of wearable technology in sports medicine describe diverse applications including training, rehabilitation, and performance research, with systems using accelerometers, gyroscopes, force sensors, GPS/GNSS, and IMUs [1]. A systematic review of wearable microtechnology in intermittent team sports also found that validity and reliability depend on device model, sampling rate, movement type, and intensity [2]. Broader reviews of consumer wearables similarly warn that evidence for one biometric outcome does not establish accuracy for all metrics measured by the same device [3]. These findings make a technology-focused synthesis necessary for practitioners and engineers alike.

The present review focuses on the engineering pathway from wearable data acquisition to athlete-monitoring decisions. It asks: (1) which sensor modalities are used to capture athlete movement and physiological responses; (2) how signal processing and ML methods transform raw data into useful indicators; (3) what embedded and edge-computing constraints affect real-time deployment; (4) how the validity and generalizability of reported systems should be evaluated; and (5) which technical and methodological gaps should guide future research. The intended contribution is not to imply that every AI system improves performance, but to identify the conditions under which wearable-AI systems can provide reliable, interpretable, and practically useful information.

2. Review Scope and Approach

2.1. Review Design

This article is a critical, structured review of engineering and sports-science literature on wearable sensors, signal processing, machine learning (ML), and embedded implementation for athlete monitoring. It synthesizes selected peer-reviewed reviews and technical studies to compare sensor modalities, processing pipelines, model-validation practices, real-time constraints, and implementation risks. It is not presented as a completed PRISMA systematic review, because a full multi-database search, duplicate removal, independent screening, and study-level quality appraisal have not been completed and documented for this manuscript. This distinction is important: no database yield, duplicate count, screening count, or final included-study count is claimed.

2.2. Literature Identification and Selection Principles

The literature discussed here was selected for direct relevance to wearable sensing in sports medicine, athlete workload monitoring, and wearable micro technology in team sports, consumer wearable accuracy, and deep-learning-based activity recognition. Priority was given to peer-reviewed review articles and technical papers that describe sensing modalities, validation, or computational methods. The cited works

provide a foundation for the engineering discussion, but the selection process should not be interpreted as exhaustive. The article therefore uses a structured critical-review approach rather than a completed systematic-search claim. For a future systematic version, the recommended sources are IEEE Xplore, Scopus, Web of Science Core Collection, PubMed/MEDLINE, and SPORTDiscus, with exact database-specific strings, dates, filters, exported records, and screening decisions retained in a reproducible log.

2.3. Synthesis Framework

The discussion is organized around five linked questions: (1) what physical or physiological signal does each sensor measure; (2) how is that signal conditioned, segmented, and transformed into features; (3) how are ML models trained and validated; (4) what hardware, latency, energy, and communication constraints affect deployment; and (5) how well do the resulting outputs support real decisions in sport and physical education? Evidence is compared thematically rather than pooled statistically because devices, sports, outcomes, reference standards, and evaluation metrics differ substantially across the cited literature. The article distinguishes direct measurements from algorithm-derived estimates and treats reported accuracy as task-specific rather than universally transferable.

2.4. Scope Limitations

Because this is a critical review and not an exhaustive systematic review, relevant studies may have been missed and selection bias is possible. No PRISMA flow diagram or quantitative study-selection totals are reported. Claims about the overall size of the evidence base are avoided unless supported by a cited review. Before changing this article into a systematic review, the author would need to complete the documented database search, screening, study-level extraction, and quality appraisal and then update the Methods and Results sections accordingly.

3. Wearable Sensor Modalities and Measurement Characteristics

3.1. Inertial Measurement Units

Accelerometers measure linear acceleration, gyroscopes measure angular velocity, and magnetometers can provide an estimate of orientation relative to the magnetic field. Combined IMUs are commonly used for step detection, running mechanics, jump characterization, change-of-direction analysis, stroke or swing recognition, and activity classification. Sensor placement influences what can be inferred: a trunk-mounted IMU may capture gross acceleration and orientation, while a wrist device is convenient but can be strongly affected by arm movement unrelated to locomotion. Sampling frequency should be selected in relation to the temporal dynamics of the task, sensor bandwidth, storage capacity, and energy budget. Higher sampling rates can preserve rapid events but increase data volume and power use; low rates may miss brief impacts or fast changes in movement.

IMU signals are affected by sensor drift, gravity components, soft-tissue movement, attachment looseness, axis misalignment, and differences in device coordinate systems. Calibration, orientation estimation, filtering, and consistent placement protocols are therefore essential. Algorithms trained on one sensor location or one sport may not transfer to another without adaptation. In field settings, a model must also distinguish the target action from incidental movement, equipment vibration, and contact events. Reports should describe sensor placement, attachment, sampling rate, filtering, and the coordinate frame, rather than reporting only the commercial device name.

3.2. GNSS and Local Positioning

GNSS-based tracking is used to estimate position, distance, speed, and movement trajectories in outdoor sports. Some systems combine GNSS with inertial sensing and local positioning technologies. The validity of speed and distance estimates depends on satellite visibility, sampling frequency, antenna placement, movement geometry, and the algorithms used to smooth noisy trajectories. Short sprints, rapid acceleration, sharp turns, and indoor activity are particularly challenging because position estimates can lag or smooth away high-frequency changes. GNSS metrics should be interpreted as estimates, not direct measures of physiological load. Device models and sampling configurations should not be assumed interchangeable without validation [2]. Recent team-sport validation evidence further illustrates that validity differs by physiological outcome and testing environment, reinforcing the need to avoid generalizing one device's performance to all metrics [5].

3.3. Cardiovascular and Respiratory Sensing

PPG uses changes in optical absorption associated with blood-volume variation to estimate pulse rate and, in some devices, derived measures of heart-rate variability. It is compact and suited to wrist-worn devices, but motion, skin-device contact, perfusion, ambient light, temperature, and algorithmic filtering can distort the signal. ECG measures electrical cardiac activity and may provide a more direct reference for heart-rate validation, although electrode contact and motion can still introduce artifacts. Respiratory rate may be estimated using chest expansion, impedance, PPG-derived features, or dedicated sensors; each approach has different error characteristics. Evidence on wearable validity indicates that performance varies by outcome and setting, so the method used to estimate a metric must be stated explicitly [3], [5].

A critical distinction is needed between a directly measured signal and a derived physiological estimate. A watch may measure an optical waveform but calculate heart rate, energy expenditure, recovery scores, or readiness through proprietary algorithms. The latter outputs are not equivalent to direct measurements and may depend on undisclosed assumptions. Studies should identify whether the device provides raw signals, processed signals, or only summary metrics, and should compare derived values with appropriate reference standards under sport-specific movement conditions.

3.4. Electromyography, Pressure, Force, and Smart Textiles

Surface EMG captures electrical activity associated with muscle activation and can inform movement timing, fatigue-related features, or technique analysis. Interpretation is complicated by electrode placement, skin preparation, cross-talk, movement artifacts, and between-athlete variability. Pressure insoles and force sensors can estimate loading patterns, contact time, or distribution of plantar pressure, while textile-based sensors may integrate strain, pressure, or biopotential sensing into garments. Such systems can be useful where conventional laboratory instrumentation is impractical, but they require mechanical durability, stable contact, repeatable calibration, washability where relevant, and validation against accepted reference systems. The engineering challenge is to balance comfort and unobtrusiveness against signal quality and long-term stability.

3.5. Multimodal Sensing and Sensor Fusion

A single sensor rarely captures the full context of athletic performance. IMUs describe movement, GNSS provides outdoor trajectory information, and physiological sensors estimate internal responses. Sensor fusion can combine these complementary signals to improve activity recognition or contextualize workload. Fusion can occur at the raw-signal level, at the feature level, or at the decision level. Early fusion may preserve detailed relationships but requires synchronized sampling and compatible data representations. Late fusion is

easier to modularize and may tolerate missing modalities, but can discard information. Fusion design should account for clock drift, asynchronous sampling, sensor dropouts, differences in latency, and uncertainty in each modality. Adding more sensors is not automatically beneficial: extra hardware increases energy consumption, cost, data-management burden, and opportunities for failure.

4. Signal Processing Pipeline

4.1. Acquisition, Synchronization, and Quality Control

The reliability of an AI model is constrained by the quality of the input signal. A robust pipeline begins with documented sensor configuration, calibration, timestamps, sampling rates, and data-quality checks. In multimodal systems, synchronization is essential because a small timing offset can misalign an impact measured by an IMU with a physiological response or a video annotation. Systems should record missing packets, sensor saturation, implausible values, and periods of poor contact. Automated quality flags can prevent downstream models from generating confident outputs from unusable data. Raw data should be retained where feasible so that preprocessing choices can be audited and revised.

4.2. Filtering and Artifact Reduction

Filtering may reduce high-frequency noise, baseline drift, or movement artifacts, but filters must be selected according to the signal and task. An overly aggressive low-pass filter can suppress meaningful impact peaks; smoothing a PPG signal may remove motion noise while also distorting pulse morphology. Common approaches include finite-impulse-response and infinite-impulse-response filters, adaptive filtering, wavelet denoising, and artifact-rejection rules. Filtering parameters should be reported, including cutoff frequencies, filter order, phase characteristics, and whether filtering is causal. Offline zero-phase filtering may be appropriate for retrospective analysis but cannot be assumed to be available in a real-time embedded system without introducing latency or requiring future samples.

4.3. Segmentation and Feature Extraction

Many wearable pipelines divide continuous signals into fixed or overlapping windows. Window length determines the trade-off between temporal resolution and stability of extracted features. Short windows may detect rapid events but produce noisy estimates; long windows may average away transitions and delay feedback. Traditional features include mean, standard deviation, root-mean-square amplitude, signal magnitude area, peak values, frequency-domain power, spectral entropy, and cross-axis correlations. Domain-specific features may include cadence, stride time, jump height estimates, acceleration load, heart-rate recovery, or EMG amplitude and frequency characteristics. Feature selection should be driven by the task and physical interpretation rather than solely by statistical association.

4.4. Validation Metrics and Statistical Reporting

Classification accuracy alone can be misleading, especially when a dataset contains many examples of common activities and few examples of rare but important events. Reports should include precision, recall, F1-score, sensitivity, specificity, and confusion matrices where appropriate. Regression tasks should report measures such as mean absolute error, root mean square error, bias, limits of agreement, and calibration across the relevant range. Correlation does not demonstrate agreement: two devices can correlate strongly while systematically over- or underestimating a metric. Validation should therefore use metrics matched to the intended decision and, where possible, compare estimates against a suitable reference method.

Confidence intervals and subject-level variation are important for determining whether a reported improvement is meaningful in practice.

5. Machine Learning and Deep Learning

5.1. *Classical Machine Learning*

Support vector machines, random forests, gradient-boosted trees, logistic regression, and k-nearest-neighbour methods remain useful for wearable data, particularly when datasets are small or features have been engineered from known signal characteristics. These models can be computationally efficient and easier to deploy on microcontrollers or mobile devices. They also offer baselines against which deep learning should be compared. A complex model is not necessarily superior when data are limited, labels are noisy, or the target outcome is poorly defined. Researchers should compare models under identical data splits and report hyperparameter selection procedures to avoid optimistic comparisons.

5.2. *Deep Learning for Temporal and Multimodal Data*

Convolutional neural networks can learn local patterns from raw or transformed sensor windows; recurrent architectures such as long short-term memory networks and gated recurrent units can model temporal dependencies; and transformer-based models can represent longer-range relationships when sufficient data and computation are available. Deep learning has become prominent in wearable human-activity recognition, but its effectiveness depends on the size and diversity of the dataset, the quality of labels, preprocessing, and evaluation design [6], [7]. Sports applications add complexity because movement patterns vary across athletes, playing positions, skill levels, fatigue states, surfaces, and equipment.

A common methodological risk is data leakage. If windows from the same athlete or the same training session are randomly divided between training and test sets, highly similar patterns can appear in both sets, inflating apparent performance. When the intended use is to generalize to unseen athletes, the split should be participant-independent. For future-session prediction, sessions should be separated chronologically. External validation at another institution, club, device type, or competition setting provides stronger evidence of transportability than repeated random splits of a single dataset. Studies should state precisely what unit was separated: samples, windows, sessions, athletes, teams, or sites.

5.3. *Personalization, Transfer Learning, and Explain Ability*

Athletes differ in anthropometry, technique, fitness, movement strategy, and physiological response. A population-level model may therefore be less accurate for an individual athlete than a calibrated personalized model. Personalization can use a short calibration session, athlete-specific normalization, transfer learning, or online adaptation. However, personalization introduces practical questions about how much data are needed, how often the model should be updated, and whether changes in performance reflect learning, fatigue, injury, or sensor placement. Explainability methods can identify influential features or provide interpretable rules, but explanations should be checked for stability and physiological plausibility. An explanation that merely reflects a spurious correlation does not establish a causal mechanism.

5.4. *Injury-Risk Prediction and Responsible Interpretation*

Injury-risk prediction is an attractive application but particularly vulnerable to overclaiming. Injuries are relatively infrequent, multifactorial, and affected by exposure, previous injury, workload, recovery, playing context, and individual susceptibility. Models may perform well on retrospective data while producing many false alarms in prospective use. Risk estimates should be communicated as uncertain decision-support

information rather than definitive predictions. Before a model is used to change training or restrict participation, prospective validation, calibration, fairness assessment, and clear responsibility for decisions are needed. A wearable metric should not be treated as a diagnosis or as proof that an athlete will be injured.

6. Real-Time Embedded Systems and Edge AI

6.1. System Architecture

A practical athlete-monitoring system can be organized into five layers: (1) sensing and analog front-end; (2) embedded acquisition and timestamping; (3) local preprocessing and feature extraction; (4) on-device or edge inference; and (5) user-facing feedback and secure data management. Sensor data may be transmitted through Bluetooth Low Energy, Wi-Fi, or another radio technology to a phone, gateway, or server. The appropriate architecture depends on whether immediate feedback is needed, whether connectivity is reliable, and how sensitive the data are. A cloud-only design may support complex analysis but depends on network availability and creates latency and privacy concerns. On-device inference reduces data transmission and can continue offline, but is constrained by memory, compute, battery, and thermal limits.

6.2. Latency, Energy, and Computational Cost

Real-time suitability must be measured rather than asserted. Relevant metrics include end-to-end latency, inference time, throughput, and memory footprint, energy per inference, battery life, packet loss, and recovery from connection failures. The computational cost of a model includes preprocessing and communication, not just neural-network inference. Quantization, pruning, efficient feature representations, and hardware-aware model design may reduce resource requirements, although these optimizations can affect accuracy. For athlete feedback, the acceptable latency depends on the task: technique feedback during a movement may require rapid processing, while post-session workload summaries can tolerate delayed computation. Authors should specify the use case and the threshold that defines “real time.”

6.3. Edge-Cloud Partitioning

A hybrid architecture can perform signal cleaning, event detection, or basic alerts locally while sending selected features or encrypted summaries to a cloud platform for longitudinal analysis. This reduces continuous raw-data transmission and can improve resilience when connectivity is poor. Nevertheless, a distributed design introduces synchronization, software-versioning, model-update, and cybersecurity requirements. Model updates should be traceable, reversible, and tested against known datasets before deployment. When the system operates across clubs or institutions, interoperability and stable data definitions become important; otherwise, nominally identical workload metrics may not be comparable. Reviews of wearable sensors for sports performance and rehabilitation discuss the underlying transduction technologies and the practical trade-offs involved in body-worn monitoring [8]. Systems-level work on internal and external workload monitoring highlights the importance of combining physiological and movement measures with context [9]. A sports-medicine review of AI-enhanced sensing also emphasizes that sensor data must be connected to clearly defined performance or monitoring decisions [10]. Recent field-based synthesis of wearable sensors and machine learning for biomechanical load assessment further points to gaps in sport diversity and external validation [11].

7. Synthesis of Evidence and Comparative Technical Considerations

The literature supports a qualified rather than universal conclusion: wearable systems can capture useful information for athlete monitoring, but validity is metric- and context-dependent. Reviews of sports-medicine wearables document a broad range of devices and applications [1], while the review of microtechnology in intermittent team sports emphasizes the influence of device model, movement type, and intensity [2]. A review of consumer wearable validation found that evidence is uneven across outcomes and devices, reinforcing the need for outcome-specific validation [3]. Deep-learning reviews show that wearable signals can support activity recognition, but generalization and resource constraints remain central issues [6], [7]. Recent work focused specifically on field-based biomechanical load assessment likewise highlights the promise of sensor-plus-ML systems while identifying limited sport diversity and the need for stronger real-world validation [11].

Technology / Data Stream	Typical Outputs	Main Engineering Risks	Recommended Validation
IMU (accelerometer/gyroscope)	Activity, acceleration, cadence, jump or movement features	Placement, drift, attachment, orientation, impact saturation	Motion capture, force plates, annotated video, or validated reference sensor
GNSS / positioning	Distance, speed, trajectory, external load	Signal loss, smoothing, low sampling rate, indoor limitations	Reference tracking, timing gates, course measurements, sport-specific protocols
PPG / ECG	Heart rate and cardiac waveform features	Motion artifact, contact, perfusion, proprietary processing	ECG or validated reference under rest and sport-specific movement
sEMG	Muscle activation timing and amplitude features	Electrode placement, cross-talk, skin impedance, motion artifact	Standardized placement and synchronized reference/clinical protocol
Pressure / force / smart textiles	Contact, loading distribution, strain, movement-related pressure	Calibration drift, durability, fit, hysteresis	Force plates, pressure platforms, mechanical calibration
Multimodal sensor fusion	Contextualized workload, activity and physiological	Synchronization, missing modalities, complexity, energy	Modality ablation, synchronization checks, external

estimates use and field validation

A useful comparison should separate measurement validity from predictive performance. For example, a model may classify movement patterns accurately while using an input signal that is not sufficiently accurate for quantitative biomechanical inference. Conversely, a sensor can measure a signal reliably but the model may not transform it into a useful outcome. Four questions should therefore be assessed separately: (a) does the sensor measure the intended physical or physiological signal; (b) does preprocessing preserve relevant information; (c) does the model generalize to the intended athlete and environment; and (d) does the final output improve a real decision? A strong evaluation addresses all four.

8. Key Challenges and Research Gaps

8.1. *Lack of Standardization*

Studies often differ in sensor placement, sampling rate, device firmware, preprocessing, task definitions, outcome labels, and evaluation metrics. This makes direct comparison difficult and can produce apparently conflicting findings. The field would benefit from shared reporting templates that include hardware specifications, calibration, signal-quality criteria, exact data splits, code or pseudocode, and model evaluation protocols. Benchmark datasets should document sport, position, age group, sex, training status, surface, and environmental conditions while protecting participant privacy.

8.2. *Generalization and Representation*

Many datasets are collected from small convenience samples or from a narrow range of sports and skill levels. Models may learn participant identity, equipment characteristics, or site-specific artifacts instead of the intended movement or physiological relationship. Data should be evaluated across athletes and, where possible, across teams, seasons, devices, and environmental conditions. Sex, age, disability, skin tone where optical sensing is involved, and different levels of fitness should be considered when relevant to sensor performance. Fairness is not achieved merely by including demographic variables; subgroup error and calibration need to be examined.

8.3. *Reliability, Uncertainty, and Decision Thresholds*

A prediction should be accompanied by an indication of uncertainty when it informs consequential decisions. Model confidence may be poorly calibrated under distribution shift, such as a new sport, device, or athlete population. Systems should detect out-of-distribution inputs, poor signal quality, and sensor failure, and should be able to abstain rather than issue a misleading estimate. Decision thresholds should be chosen with coaches and sports scientists based on the relative cost of false alarms and missed events. The threshold should be prospectively evaluated, not selected solely to maximize performance on a test dataset.

8.4. *Privacy, Ownership, and Cybersecurity*

Wearable data can reveal sensitive information about health, routines, location, fatigue, and performance. Teams and institutions should establish clear rules about who owns the data, who can access raw signals, how long data are retained, and whether they may be used for model training. Consent should distinguish monitoring for immediate coaching from secondary research or commercial reuse. Secure pairing, encryption in transit and at rest, access control, software updates, and incident response are relevant to connected systems. Privacy-preserving analytics and on-device processing can reduce exposure, but do not eliminate governance obligations.

8.5. From Laboratory Metrics to Practical Usefulness

Laboratory validation is necessary but not sufficient. A wearable may perform well during a controlled treadmill test yet fail during contact, sweat, rapid changes of direction, equipment collisions, or irregular competition conditions. Field validation should include representative movement patterns and report failures, missingness, comfort, adherence, battery performance, and coach interpretation. The outcome of interest should also be relevant to practice. A small improvement in classification accuracy may have little value if it does not change a training decision, while a modestly accurate early-warning indicator may be useful if it is calibrated and integrated into a broader decision process.

9. Proposed Reference Framework for Future Systems

Based on the reviewed engineering considerations, a robust athlete-monitoring framework should include six connected components. First, a sensor layer should select modalities according to the target construct rather than device availability alone. Second, a signal-quality layer should detect missing, saturated, noisy, or poorly synchronized data. Third, a processing layer should document filtering, segmentation, and feature extraction. Fourth, a model layer should use a baseline-first strategy, athlete-independent validation where appropriate, and uncertainty estimates. Fifth, a deployment layer should measure latency, energy, communication reliability, and failure recovery on the actual target hardware. Sixth, a governance layer should address consent, data minimization, access controls, retention, and transparent communication of limitations.

The framework should be tested using a staged evaluation. Stage 1 validates sensor outputs against reference equipment under controlled conditions. Stage 2 evaluates the processing pipeline on annotated tasks. Stage 3 tests model generalization on unseen athletes and sessions. Stage 4 deploys the system in representative training environments and measures reliability, latency, battery life, and usability. Stage 5 evaluates whether the output changes decisions or improves an agreed outcome without unacceptable harms. Each stage should have predefined acceptance criteria and a record of failure cases. This approach connects electronic measurement quality to practical sports-science value.

10. Recommendations for Researchers, Engineers, and Practitioners

Researchers should publish enough information to reproduce data preparation, feature extraction, and model evaluation. Where licensing permits, code, model configurations, and de-identified benchmark data should be shared. Engineers should prioritize robust attachment, calibration, power management, and fault detection alongside algorithmic accuracy. Sports scientists should define the decision that the metric is meant to support and avoid treating proprietary readiness scores as direct physiological measurements. Coaches should interpret wearable outputs alongside athlete feedback, training context, and professional judgment. Institutions should create governance policies before collecting sensitive continuous data, particularly when monitoring youth athletes or using data for selection and contractual decisions.

For physical education and sports science settings with limited resources, the most effective solution may not be the most complex. A validated low-cost IMU with transparent signal processing may be more useful than a proprietary multimodal system whose algorithms and limitations are opaque. Open-source software, modular hardware, and local processing can reduce cost and dependence on connectivity. However, low-cost does not mean automatically valid: devices should still be tested against an appropriate reference and for the intended task. Implementation plans should include training for teachers and coaches, device maintenance, data literacy, and safeguards against overinterpretation.

10.1. *Sport-Specific Interpretation and Decision Context*

The meaning of a wearable metric depends on the sport, event, position, training phase, and decision it is intended to support. In endurance sports, longitudinal trends in heart rate, pace, elevation, and recovery may be more informative than isolated peaks. In intermittent team sports, high-intensity accelerations, decelerations, changes of direction, and repeated-sprint patterns may be central, but external-load measures should be interpreted alongside playing time, tactical role, and contextual demands. In strength and power training, repetition count, movement velocity, range of motion, and technique may be more relevant than GPS-derived metrics. A review that treats all sports as a single population risks obscuring these differences.

The same measured feature can have different meanings in different settings. A high acceleration peak may represent a purposeful sprint, a collision, a landing, or sensor impact. A decrease in movement intensity may indicate fatigue, tactical instructions, recovery, or a change in participation. Therefore, a model should not convert a signal into a coaching recommendation without context. Systems should combine quantitative outputs with session metadata, event annotations, and the coach's stated monitoring objective. This does not mean that subjective judgment replaces measurement; rather, context determines whether a technically valid measurement answers the practical question.

For physical education settings, the priority may be safe participation, skill acquisition, inclusive engagement, and feedback that helps learners understand movement. A system designed for elite athlete workload management may be unnecessarily complex, costly, or intrusive for school-based physical education. In educational environments, the selection of wearable technologies should be proportional to the learning objective, age group, available infrastructure, and data-protection requirements. Teachers need understandable summaries rather than unexplained risk scores, and students should not be disadvantaged because they decline optional biometric monitoring. These considerations should be part of system design rather than added after technical development.

10.2. *Reference Standards, Calibration, and Measurement Traceability*

A credible wearable system requires a defined reference against which its measurements or predictions can be evaluated. Reference methods should be selected according to the construct: motion-capture systems or force plates may be appropriate for selected biomechanical outcomes; timing gates may be useful for sprint time; ECG may serve as a reference for heart-rate validation; and manually annotated video may support event recognition when annotation rules are clear. No reference method is perfect for every task, and the manuscript should state the limitations of the chosen standard. Comparing a new system only with another consumer wearable, without explaining the reference device's validity, can create circular validation.

Calibration should be reported as an operational procedure, not simply as a manufacturer feature. Authors should state whether calibration was performed before each session, at the start of data collection, or only in the laboratory; whether calibration was individual or device-level; and how drift was monitored. Where devices are reused across athletes, attachment and repositioning can introduce additional variability. The same device may produce different outputs if worn on a different body segment or secured with a different degree of tightness. A reproducible protocol should describe placement landmarks, orientation, attachment materials, firmware version, and any exclusions for poor contact or signal failure.

Measurement traceability also requires attention to units and definitions. Terms such as workload, intensity, fatigue, readiness, and recovery are often used inconsistently. A paper should define each target outcome operationally and avoid treating a proprietary composite score as if it were a directly measured physical quantity. If the algorithm behind a device is undisclosed, that limitation should be reported. Clear definitions

allow readers to distinguish a validated measurement from a predictive proxy and make comparisons across studies more meaningful.

10.3. Data Leakage, Reproducibility, and Reporting Standards

The design of training and test partitions is one of the most consequential choices in wearable-AI research. Continuous recordings are commonly segmented into many overlapping windows. If these windows are randomly distributed across training and test sets, adjacent windows from the same athlete and session can occur on both sides of the split. The resulting test score may measure recognition of familiar patterns rather than generalization to new people or conditions. Participant-independent splitting is therefore appropriate when the intended deployment includes athletes not represented during training. If the system is intended to personalize to a known athlete, evaluation should instead make the calibration and prediction periods explicit and maintain temporal separation between them.

Reproducibility also depends on documenting preprocessing and model selection. All transformations that estimate parameters from the data—including normalization, imputation, feature selection, and dimensionality reduction—should be fitted only on training data and then applied to validation and test data. Hyperparameters should be selected without repeatedly inspecting the final test set. Where class imbalance is addressed through resampling or synthetic examples, these procedures must be applied only within the training data. Reporting only the best score among many undocumented model configurations creates a risk of selection bias.

A useful minimum reporting package should include the dataset source and collection protocol; participant and session counts; sensor specifications and placements; preprocessing parameters; window length and overlap; label definitions; split strategy; model architecture and hyperparameters; baseline comparisons; evaluation metrics with uncertainty; missing-data handling; and code or pseudocode where feasible. Negative or null findings should not be omitted. Such reporting improves the ability to distinguish genuine technical progress from differences caused by data composition or evaluation procedure.

10.4. Implementation, Usability, and Human Factors

Technical performance does not guarantee successful adoption. Wearable systems may be rejected because they are uncomfortable, difficult to charge, unreliable during contact, slow to configure, or burdensome to interpret. Usability testing should include athletes, coaches, sports scientists, teachers, and technical staff, because each group interacts with the system differently. Athletes may prioritize comfort and transparency about data use; coaches may need quick summaries; technical staff may require diagnostics and export options; and researchers may need access to raw data and configuration details. The design should be evaluated across a complete workflow, from device preparation and fitting through synchronization, analysis, feedback, and data export.

A system should provide clear warnings when data quality is inadequate and distinguish missing measurements from low values. For example, an absent heart-rate signal should not be interpreted as an unusually low heart rate, and a disconnected IMU should not be treated as zero movement. Interfaces should communicate the time window, units, confidence or uncertainty, and limitations of each estimate. Where recommendations are generated, users should be able to identify the evidence supporting them and override the output when context requires. Human oversight is particularly important when metrics could influence selection, training restriction, return-to-play decisions, or educational assessment.

Deployment studies should measure adherence and data completeness in addition to algorithm accuracy. A model evaluated on carefully curated laboratory recordings may fail when athletes remove devices, change

placement, sweat heavily, or move through areas with poor connectivity. Reports should provide the proportion of usable recording time, number of failed sessions, reasons for missing data, charging frequency, and maintenance needs where these are available. These measures reveal whether a system is practical outside a controlled demonstration.

10.5. Cost, Accessibility, and Sustainability

The cost of a wearable-AI system extends beyond the purchase price of sensors. It includes replacement straps and electrodes, charging equipment, software subscriptions, data storage, connectivity, technical support, training, calibration equipment, and the staff time needed to interpret outputs. Institutions should compare total cost of ownership with the value of the decision the system is expected to improve. A more expensive platform may be justified if it provides validated measurements and dependable support, while a lower-cost device may be preferable for broad educational use if it meets the required level of validity. Neither price nor technological sophistication is a substitute for evidence.

Accessibility is also a design requirement. Systems should accommodate different body sizes, movement abilities, skin and hair characteristics relevant to optical or electrode contact, and diverse sporting contexts. A device validated only among highly trained adult men should not automatically be assumed to perform equally in children, women, older adults, para-athletes, or recreational participants. Researchers should report who was included and where the system has not been tested. In physical education, monitoring should support participation and learning rather than create unnecessary surveillance or ranking of students.

Sustainability considerations include battery replacement, device lifetime, repairability, electronic waste, and the energy cost of continuous data transmission and model inference. Edge processing may reduce the need to transmit raw data continuously, but the energy and lifecycle implications depend on the actual hardware and usage pattern. Future studies could report battery life under realistic sampling and transmission settings, device repair options, and the amount of data stored or transferred per session. These factors are particularly important for schools, community sports programmes, and institutions with limited technical infrastructure.

10.6. Proposed Minimum Technical Reporting Checklist

To improve comparability, future papers should report a minimum set of technical details. The following checklist is proposed as a practical synthesis for wearable-AI athlete-monitoring studies:

1. Intended use: sport, population, decision supported, and whether the system is for laboratory analysis, training feedback, competition, rehabilitation, or physical education.
2. Hardware: sensor type, manufacturer and model where relevant, placement, orientation, sampling rate, resolution, firmware, calibration procedure, and attachment method.
3. Data quality: synchronization method, signal-quality thresholds, missing-data rules, artifact handling, and proportion of unusable data.
4. Processing: filtering, segmentation window and overlap, feature definitions, normalization, and whether the pipeline is causal or offline.
5. Model: algorithm, baseline models, hyperparameter selection, handling of imbalance, and computational requirements.
6. Evaluation: reference standard, participant/session/site split, external validation, confidence intervals, relevant metrics, and subgroup performance where appropriate.
7. Deployment: inference latency, memory footprint, energy use, battery life, communication requirements, and failure recovery on the intended hardware.
8. Governance: consent, data ownership, storage and retention, access controls, security safeguards, and

intended limits on use.

9. Reproducibility: protocol, code or pseudocode, model version, data dictionary, and sufficient information to recreate the analysis where ethically and legally possible.

This checklist is not a validated reporting guideline and should be presented as a proposed framework, not as an established consensus standard. Its purpose is to help authors expose the information needed to judge whether a result is reproducible and whether a system can be implemented in a real sports setting. It can also help reviewers separate evidence of algorithmic performance from evidence of reliable measurement and useful deployment.

10.7. *Prioritized Future Research Agenda*

Several research priorities follow from the synthesis. First, multi-site datasets should be developed using harmonized protocols and clear descriptions of sport, position, training level, sensor placement, and reference methods. Data sharing should use appropriate consent and governance controls, but where raw data cannot be released, authors can still share data dictionaries, analysis code, preprocessing procedures, and aggregate results. Second, external validation should become a routine expectation for claims of generalization. A model that performs well on one dataset should be tested on new athletes, sessions, devices, and realistic environmental conditions before it is promoted as broadly applicable.

Third, energy-aware model development should be integrated into the research question from the beginning. Rather than reporting accuracy alone and optimizing hardware later, studies should evaluate the trade-off between predictive performance, latency, memory, energy, and robustness. Fourth, uncertainty estimation and out-of-distribution detection should be explored for systems used in high-stakes contexts. Fifth, prospective studies should evaluate whether the system changes a decision or improves a relevant outcome, not merely whether it classifies a signal correctly. Finally, interdisciplinary collaboration is necessary: engineers can optimize sensing and computation, sports scientists can define meaningful constructs, coaches can specify workflow needs, and data-protection specialists can assess governance and risk.

A useful future study could compare a low-power IMU-only pipeline with a multimodal IMU-plus-physiological pipeline under the same participant-independent evaluation protocol. The study could quantify not only classification or regression performance but also the incremental value of each modality, battery consumption, data transmission, latency, and usability. Such ablation and deployment analysis would help determine whether the added complexity of multiple sensors produces a meaningful practical benefit. This is a proposed research design, not a claim that such an experiment has already been completed.

10.8. *A Practical Benchmarking Protocol for Wearable-AI Systems*

A benchmarking protocol should be designed around the intended application before a device or model is evaluated. The first step is to define the target construct precisely—for example, sprint time, jump height, movement class, heart rate, or an estimate of external workload—and specify the decision the output is intended to support. The protocol should then identify a criterion reference, define acceptable error in units meaningful to the task, and pre-register the primary metrics where possible. A model that classifies an activity correctly is not automatically a valid instrument for estimating the magnitude of the movement or physiological response.

The second step is to separate development from validation. Data should be partitioned at the participant, session, or site level in accordance with the intended deployment. A locked test set should not be used for repeated model selection. Where the system is designed for a known athlete, the protocol should define the amount of calibration data, how personalization occurs, and when the model may be updated. External

validation should test the system under different but relevant conditions, including changes in device placement, competition intensity, environment, and athlete characteristics. The third step is to evaluate failure modes as carefully as average performance: poor contact, missing packets, sensor saturation, drift, battery depletion, and unexpected movement patterns should be recorded and quantified.

The fourth step is to measure system-level performance on the target hardware. Report the time from sensing to feedback, memory use, energy consumption per inference or per hour, data transmission volume, and battery life under realistic sampling and communication settings. If inference is divided between a wearable, smartphone, and cloud service, report latency and failure behaviour for the complete pipeline rather than only the model's isolated runtime. Finally, conduct a usability evaluation with representative athletes and coaches. The system should communicate units, confidence, missingness, and limits in language users can understand. This protocol links sensor validity, model generalization, hardware efficiency, and decision usefulness, which should be evaluated as distinct but interdependent properties.

11. Limitations of the Review

This review focuses on the engineering chain linking wearable sensors, signal processing, ML, and embedded deployment, rather than providing a pooled meta-analysis of a single outcome. The cited literature varies in devices, sports, populations, outcomes, and evaluation methods, limiting direct quantitative comparison. The article uses a structured critical-review approach and does not claim exhaustive systematic coverage. Selection bias and missed studies are possible. The conclusions should therefore be interpreted as a critical synthesis of selected literature rather than a complete census of all available evidence. A future systematic review should follow PRISMA 2020 and PRISMA-S, retain the exact search strategies and exported records, document selection decisions, assess study quality, and provide a completed flow diagram.

12. Conclusion

AI-enabled wearable sensor systems offer a promising means of monitoring athlete movement and physiological response, but their usefulness depends on the integrity of the entire measurement and computation pipeline. Sensor choice, placement, calibration, signal processing, model validation, embedded resource constraints, and data governance jointly determine whether a reported metric is trustworthy. Current evidence supports the use of wearable systems for selected monitoring tasks while warning against assuming that accuracy transfers across metrics, devices, athletes, and competitive settings. Future progress requires stronger external validation, standardized reporting, multimodal synchronization, participant-independent testing, interpretable and uncertainty-aware models, and efficient edge deployment. The central engineering priority is to build systems that are not merely intelligent, but measurable, reproducible, robust, secure, and useful for real sports decisions.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

No new participant data were collected or analysed for this critical review. The evidence discussed is drawn from the cited published literature. No separate dataset was generated for this manuscript.

AI Usage Disclosure

Generative AI tools were used to assist with language drafting, organization, and editing. The author must independently verify all factual claims, citations, reference metadata, and interpretations and should revise this statement to name the tool/version and purpose in accordance with the journal's policy.

Author Contributions

Dr. Vikas Kundu: Conceptualization; Literature Interpretation; Writing – Original Draft; Writing – Review and Editing. The Author has Read and Approved the Manuscript.

References

- [1] R. J. Olsen, S. S. Hasan, J. J. Woo, D. H. Nawabi, and P. N. Ramkumar, "The fundamentals and applications of wearable sensor devices in sports medicine: A scoping review," *Arthroscopy*, vol. 41, no. 2, pp. 473–492, 2025, doi: 10.1016/j.arthro.2024.01.042.
- [2] Z. L. Crang, G. Duthie, M. H. Cole, J. Weakley, A. Hewitt, and R. D. Johnston, "The validity and reliability of wearable microtechnology for intermittent team sports: A systematic review," *Sports Medicine*, vol. 51, no. 3, pp. 549–565, 2021, doi: 10.1007/s40279-020-01399-1.
- [3] C. Doherty, M. Baldwin, A. Keogh, B. Caulfield, and R. Argent, "Keeping pace with wearables: A living umbrella review of systematic reviews evaluating the accuracy of consumer wearable technologies in health measurement," *Sports Medicine*, vol. 54, pp. 2907–2926, 2024, doi: 10.1007/s40279-024-02077-2.
- [4] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.
- [5] N. Čokorilo, N. Manolopoulos, T. Matijević, and R. Rajović, "How valid are wearable devices in team sports? A systematic review," *Sports*, vol. 14, no. 7, p. 264, 2026, doi: 10.3390/sports14070264.
- [6] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng, and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: A review on advances," *Sensors*, vol. 22, no. 4, p. 1476, 2022, doi: 10.3390/s22041476.
- [7] S. Gupta, "Deep learning based human activity recognition (HAR) using wearable sensor data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100046, 2021, doi: 10.1016/j.jjime.2021.100046.
- [8] R. De Fazio, V. M. Mastronardi, M. De Vittorio, and P. Visconti, "Wearable sensors and smart devices to monitor rehabilitation parameters and sports performance: An overview," *Sensors*, vol. 23, no. 4, p. 1856, 2023, doi: 10.3390/s23041856.
- [9] D. R. Seshadri, R. T. Li, J. E. Voos, J. R. Rowbottom, C. M. Alfes, C. A. Zorman, and C. K. Drummond, "Wearable sensors for monitoring the internal and external workload of the athlete," *npj Digital Medicine*, vol. 2, Art. no. 71, 2019, doi: 10.1038/s41746-019-0149-2.
- [10] S. Chidambaram et al., "Using artificial intelligence-enhanced sensing and wearable technology in sports medicine and performance optimisation," *Sensors*, vol. 22, no. 18, p. 6920, 2022, doi: 10.3390/s22186920.
- [11] B. J. Stetter, J. Unger, S. Sell, and T. Stein, "Wearable sensors and machine learning for field-based biomechanical load assessment in sports: A structured overview of current research," *Current Issues in Sport Science*, vol. 11, no. 3, Art. no. 005, 2026, doi: 10.36950/2026.3ciss005.