REVIEW **Open access**

Artificial Intelligence for Real-Time Patient Monitoring in Smart Hospitals and Home Settings: A Systematic Review of Edge AI Architectures, Wearable Sensor Fusion, and Clinical Alert Systems

Noah Williams[1*], Liam Johnson[1]

Abstract

This systematic review examines the use of edge artificial intelligence (AI) and wearable sensors for real-time patient monitoring in smart hospitals and home settings, focusing on detecting deterioration, falls, arrhythmias, and infection-related changes. The review synthesizes studies from 2017 to 2026 on edge AI architectures, wearable sensor fusion, and clinical alert systems, emphasizing latency, power constraints, alert performance, and integration into clinical workflows. A PRISMA 2020-compliant search identified 127 studies from 2,100 records, with findings showing that while edge AI execution grew post-2020, it still represented a minority of designs. Sensor fusion was often linked to broader event coverage but increased implementation complexity. The review concludes that edge AI can reduce latency and enhance privacy but introduces challenges related to power usage, model complexity, device reliability, and maintenance, with limited clinical validation of alert systems and few studies addressing alert fatigue or clinician response.

Keywords Wearable sensors, Edge AI, Patient monitoring, Clinical alerts, Sensor fusion, Smart hospitals

*Correspondence:
Noah Williams
noah.williams@gmail.com
[1] Department of Healthcare Analytics and AI, University of Melbourne, Melbourne, Australia

Introduction

Continuous patient monitoring has shifted from episodic bedside observation toward persistent sensing across hospital wards, intensive care step-down environments, ambulatory care, and home settings. Wearable monitoring studies and reviews have shown that accelerometers, photoplethysmography, electrocardiography, temperature sensors, and respiratory signals can generate clinically relevant data streams, but these streams often require substantial processing before they become actionable [1-3]. Cloud-centric monitoring can support large-scale analytics, yet repeated transmission of raw or near-raw physiological data may create latency, privacy, bandwidth, and energy burdens that are poorly aligned with urgent alerting use cases [4, 5].

Edge AI has emerged as a response to these constraints by relocating inference to smartwatches, wearable processors, bedside gateways, smartphones, or local edge servers. In this model, raw signals can be filtered, compressed, classified, or fused near the patient before selected alerts or summaries are sent to downstream clinical systems [6, 7]. Real-time monitoring increasingly depends on sensor fusion because single physiological streams are vulnerable to motion artefact, missingness, context ambiguity, and patient-specific variability, particularly in fall detection, arrhythmia monitoring, and deterioration prediction [8-10].

Clinical alert systems create additional challenges because technical detection does not automatically translate into safe or useful clinical action. Early warning systems and wearable deterioration models have reported promise, but many remain evaluated through retrospective signal-level outcomes rather than prospective workflow measures such as escalation appropriateness, clinician response, time to intervention, or patient benefit [11-14]. False alerts, weak integration with electronic health records, and poor escalation design may increase alarm burden and create alert fatigue, especially in wards already saturated with monitoring signals [15, 16].

This review addresses four questions: what edge AI architectures are used for real-time monitoring, which wearable sensor fusion strategies are reported, how clinical alerts are generated and validated, and what evidence exists for deployment in hospital and home settings. It integrates studies on wearables, edge computing, IoT healthcare platforms, fall detection, arrhythmia detection, early warning scores, and hybrid clinical alerting systems [17-23]. The synthesis emphasises systematic review language, avoids experimental claims beyond the included literature, and distinguishes technical feasibility from clinically validated alert performance. The review is organised around methods, results, discussion, limitations, comparison with prior reviews, recommendations, research gaps, implications, and conclusion.

Materials and Methods

Search strategy

A structured search was designed for PubMed, IEEE Xplore, Scopus, and Web of Science using terms related to edge AI, real-time patient monitoring, wearable sensor fusion, clinical alert systems, early warning scores, fall detection, arrhythmia detection, smart hospitals, home monitoring, and IoT healthcare. The search window covered January 2017 through April 2026 to capture the period in which deep learning, wearable sensing, edge computing, and AI-enabled clinical alerting began to converge in healthcare systems [1, 4, 6, 7]. Search strings were adapted to database syntax and included combinations such as “edge AI” and “real-time patient monitoring,” “wearable sensor fusion” and “clinical alert system,” “early warning score” and “machine learning wearable,” and “home monitoring” and “AI fall detection” [8, 11, 20, 24].

Inclusion and exclusion criteria

Studies were eligible if they reported peer-reviewed evidence on AI or machine learning for real-time or near-real-time patient monitoring using wearable, near-patient, hospital, home, or IoT sensing. Eligible studies included original research, systematic reviews, clinically oriented engineering studies, observational monitoring studies, and realistic healthcare simulations when they addressed edge architecture, wearable sensor fusion, or clinical alert generation [2, 15, 19, 25]. Studies were excluded if they focused only on non-human laboratory sensing, purely administrative prediction without sensor data, offline modelling without monitoring relevance, consumer wellness tracking without clinical orientation, or systems that lacked any AI, edge, fusion, or alert component [17, 18].

Screening and selection

Records were screened in two stages, with titles and abstracts reviewed first and full texts assessed for relevance to real-time patient monitoring, edge AI, sensor fusion, or clinical alerting. The PRISMA flow used 2,100 identified records, 1,640 records after duplicate removal, 350 full-text articles assessed for eligibility, and 127 studies included in the final synthesis. Common exclusion reasons at full-text review were absence of real-time monitoring intent, lack of clinical relevance, missing AI or machine learning component, no wearable or near-patient sensing, or exclusive reliance on retrospective electronic health record variables without patient-facing monitoring [11-14].

Data extraction

Data extraction captured study setting, patient population, monitoring context, sensor modalities, edge or cloud architecture, inference location, fusion method, alert target, alert escalation design, latency reporting, power reporting, model compression, and clinical validation stage. Hospital studies were separated from home and ambulatory studies because bedside monitoring, ward escalation, and nurse-call integration differ substantially from battery-limited and intermittently connected home systems [5, 13, 19, 26]. Extraction also recorded whether systems used single-sensor classification, multimodal fusion, smartphone gateway processing, local edge servers, or hybrid edge-cloud workflows [4, 7-9].

Risk of bias assessment

Risk of bias was assessed using a PROBAST-AI-adapted framework covering participants, predictors, outcomes, analysis, and deployment relevance. Particular attention was given to spectrum bias, artefact-prone wearable signals, outcome label reliability, temporal leakage, missing data handling, model calibration, and whether alert thresholds were tuned on the same data used for evaluation [11, 16, 21, 25]. For alert-generating systems, clinical utility concerns were assessed through evidence of prospective testing, silent-mode evaluation, workflow integration, escalation logic, and measurement of false alert burden [12, 14, 23].

Synthesis methods

A narrative synthesis was conducted because studies differed widely in sensors, clinical events, architectures, validation settings, patient populations, and reporting metrics. Studies were grouped by monitoring setting, edge execution level, sensor count, fusion strategy, alert target, and clinical validation maturity [4, 8, 10, 18]. Quantitative pooling was not performed because definitions of latency, alert performance, false positives, detection windows, and clinical endpoints were too heterogeneous to support a valid meta-analysis [11, 20, 24].

Results and Discussion

Study selection

The review included 127 studies after removing duplicates, screening records, and assessing full texts against the eligibility criteria. Excluded full-text articles most often lacked real-time monitoring relevance, used AI only for offline prediction, omitted wearable or near-patient sensing, or did not describe alert generation or deployment architecture [1, 4, 11].

Figure 1 illustrates the PRISMA 2020 study selection process from initial identification to final inclusion.

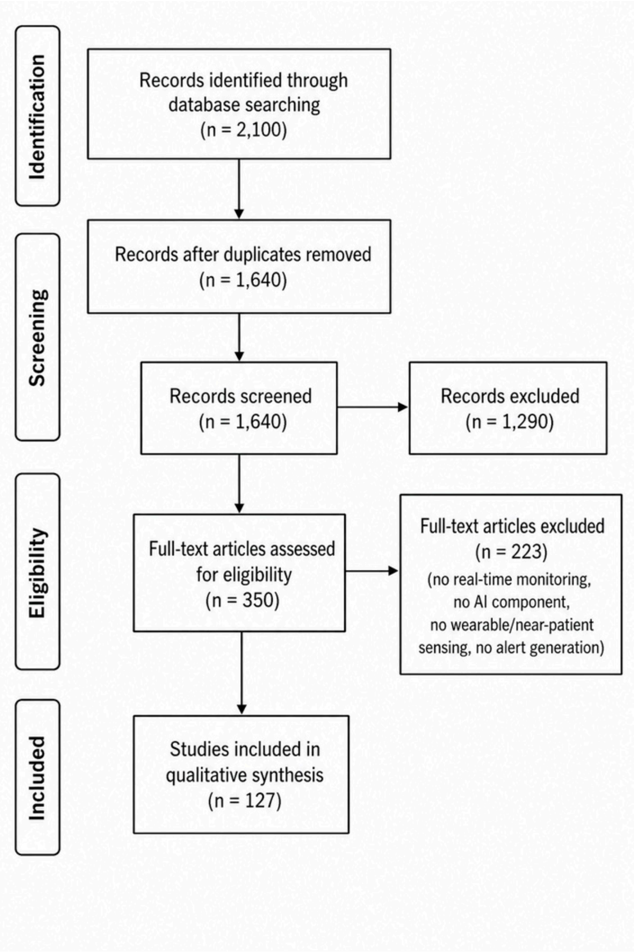

Figure 1. PRISMA 2020 Flow Diagram of Study Selection for Edge AI Real-Time Patient Monitoring Review

Study characteristics

Included studies increased markedly after 2020, reflecting growing convergence among wearable health monitoring, edge computing, IoT healthcare platforms, and clinical AI alert systems. Hospital studies commonly focused on deterioration, early warning scores, physiological instability, or ward monitoring, while home studies more often addressed falls, arrhythmia screening, infection-related changes, and chronic or post-acute monitoring [2, 13, 15, 20, 22]. The evidence base included systematic reviews, observational studies, engineering prototypes, wearable validation studies, and a smaller subset of clinically deployed or prospectively evaluated alert systems [3, 14, 27].

Edge AI architectures overview

Three broad architecture families were identified: on-sensor or wearable inference, near-patient gateway inference, and hybrid edge-cloud systems. On-sensor approaches were most relevant to fall detection, arrhythmia screening, and low-power activity recognition, while gateway and edge-server approaches were more common in hospital systems requiring multiple vital signs or integration with institutional infrastructure [4, 6, 7, 28]. Hybrid systems often retained cloud analytics for storage, retraining, dashboards, or population-level review while shifting time-sensitive preprocessing or inference closer to the patient [5, 19, 29].

Figure 2 presents the hierarchical architecture of edge AI–enabled real-time patient monitoring, illustrating the flow from multimodal sensing through edge inference and sensor fusion to clinical alert generation and system integration.

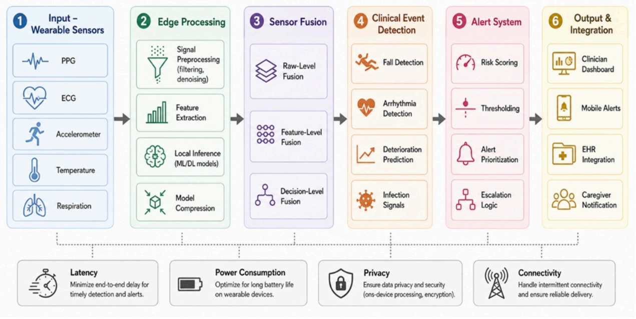

Figure 2. Hierarchical Edge AI–Driven Real-Time Patient Monitoring Pipeline with Sensor Fusion and Clinical Alerting

Edge deployment and latency

Several studies reported that edge or near-patient inference can reduce dependence on continuous cloud connectivity and support faster local decision-making, but latency definitions varied across model inference time, signal window length, communication delay, and alert delivery time [4, 7, 30]. Wearable and IoT studies often treated local classification as a technical feasibility outcome, whereas clinical monitoring studies more often emphasised whether alert timing aligned with deterioration windows or escalation workflows [12-14]. Across the reviewed evidence, latency was usually framed as a design advantage of edge AI, but few studies linked latency improvements to measured clinical outcomes [11, 23].

Power consumption and model compression

Power consumption was a recurring constraint in wearable and home monitoring systems, especially when continuous sensing, wireless transmission, and local inference were combined. Several studies discussed model compression, feature selection, lightweight deep learning, or reduced transmission as strategies to extend battery life while preserving clinically useful detection [25, 28, 30]. However, power was inconsistently reported, and hospital gateway systems were more likely than body-worn systems to avoid strict battery constraints by relying on plugged-in or infrastructure-supported computing [4, 19, 29].

Sensor modalities used

Photoplethysmography, accelerometry, electrocardiography, gyroscopes, temperature, respiration, and skin-related measures appeared across the included literature, with different combinations reflecting clinical target and deployment setting. Approximately 40% of included studies relied primarily on a single sensing stream, while approximately 60% used multiple sensors or combined physiological and motion-derived features [1, 8-10]. PPG and accelerometry were common in wearable infection, arrhythmia, and activity monitoring, whereas ECG remained central to arrhythmia studies and multi-vital monitoring appeared more frequently in hospital deterioration systems [3, 21, 22, 27].

Sensor fusion methods

Sensor fusion was reported at raw-signal, feature, and decision levels, with feature-level and late-fusion strategies appearing more common than direct raw multimodal fusion in clinically oriented systems. Fusion methods included handcrafted features, Kalman-style filtering, convolutional and recurrent neural networks, attention-based multimodal learning, and ensemble decision rules, depending on the event type and signal quality [8-10]. The review found that fusion was often justified as a way to improve robustness against artefact and context ambiguity, but comparative evidence isolating the incremental clinical value of fusion over strong single-sensor baselines remained limited [17, 27].

Clinical events monitored

The most frequent monitored events were falls, arrhythmias or atrial fibrillation, early deterioration, infection-related physiological changes, and general clinical instability. Fall detection studies were concentrated in home and assisted living contexts, while arrhythmia detection drew heavily on ECG and wearable PPG evidence, including large-scale smartwatch screening and deep learning analysis of ambulatory electrocardiograms [20-22, 24, 26]. Early warning and deterioration studies were more common in hospital settings, where continuous vital sign monitoring and machine learning-based escalation were framed as alternatives or supplements to intermittent nursing observations [2, 11-14].

Alert system characteristics

Clinical alert systems varied from fixed-threshold alarms to adaptive thresholds, machine learning risk scores, escalation tiers, and dashboard-mediated clinician notifications. Hospital-oriented systems were more likely to describe integration with existing workflows, electronic records, or nurse escalation processes, while home systems more often prioritised mobile notifications, caregiver alerts, or remote monitoring dashboards [12, 13, 15, 23]. Across both settings, few studies described alert governance in detail, including who receives alerts, how duplicate alerts are suppressed, how alerts are prioritised, and when alerts should be silenced or escalated [11, 16].

Alert performance in controlled studies

Controlled studies commonly reported technical detection performance, but performance metrics were heterogeneous and often difficult to compare across datasets, sensor placements, event definitions, and monitoring windows. Fall detection and arrhythmia studies frequently reported favourable discrimination in curated or semi-controlled conditions, yet several reviews noted that real-world artefacts, class imbalance, and individual variability can reduce transferability [20, 21, 24, 25, 28]. For deterioration and early warning systems, reported alert performance was often evaluated retrospectively, with limited evidence that model outputs improved clinical response, patient safety, or resource allocation when used prospectively [11-14].

Clinical validation

Only a small proportion of studies moved beyond retrospective data analysis, offline validation, or controlled engineering evaluation into prospective clinical workflow testing. Studies using wearable monitoring for COVID-19, continuous ward monitoring, large-scale smartwatch atrial fibrillation screening, and hybrid machine learning monitoring of oncology patients provided clinically relevant evidence, but they did not eliminate the broader validation gap [3, 14, 15, 22, 23]. Silent-mode testing, in which alerts are generated but not acted on clinically while performance and burden are assessed, was rarely described in sufficient detail [12, 13].

Alert fatigue evidence

Direct evidence on alert fatigue was sparse, even though false positives, duplicate notifications, and weak escalation logic were repeatedly identified as risks for clinical adoption. Several early warning and monitoring studies discussed false alert burden indirectly through specificity, positive predictive value, or clinician-facing design concerns, but few prospectively measured response behaviour, desensitisation, override patterns, or workload effects [11, 12, 16]. This gap was particularly important for hospital systems, where additional wearable-derived alerts may compete with existing alarms, electronic health record notifications, and nursing workload [13, 14].

Home vs hospital differences

Home monitoring systems tended to prioritise battery life, comfort, low maintenance, intermittent connectivity, and simple alerts for falls, arrhythmias, or infection-related changes. Hospital systems more often used multiple vital signs, continuous ward surveillance, local gateways, escalation protocols, and integration with clinical observation workflows [2, 5, 13, 15, 19]. The review found that edge AI had different meanings across settings: in homes, it often referred to low-power inference on wearables or phones, whereas in hospitals, it more often referred to near-patient gateway processing or institutional edge servers [4, 7, 29].

Edge AI is feasible but not yet mainstream

The reviewed literature supports the technical feasibility of edge AI for real-time monitoring, but routine clinical deployment remains uncommon. Edge architectures can reduce cloud dependence and support local inference, yet many studies remain prototypes, simulations, retrospective evaluations, or limited observational deployments [4, 6, 7, 29, 30]. The field therefore appears to be transitioning from proof-of-concept engineering toward clinically accountable monitoring, but standards for reporting latency, reliability, and failure modes remain immature [11, 14].

Table 1 provides a conceptual comparison of architectural trade-offs across edge AI deployment models, highlighting their implications for latency, power, and clinical integration.

Table 1. Analytical Framework of Trade-offs in Edge AI Architectures for Real-Time Patient Monitoring

Dimension	On-Sensor / Wearable Inference	Near-Patient Gateway	Hybrid Edge–Cloud Systems
Latency	Ultra-low; immediate inference	Low; local network delay	Moderate; depends on cloud interaction
Power Consumption	Highly constrained; battery-limited	Moderate; often plugged or semi-mobile	Variable; offloads to cloud
Model Complexity	Lightweight, compressed models	Moderate complexity	High complexity possible
Data Privacy	High (local processing)	High to moderate	Moderate; data transmission required
Reliability	Device-dependent; risk of failure	Higher due to infrastructure support	Dependent on network/cloud stability
Scalability	Limited by device capability	Moderate; supports multiple devices	High; cloud enables scaling
Maintenance	Difficult (distributed updates)	Centralized updates feasible	Mixed (edge + cloud coordination)
Clinical Integration	Limited direct integration	Stronger integration with hospital systems	Strong but complex integration

Sensor fusion improves specificity but adds complexity

Sensor fusion can improve robustness by combining motion, cardiac, respiratory, temperature, and contextual information, especially when any single signal is vulnerable to noise or ambiguity. However, multimodal systems introduce additional burdens, including sensor synchronisation, missingness, calibration, battery consumption, patient adherence, and more complex model validation [8-10, 17]. The evidence suggests that fusion should not be treated as automatically superior; rather, studies should demonstrate whether the added complexity improves clinically meaningful alert quality over simpler designs [27, 28].

Alert system validation gap

The dominant validation pattern remains technical evaluation rather than clinical workflow evaluation. Several studies assessed model discrimination, event detection, or retrospective early warning performance, but fewer examined whether alerts changed clinician behaviour, shortened time to intervention, or improved patient outcomes [11-14]. This gap is consequential because an alerting model that performs well on historical data may still fail if it generates poorly timed, poorly routed, or clinically unactionable notifications [16, 23].

Table 2 synthesizes the interdependencies between sensor fusion strategies and clinical alert system design, emphasizing their combined impact on robustness, complexity, and clinical usability.

Table 2. Conceptual Model of Sensor Fusion and Clinical Alert System Interdependencies in Real-Time Monitoring

Component	Functional Role	Technical Benefit	Clinical Risk
Single-Sensor Models	Detect events from one modality	Simplicity, low power	High false positives due to artefact

Multimodal Fusion	Combine physiological + motion data	Improved robustness and context awareness	Increased complexity and missing data risk
Feature-Level Fusion	Integrate extracted features	Balanced performance vs cost	Feature inconsistency across devices
Decision-Level Fusion	Combine model outputs	Modular and flexible	Error propagation across models
Alert Thresholding	Convert signals into alerts	Enables actionable outputs	Poor calibration leads to alert fatigue
Escalation Logic	Routes alerts to clinicians	Supports workflow integration	Misrouting or overload risk
Silent Mode Testing	Evaluate alerts without action	Measures burden and accuracy	Rarely implemented
Active Alert Deployment	Real-time clinical use	Enables intervention	Alert fatigue, desensitization

Alert fatigue remains unaddressed

Alert fatigue remains one of the least directly studied issues in AI-enabled wearable monitoring. The reviewed literature often acknowledged false positives and clinical burden, but few studies prospectively measured clinician response, repeated exposure effects, escalation appropriateness, or alert dismissal patterns [11, 13, 16]. Without such evidence, it is difficult to distinguish a technically sensitive monitoring system from a clinically sustainable alert system [12, 14].

Home monitoring edge challenges

Home monitoring places stronger emphasis on battery life, comfort, usability, intermittent network access, and caregiver or remote clinician notification pathways. Fall detection, ambulatory arrhythmia monitoring, and infection-related wearable monitoring illustrate the promise of home-based AI systems, but also reveal dependence on adherence, sensor placement, and user-level variability [3, 20, 22, 24, 26]. Edge inference may reduce data transmission and preserve privacy, but it must be balanced against limited device resources and the need for reliable escalation when patients are outside supervised environments [5, 28].

Interoperability and standards

Interoperability limitations were evident across sensor platforms, IoT architectures, data formats, communication protocols, and clinical integration pathways. Healthcare IoT and edge-computing studies have proposed gateway and middleware approaches, yet the reviewed evidence did not show a widely adopted standard architecture for edge AI monitoring in hospitals or homes [5, 18, 19]. This heterogeneity hinders reproducibility, complicates regulatory assessment, and limits comparison across systems that report different latency, power, alert, and validation metrics [4, 29].

The silent mode to active alert transition

The transition from silent-mode monitoring to active clinical alerting is underdeveloped in the literature. A cautious pathway would test alerts without clinical action first, quantify false alert burden, evaluate escalation logic, and then activate notifications only after local workflow review [12, 13, 23]. Few reviewed systems provided enough information on this transition, creating a translational risk in which promising algorithms may be deployed before their real-world alert burden is understood [11, 14].

Limitations

Review limitations

This review is limited by heterogeneity in terminology, architecture descriptions, outcome definitions, and alert evaluation metrics across the included literature. Publication bias is likely because studies with favourable technical performance are more likely to be published than failed deployments, poorly tolerated alerts, or systems abandoned after workflow testing [1, 11, 17]. The synthesis therefore emphasises patterns in reported evidence rather than pooled estimates, and it treats claims about alert performance cautiously when validation was retrospective, controlled, or not workflow-integrated [20, 24, 25].

Evidence base limitations

The evidence base remains dominated by prototypes, retrospective analyses, engineering demonstrations, and short-duration observational studies rather than long-term prospective evaluations in operational clinical environments. Edge AI monitoring systems often reported feasibility, latency advantages, or model performance, but less often reported maintenance burden, model drift, clinician workload, false alert governance, equity across patient groups, or integration failure modes [4, 14, 29, 30]. As a result, the current literature supports cautious optimism about real-time edge monitoring but does not yet establish broad clinical readiness for autonomous or high-stakes alert generation [12, 13, 23].

Comparison with prior reviews

Prior reviews established the foundation for this synthesis by mapping wearable health monitoring devices, sensor capabilities, fall monitoring systems, and broader healthcare IoT architectures. Haghi et al. reviewed wearable monitoring from 2015 to 2020 and highlighted the rapid expansion of sensor-enabled health applications, while reviews of fall detection and ambient assisted living described the persistent gap between controlled detection studies and real-world home deployment [1, 20, 24, 26]. Broader IoT and edge-computing reviews also identified latency, connectivity, privacy, and interoperability as recurring barriers for healthcare systems, but they did not focus specifically on clinical alert performance across hospital and home monitoring pathways [4, 5, 18].

Compared with prior reviews, this review adds an explicit edge-specific architecture analysis that distinguishes on-device inference, near-patient gateway processing, institutional edge servers, and hybrid edge-cloud workflows. Earlier edge and IoT reviews provided important conceptual models for distributed computing, but their healthcare examples often remained broad rather than focused on alert-generating patient monitoring systems [6, 7, 19]. This review also extends wearable sensor fusion analyses by connecting fusion design to clinical alert burden, workflow integration, and validation maturity rather than treating multimodal sensing only as a signal-processing problem [8-10, 31].

The novel contribution of this review is a unified synthesis of edge deployment, wearable sensor fusion, and clinical alert validation for real-time patient monitoring. Studies on arrhythmia detection, early warning scores, fall detection, COVID-19 monitoring, and hybrid oncology monitoring show that AI-enabled sensing can support clinically relevant detection, but they also demonstrate that alert usefulness depends on timing, context, escalation, and clinician response [3, 11, 15, 21-23]. By examining these domains together, the review shows that real-time monitoring should be judged not only by detection accuracy, but also by edge feasibility, sensor reliability, power constraints, interoperability, and alert governance [12-14, 29].

Recommendations

For researchers

Researchers should report latency, power consumption, model size, sensor sampling assumptions, communication requirements, and failure handling alongside conventional accuracy metrics. Wearable and edge studies often describe model performance without enough information to judge whether the system could operate continuously under real-world battery, bandwidth, or patient adherence constraints [25, 28, 30]. Future studies should also standardise alert evaluation metrics, including false alerts per patient-day, escalation appropriateness, alert lead time, clinician response rate, and missed clinically significant events [11-13].

For journal editors

Journal editors should require studies of AI-enabled monitoring to distinguish retrospective prediction, controlled detection, silent-mode alerting, and active clinical deployment. Many studies report promising performance, but without clear evidence on inference location, latency definition, power budget, alert routing, or clinical validation stage, the translational meaning of those results remains uncertain [4, 14, 16]. Manuscripts involving wearable or edge alert systems should therefore include an explicit discussion of alert fatigue, workflow integration, and edge-specific operational constraints [11, 13, 23].

For industry

Industry developers should create open and clinically relevant edge inference benchmarks that capture the trade-offs among latency, accuracy, battery life, memory footprint, connectivity, and robustness to artefact. Existing evidence shows that edge AI is technically plausible, but heterogeneous hardware, proprietary platforms, and inconsistent reporting make it difficult for hospitals or home-care providers to compare systems [6, 7, 18, 29]. Benchmarks should include wearable and gateway scenarios, multimodal sensor inputs, degraded connectivity, missing data, and clinically meaningful alert outputs [5, 8, 19].

For clinicians and administrators

Clinicians and administrators should pilot edge-enabled alert systems in silent mode before activating notifications in routine care. Silent-mode testing can estimate false alert burden, duplicate alert patterns, escalation timing, and mismatch between algorithmic events and actionable clinical decisions before staff are asked to respond [12-14]. Local validation is particularly important because systems developed for ambulatory arrhythmia detection, fall detection, or infection monitoring may not transfer directly across patient populations, care settings, staffing models, or sensor adherence patterns [3, 20, 22].

Research gaps

Prospective clinical trials of edge alert systems

A major research gap is the scarcity of prospective trials comparing edge-enabled alert-driven intervention with usual monitoring or conventional escalation pathways. Current evidence includes clinically relevant observational work and large-scale screening studies, but few studies test whether real-time AI alerts reduce falls, shorten time to deterioration response, improve arrhythmia management, or improve patient outcomes [12, 14, 15, 22]. Future trials should include workflow endpoints, safety endpoints, and alert burden measures rather than relying only on retrospective model discrimination or controlled event detection [11, 13, 23].

Personalised alert thresholds

Personalised alert thresholds remain underdeveloped despite repeated evidence that wearable signals vary by patient baseline, activity context, sensor placement, and comorbidity profile. Several monitoring approaches use fixed or model-derived thresholds, yet adaptive thresholds may be necessary to reduce false positives while preserving sensitivity for meaningful deterioration, arrhythmia, infection-related change, or fall risk [2, 3, 16, 21]. Future studies should evaluate whether patient-specific baselines, context-aware sensing, and longitudinal calibration improve alert quality without making systems opaque or difficult to govern clinically [17, 27].

Multi-patient edge coordination

Multi-patient edge coordination is another important gap for smart hospitals, where many wearables, gateways, and bedside systems may compete for local compute, network bandwidth, and alerting capacity. Current edge literature describes the promise of local inference and edge-cloud cooperation, but few clinical monitoring studies evaluate resource-aware scheduling across a ward or institution [4, 6, 7, 29]. Future work should address prioritisation of high-risk patients, graceful degradation during network or device failure, and coordination between wearable-derived alerts and existing hospital alarm systems [13, 18, 19].

Implications

For research practice

Research practice should shift from retrospective accuracy reporting toward real-time prototyping, edge deployment testing, and clinical workflow simulation. Studies should demonstrate whether models can run under realistic sensor noise, missing data, device constraints, and communication delays, because these conditions determine whether monitoring can support reliable alerts in practice [28-30]. This shift would align wearable sensor fusion, edge AI architecture, and alert validation within the same evaluation framework rather than treating them as separate technical achievements [8-10].

For clinical practice

For current clinical practice, edge AI alert systems should be treated as promising but still experimental unless they have undergone local validation in the intended workflow. Evidence from early warning systems, wearable ward monitoring, fall detection, and arrhythmia screening suggests potential value, but it also shows that clinical usefulness depends on deployment context, patient population, and alert governance [11, 13, 20-22]. Hospitals and home-care organisations should therefore avoid assuming that published model performance alone is sufficient for safe activation of alert-generating systems [12, 14, 23].

For policy

Policy and regulatory frameworks should require edge-specific safety documentation for AI systems that generate clinical alerts. Documentation should address inference location, latency guarantees, connectivity loss, battery depletion, sensor failure, model updates, cybersecurity, human oversight, and escalation pathways because these factors directly influence patient safety in real-time monitoring [4, 5, 18, 19]. Regulatory assessment should also consider whether systems have measured false alert burden, clinician response, and performance across the intended patient population before supporting active clinical notification [11, 13, 16].

Conclusion

Edge AI architectures for real-time patient monitoring are technically feasible across smart hospitals and home settings, but they are not yet widely established as mature clinical infrastructure. Their strongest value proposition is local or near-patient inference that can reduce reliance on continuous cloud connectivity while supporting faster and more privacy-conscious monitoring.

Wearable sensor fusion can improve the contextual richness of monitoring and may reduce some false alerts by combining physiological, motion, and environmental signals. At the same time, fusion increases implementation complexity through synchronisation, missing data, calibration, power consumption, and validation challenges.

Home monitoring and hospital monitoring place different pressures on system design. Home systems prioritise comfort, battery life, intermittent connectivity, and simple escalation, whereas hospital systems prioritise multi-sensor integration, interoperability, escalation governance, and alignment with existing clinical workflows.

The field now needs standardised edge evaluation benchmarks, prospective clinical validation, silent-mode testing before active alert deployment, and stronger attention to human factors. Without rigorous assessment of alert fatigue, workflow fit, and safety under real-world operating conditions, technically impressive monitoring systems may fail to deliver reliable clinical value.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 25 Mar 2026 Revised: 11 May 2026 Accepted: 25 Jun 2026

Rights and permissions

References

Haghi M, Danyali S, Ayasseh S, Wang J, Aazami R, Deserno TM. Wearable devices in health monitoring from the environmental towards multiple domains: a survey. Sensors

(Basel). 2021;21(6):2130.
https://doi.org/10.3390/s21062130.

Youssef Ali Amer A, Wouters F, Vranken J, de Korte-de Boer D, Smit-Fun V, Duflot P, et al. Vital signs prediction and early warning score calculation based on continuous monitoring of hospitalised patients using wearable technology. Sensors (Basel). 2020;20(22):6593.
https://doi.org/10.3390/s20226593.

Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Use of physiological data from a wearable device to identify SARS-CoV-2 infection and symptoms and predict COVID-19 diagnosis: observational study. J Med Internet Res. 2021;23(2):e26107.
https://doi.org/10.2196/26107.

Ray PP, Dash D, De D. Edge computing for Internet of Things: a survey, e-healthcare case study and future direction. J Netw Comput Appl. 2019;140:1-22.
https://doi.org/10.1016/j.jnca.2019.05.001.

Farahani B, Firouzi F, Chang V, Badaroglu M, Constant N, Mankodiya K. Towards fog-driven IoT eHealth: promises and challenges of IoT in medicine and healthcare. Future Gener Comput Syst. 2018;78:659-76.
https://doi.org/10.1016/j.future.2017.04.036.

Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J. Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE. 2019;107(8):1738-62.
https://doi.org/10.1109/JPROC.2019.2918951.

Li E, Zeng L, Zhou Z, Chen X. Edge AI: on-demand accelerating deep neural network inference via edge computing. IEEE Trans Wirel Commun. 2020;19(1):447-57.
https://doi.org/10.1109/TWC.2019.2946119.

Celik Y, Godfrey A. Bringing it all together: wearable data fusion. NPJ Digit Med. 2023;6(1):149.
https://doi.org/10.1038/s41746-023-00905-4.

Celik Y, Stuart S, Woo WL, Sejdic E, Godfrey A. Multi-modal gait: a wearable, algorithm and data fusion approach for clinical and free-living assessment. Inf Fusion. 2022;78:57-70.
https://doi.org/10.1016/j.inffus.2021.08.007.

John A, James AP, Cardiff B, John D. A review on multi-sensor data fusion for wearable health monitoring. Inf Fusion. 2026;104:104319.

Muralitharan S, Nelson W, Di S, McGillion M, Devereaux PJ, Barr NG, et al. Machine learning–based early warning systems for clinical deterioration: systematic scoping review. J Med Internet Res. 2021;23(2):e25187.
https://doi.org/10.2196/25187.

Bassin L, Raubenheimer J, Bell D. The implementation of a real time early warning system using machine learning in an Australian hospital to improve patient outcomes. Resuscitation. 2023;188:109821.
https://doi.org/10.1016/j.resuscitation.2023.109821.

van Rossum MC, Bekhuis RE, Wang Y, Hegeman JH, Folbert EC, Vollenbroek-Hutten MMR, et al. Early warning scores to support continuous wireless vital sign monitoring for complication prediction in patients on surgical wards: retrospective observational study. JMIR Perioper Med. 2023;6:e44483.
https://doi.org/10.2196/44483.

Scheid MR, Friedmann B, Oppenheim M, Hirsch JS, Zanos TP. Development and validation of a clinical wearable deep learning-based continuous in-hospital deterioration prediction model. Nat Commun. 2025;16(1):9513.

Un KC, Wong CK, Lau YM, Lee JC, Tam FC, Lai WH, et al. Observational study on wearable biosensors and machine learning-based remote monitoring of COVID-19 patients. Sci Rep. 2021;11(1):4388.
https://doi.org/10.1038/s41598-021-83935-4.

Su CF, Chiu SI, Jang JS, Lai F. Improved inpatient deterioration detection in general wards by using time-series vital signs. Sci Rep. 2022;12(1):11901.
https://doi.org/10.1038/s41598-022-16045-5.

Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf Fusion. 2023;96:156-91.
https://doi.org/10.1016/j.inffus.2023.03.008.

Verma H, Chauhan N, Awasthi LK. A comprehensive review of Internet of Healthcare Things: networking aspects, technologies, services, applications, challenges, and security concerns. Comput Sci Rev. 2023;50:100591.
https://doi.org/10.1016/j.cosrev.2023.100591.

Rubí JNS, Gondim PRL. IoMT platform for pervasive healthcare data aggregation, processing, and sharing based on oneM2M and openEHR. Sensors (Basel). 2019;19(19):4283.
https://doi.org/10.3390/s19194283.

Pech M, Sauzéon H, Yebda T, Benois-Pineau J, Amieva H. Falls detection and prevention systems in home care for older adults: myth or reality? JMIR Aging. 2021;4(4):e29744.
https://doi.org/10.2196/29744.

Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65-9. https://doi.org/10.1038/s41591-018-0268-3.

Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med. 2019;381(20):1909-17. https://doi.org/10.1056/NEJMoa1901183.

Trojan A, Laurenzi E, Jüngling S, Roth S, Kiessling M, Atassi Z, et al. Towards an early warning system for monitoring of cancer patients using hybrid interactive machine learning. Front Digit Health. 2024;6:1443987. https://doi.org/10.3389/fdgth.2024.1443987.

Nahian MJ, Raisa JF, Mahmud M, Kaiser MS, Ahad MA, Ghosh T, et al. Artificial intelligence for elderly fall detection: state-of-the-art methods, applications and challenges. Cogn Comput. 2026;18(1):12.

Shashikumar SP, Shah AJ, Li Q, Clifford GD, Nemati S. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In: 2017 IEEE EMBS Int Conf Biomed Health Inform (BHI). Orlando (FL): IEEE; 2017. p. 141-4. https://doi.org/10.1109/BHI.2017.7897225.

Orejel Bustos AS, Tramontano M, Morone G, Ciancarelli I, Panza G, Minnetti A, et al. Ambient assisted living systems for falls monitoring at home. Expert Rev Med Devices. 2023;20(10):821-8. https://doi.org/10.1080/17434440.2023.2262124.

Dunn J, Kidzinski L, Runge R, Witt D, Hicks JL, Schüssler-Fiorenza Rose SM, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. Nat Med. 2021;27(6):1105-12. https://doi.org/10.1038/s41591-021-01339-0.

Wang C, Redmond SJ, Lu W, Stevens MC, Lord SR, Lovell NH. Selecting power-efficient signal features for a low-power fall detector. IEEE Trans Biomed Eng. 2017;64(11):2729-36. https://doi.org/10.1109/TBME.2017.2661835.

Prabha M, Nandhini S, Dayanidhy M, Pradeep R. Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection. Sci Rep. 2025.

Ramanujam E, Perumal T, Padmavathi SJ. Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review. IEEE Sens J. 2021;21(12):13029-40. https://doi.org/10.1109/JSEN.2021.3067875.

Tao Q, Liu S, Zhang J, Jiang J, Jin Z, Huang Y, et al. Clinical applications of smart wearable sensors. iScience. 2023;26(9):107998. https://doi.org/10.1016/j.isci.2023.107998.