A Review of Edge Intelligence–Multimodal Data Fusion Methodologies for AI-IoT-Based Patient's Predictive Early Warning Systems

Dr. Narendra G. Narole[1], Dr. Lowlesh Nandkishor Yadav[2]

[1] Assistant Professor, ECE Department, Visvesvaraya National Institute of Technology, Nagpur-440010, M.S.; Email ID: pdf.narendra@lincoln.edu.my; drnarendranarole@gmail.com

[2] Associate Professor, Dept. of Computer Engineering, Suryodaya College of Engineering and Technology, Nagpur

Email ID: pdfsv.lowlesh@lincoln.edu.my; Lowlesh.yadav@gmail.com

---

Abstract: Clinical deterioration outpaces recognition in real workflows: threshold-based monitoring floods nurses with low-value alarms, while cloud-centric analytics impose latency and privacy costs that obstruct real-time response. This review synthesizes the ten most recently published advances (2026) in AI-driven IoT-based patient monitoring, categorizing multimodal fusion architectures, edge inference strategies, early warning and alarm-management designs, and deployment-validation practice. Multimodal deep architectures reach AUROC 0.7857 over 5.7 million prediction samples (Sadanandan, 2026) and ICU mortality AUC 0.96 (Abuhamad et al., 2026); transformer-based IoT–EHR fusion cuts sensor-forecasting error by 16.1% and raises clinical-risk AUC by 3.9% (Liu et al., 2026); yet every high-accuracy system remains cloud-resident and retrospectively validated. At the bedside, machine-learning alarms (0.032 versus 0.132 alarms per patient-hour; PPV 0.181 versus 0.073) (Rosario et al., 2026) reduce but do not remove noise, and continuous-monitoring devices still present usability barriers in noncritical care (Pan et al., 2026). A comparative gap analysis of the ten works shows no system simultaneously delivering edge-resident inference, quality-gated multimodal fusion, alarm-aware alerting, and prospective multi-site validation — the design space this review formalizes into a reference architecture.

Keywords: Edge computing; multimodal data fusion; early warning systems; Internet of Things; machine learning; clinical deterioration

Introduction

1.1 Research Problem Statement & Motivation

Modern acute care suffers not from data scarcity but from failure to convert data into timely, trusted alerts. Continuous monitoring devices in noncritical care units remain hard to integrate into nurse workflows, with persistent usability barriers in their deterioration alerting systems (Pan et al., 2026). The alarm problem is quantitative: traditional monitoring emits 0.132 alarms per patient-hour at a positive predictive value of 0.073 (Rosario et al., 2026); machine learning alarms improve this roughly fourfold (0.032 alarms per patient-hour, PPV 0.181) without eliminating noise (Rosario et al., 2026). Meanwhile, the most accurate multimodal models achieve strong discrimination only in cloud resident, retrospective configurations (Abuhamad et al., 2026; Sadanandan, 2026), and prospective multi-site evidence shows performance does not transfer across hospitals without explicit domain adaptation (Papapanagiotou et al., 2026). Real-time bedside conditions therefore demand an architecture coupling multimodal accuracy with edge latency, privacy, and workflow constraints, a coupling none of the ten most recent systems achieves.

1.2 Objectives & Scope

This review categorizes the algorithms, IoT architectures, fusion strategies, and early warning designs reported in the ten most recent works; delineates tradeoffs between localized edge inference and centralized cloud analytics; and identifies gaps separating current systems from deployment-ready real-time prediction.

A standardized taxonomy is proposed for evaluating lightweight neural architectures on resource-constrained hardware under real working conditions: alarm volume, usability, false-alarm rate, and cross-site validation (Pan et al., 2026; Rosario et al., 2026).

1.3 Current Methodologies in Patient Monitoring

Recent literature converges on supervised sequence models, recurrent and transformer architectures for physiological time series, increasingly fused with clinical text and structured EHR features (Liu et al., 2026; Sadanandan, 2026). Wearable sensing is advancing toward multidomain AI pipelines with quality aware fusion and missing modality handling as explicit design elements (Liu et al., 2026). Deployment methodology lags: edge inference is demonstrated only for narrow anomaly detection tasks (Pastore et al., 2026), and alert design addresses false-alarm rates through tiered strategies (Wu et al., 2026) without coupling to continuous multimodal input.

1.4 Proposed Solution & Novelty

Rather than proposing another model, this review performs a framework-level synthesis of the ten most recent works: benchmark dimension-by-dimension, formalize eleven research gaps in five thematic clusters, and translate them into a full stack reference-architecture roadmap for an edge intelligent, multimodal, alert-aware early warning system.

1.5 Key Contributions

A consolidated taxonomy of algorithms, edge architectures, fusion approaches, and alerting designs from the ten most recent (2026) publications; a nine-dimension comparative benchmark exposing row level sparsity; structured identification of eleven gaps in five clusters; and a reference architecture roadmap prioritized for real-time working conditions.

1.6 Paper Organization

Section 2 reviews and compares the ten recent studies; Section 3 details methodology; Section 4 reports results and gap distribution; Section 5 discusses implications for real-time deployment; Section 6 concludes.

2. Literature Review

2.1 Evolution of Prior Research: The Recent Trajectory

The ten most recent works mark a shift from single-modality, cloud-bound modeling toward three currents: multimodal fusion at scale (Abuhamad et al., 2026; Sadanandan, 2026), decentralized wearable sensing (Liu et al., 2026), and alarm-aware deployment (Pan et al., 2026; Wu et al., 2026). Notably, accurate fusion systems and deployment-oriented studies belong to disjoint literatures, accuracy is retrospective, and deployment lacks high accuracy cores (Papapanagiotou et al., 2026; Pastore et al., 2026).

2.2 Early Warning Systems & Predictive Models

Multimodal deep models set the accuracy benchmark: time series EHR fused with clinical notes reaches AUROC 0.7857 over 5.7 million samples (Sadanandan, 2026); integrative hybrid fusion across the MIMIC family attains ICU mortality AUC 0.96 (Abuhamad et al., 2026); transformer-based IoT-EHR coupling reduces sensor forecasting MAE by 16.1% while improving clinical-risk AUC by 3.9% (Liu et al., 2026); and a foundation model pretrained on 1.7 million individuals enables cross device, cross modality transfer (12-lead, single-lead ECG, PPG) (Gu et al., 2026). Prospective multi-site evaluation cautions that transferability cannot be assumed: domain adaptation, not simple transfer, governs cross-hospital performance (Papapanagiotou et al., 2026).

2.3 Algorithms for Monitoring: Edge Inference, XAI, and Human Factors

Edge inference in the recent corpus is confined to anomaly detection: an intelligent fusion IoMT framework achieves 84.69 ms edge oriented fusion latency, proving on device fusion computationally feasible (Pastore et al., 2026). Wearable multidomain AI extends this vision to decentralized care with quality aware fusion (Liu et al., 2026). Human factors dominate deployment findings: usability of deterioration alerts in noncritical care is a documented barrier (Pan et al., 2026), alarm burden studies quantify residual noise in ML alarms (Rosario et al., 2026), and tiered frameworks target false alarm reduction in hospital mortality prediction (Wu et al., 2026). Explainability and federated learning are essentially absent from the ten works open gaps.

2.4 IoT Architectures & Edge Computing Methodologies

Recent architectural contributions are capability demonstrations rather than integrated systems: edge-fusion latency at 84.69 ms (Pastore et al., 2026), wearable biosensing pipelines oriented to decentralized care (Liu et al., 2026), and tiered early warning frameworks designed to address high false-alarm rates (Wu et al., 2026). No recent work delivers a full stack, end-to-end validated instantiation for critical event prediction; reference designs dominate, validated deployments do not (Pan et al., 2026).

2.5 Multimodal Data Fusion Approaches

Fusion strategies span cross modal attention over time series and notes (Sadanandan, 2026), transformer alignment of IoT and EHR streams (Liu et al., 2026), integrative hybrid fusion of waveform, ECG, and note corpora (Abuhamad et al., 2026), and foundation model transfer across sensor types (Gu et al., 2026). All operate in the cloud. Quality gated fusion with missing modality handling appears as a stated principle in the wearable literature (Liu et al., 2026) but is not demonstrated on edge hardware for early-warning endpoints.

2.6 Identified Research Gaps

Coding the ten works against nine capability dimensions, multi-site validation, continuous sensing, multimodal fusion, edge inference, federated learning, personalized alert budgeting, clinician oriented explainability, drift adaptation, full stack architecture yields eleven gaps in five clusters: (A) robustness (edge resident drift detection and recalibration (Papapanagiotou et al., 2026)); (B) fusion at the edge (quality gated ≥4-modality fusion, federated × edge × multimodal convergence, foundation model distillation (Gu et al., 2026; Liu et al., 2026; Pastore et al., 2026)); (C) human factors (personalized alert budgets, cognitive load co-optimization (Rosario et al., 2026; Wu et al., 2026)); (D) explainability (on edge, clinician naturalistic explanation); (E) rigor and coverage (full stack architecture, subgroup stratified evaluation (Abuhamad et al., 2026; Pan et al., 2026)).

2.7 Comparative Summary

Table 1 compares the ten most recent references as the basis for gap identification.

Every high accuracy multimodal system is cloud resident (Abuhamad et al., 2026; Liu et al., 2026; Sadanandan, 2026); the only edge deployed system is task narrow (Pastore et al., 2026); validation and usability evidence come from studies without high accuracy predictive cores (Pan et al., 2026; Papapanagiotou et al., 2026); alarm work addresses noise but not personalization (Rosario et al., 2026; Wu et al., 2026). No system occupies more than two capability dimensions.

Table 1. Comparative summary of the ten most recent references (2026).

Reference	Methodology	Key Strength	Limitation for Real-Time Use
Papapanagiotou et al. (Papapanagiotou et al., 2026)	Transfer learning; prospective multicenter validation	Cross-hospital validation; domain adaptation over transfer	EHR-bound; cloud inference; 539 patients, 3 hospitals
Sadanandan (Sadanandan, 2026)	Multimodal deep learning (time series EHR + notes)	AUROC 0.7857 over 5.7 M samples	Cloud resident; retrospective; no latency analysis
Gu et al. (Gu et al., 2026)	Cardiac foundation model (1.7 M individuals)	Cross device, cross-modality transfer	Cloud only; not distilled/quantized for edge

Liu et al. (Liu et al., 2026)	Multimodal wearable biosensing + multidomain AI	Decentralized care pathway; quality-aware fusion	Roadmap, not a deployed system
Rosario et al. (Rosario et al., 2026)	Traditional vs. ML alarm utility analysis 0.032 vs. 0.132 alarms/patient-hr	PPV 0.181 vs. 0.073	Alarm burden remains; no personalization
Liu et al. (Liu et al., 2026)	TwinFormer transformer fusion (IoT + EHR)	MAE −16.1%; clinical risk AUC +3.9%	Cloud digital-twin design; no edge deployment
Wu et al. (Wu et al., 2026)	AI powered tiered early warning framework	Explicitly targets high false-alarm rates	Continuous sensing integration unproven
Abuhamad et al. (Abuhamad et al., 2026)	Integrative hybrid multimodal fusion	ICU mortality AUC 0.96 Retrospective;	Cloud based; no prospective results
Pan et al. (Pan et al., 2026)	Scoping review: usability of continuous monitoring	Documents real-ward usability barriers	Descriptive; no outcome effects
Pastore et al. (Pastore et al., 2026)	Intelligent fusion for IoMT anomaly detection	84.69 ms edge-oriented fusion latency	Anomaly detection only; not EWS grade

3. Methodology

3.1 Search Strategy & Databases

The review targeted the ten most recently published works (2026) in the corpus, drawn from indexed journals and preprint repositories covering edge intelligence, multimodal fusion, IoT health architectures, and early warning systems; selection emphasized recency and direct relevance to real-time monitoring.

3.2 Inclusion & Exclusion Criteria

Included works must (i) be published in 2026, (ii) substantively engage multimodal fusion, edge/IoT inference, early warning/alarm design, or deployment validation, and (iii) report empirical performance, measured resource metrics, or systematic evidence synthesis. Earlier works were excluded to keep comparison current and tractable.

3.3 Analytical Framework

Each work was coded across nine capability dimensions with four states (addressed, partially addressed, absent, n/a); co absent dimensions were aggregated into five gap clusters and eleven specific gaps. AUROC and latency figures are reported only where the primary study states them with an explicit measurement boundary (84.69 ms edge fusion latency (Pastore et al., 2026); AUROC 0.7857 over 5.7 M samples (Sadanandan, 2026); AUC 0.96 (Abuhamad et al., 2026)).

4. Results

4.1 Algorithm Distribution & Architectural Paradigms

Transformer/attention multimodal fusion dominates the accuracy tier (Abuhamad et al., 2026; Liu et al., 2026; Sadanandan, 2026); the wearable and IoMT tier features quality-aware multidomain pipelines and lightweight edge fusion (Liu et al., 2026; Pastore et al., 2026); the clinical tier is occupied by alarm utility analysis, tiered alerting, usability synthesis, and transfer validation studies (Pan et al., 2026; Papapanagiotou et al., 2026; Rosario et al., 2026; Wu et al., 2026). No recent work pairs high accuracy multimodal prediction with on device inference.

4.2 Algorithm Landscape in Recent Studies

Cross-modal architectures achieve AUROC 0.7857 across 5.7 million samples (Sadanandan, 2026); hybrid fusion across waveform, ECG, and note corpora reaches ICU mortality AUC 0.96 (Abuhamad et al., 2026); Twin Former improves clinical risk AUC by 3.9% and cuts sensor-forecasting error by 16.1% (Liu et al., 2026); a foundation model pretrained on 1.7 million individuals transfers across ECG and PPG devices (Gu et al., 2026). Edge side, quality-gated fusion is demonstrated only as 84.69 ms latency anomaly detection (Pastore et al., 2026); wearable multidomain AI articulates the decentralized vision (Liu et al., 2026). Clinically, ML alarms yield fourfold fewer alarms and 2.5-fold higher PPV yet remain noisy (Rosario et al., 2026); tiered alerting targets false alarms (Wu et al., 2026); usability barriers persist in wards (Pan et al., 2026); and prospective transfer requires domain adaptation (Papapanagiotou et al., 2026).

4.3 Architectural & Methodological Patterns

Three patterns recur: latency feasibility at the edge is proven but endpoint scope narrow (84.69 ms for anomaly detection, not prediction (Pastore et al., 2026)); multimodal accuracy is cloud-sequestered (every high AUC system retrospective, without edge or prospective deployment (Abuhamad et al., 2026; Liu et al., 2026; Sadanandan, 2026)); and deployment evidence exists without predictive cores (usability and alarm studies characterize real conditions but attach to conventional or narrow systems (Pan et al., 2026; Rosario et al., 2026)).

4.4 Gap Pattern Distribution

Table 2. Gap-cluster distribution across the ten most recent references.

Cluster	Gaps	Evidence among the ten	Partially addressing	Fully addressing
A: Robustness	A1: edge resident drift detection and recalibration	Cross-hospital transfer requires domain adaptation (Papapanagiotou et al., 2026)	Papapanagiotou et al. (cloud-side)	None
B: Fusion × Edge	B1: ≥4-modality quality-gated edge fusion; B2: FL × edge × multimodal × critical care; B3: foundation-model distillation	Quality gated principle (Liu et al., 2026); edge latency feasible (Pastore et al., 2026); cloud fusion only (Abuhamad et al., 2026)	Liu et al.; Pastore et al.; Abuhamad et al.	None
C: Human factors	C1: personalized alert budgets; C2: cognitive-load co-optimization	Alarm burden quantified (Rosario et al., 2026); false-alarm reduction	Rosario et al.; Wu et al.	None

		targeted (Wu et al., 2026)		
D: Explainability	D1: on-edge clinician-naturalistic explanations	Absent from all ten works	None	None
E: Rigor & coverage	E1: full-stack architecture; E2: subgroup-stratified edge EWS	Prospective validation without edge (Papapanagiotou et al., 2026); usability without prediction (Pan et al., 2026)	Papapanagiotou et al	None

Every cluster's "fully addressing" cell is empty, quantitative confirmation that the most recent literature still lacks an integrated system. Drift adaptation (A1) and on edge explainability (D1) are the least occupied cells, followed by fusion-on-edge (B1) and alert personalization (C1).

5. Discussion

5.1 Algorithmic Trends & Performance Interpretation

The 2026 literature separates into an accuracy tier and a deployment tier. Accuracy advances are real but architecturally privileged: AUROC 0.7857 (Sadanandan, 2026), ICU mortality AUC 0.96 (Abuhamad et al., 2026), and cross sensor foundation model transfer (Gu et al., 2026) are achieved with cloud scale compute and retrospective cohorts. Deployment advances 84.69 ms edge fusion (Pastore et al., 2026), fourfold alarm rate reduction (Rosario et al., 2026), tiered false alarm management (Wu et al., 2026) are achieved without high accuracy multimodal cores. Neither tier alone satisfies real-time working conditions; the constraint is architectural, not computational.

5.2 Critical Assessment of Methodological Approaches

Four deficiencies recur. Validation bimodality: prospective multi-site validation exists only for EHR-bound transfer learning (Papapanagiotou et al., 2026), while clinical and usability studies lack predictive cores (Pan et al., 2026; Rosario et al., 2026) no work spans both. Unverified deployment claims: only one work reports measured edge latency (Pastore et al., 2026); wearable designs remain roadmaps without measured end to end performance (Liu et al., 2026). Alarm metrics without personalization: alarm burden is quantified precisely (Rosario et al., 2026), but no work treats cognitive load as a co-optimization target (Wu et al., 2026). Human factors disconnected from prediction: usability evidence (Pan et al., 2026) is not attached to any high-accuracy system (Abuhamad et al., 2026; Sadanandan, 2026).

5.3 Implications for Edge-Intelligent Predictive Warning Systems

For current real time conditions, the synthesis dictates four design requirements. Fusion must move to the edge with quality gating quality aware, missing modality aware design (Liu et al., 2026) must be implemented on the hardware class where 84.69 ms fusion latency is already feasible (Pastore et al., 2026), not confined to cloud digital twins (Liu et al., 2026). Robustness requires on device drift adaptation since cross site transfer demands explicit adaptation (Papapanagiotou et al., 2026), drift detection must be edge resident rather than retrospective. Alerting must be workload aware alarm volumes beyond 0.032 per patient-hour will not be sustained by nurses (Rosario et al., 2026); tiered, false alarm-aware designs (Wu et al., 2026) must couple to continuous multimodal input. Validation must be sequential and standardized silent trial deployment then prospective multicenter evaluation with subgroup stratified reporting, mirroring cross-site transfer evidence (Papapanagiotou et al., 2026) and realward usability findings (Pan et al., 2026).

6. Conclusion

6.1 Summary of Literature Review Findings

Three findings summarize the ten most recent works. First, the accuracy frontier is cloud-bound: AUROC 0.7857 (Sadanandan, 2026), AUC 0.96 (Abuhamad et al., 2026), 16.1% forecasting-error reduction (Liu et al., 2026), and 1.7-million-person pretraining (Gu et al., 2026) has no edge or prospective deployment component. Second, the deployment frontier is accuracy-free: edge fusion is limited to anomaly detection (Pastore et al., 2026), and the strongest clinical evidence concerns alarm burden (Rosario et al., 2026), usability barriers (Pan et al., 2026), and cross-site transfer requirements (Papapanagiotou et al., 2026). Third, the gap structure is reproducible: eleven gaps in five clusters, every "fully addressing" cell empty, with drift adaptation and on-edge explainability entirely unoccupied. Recent literature advances each capability in isolation; it has not joined them.

6.2 Future Research Directions

The immediate agenda should prioritize: (i) quality gated fusion of four or more physiological modalities on edge hardware, extending the 84.69 ms latency capability from anomaly detection to prediction (Liu et al., 2026; Pastore et al., 2026); (ii) edge resident drift detection and adapter-based recalibration, benchmarked against cross site domain adaptation evidence (Papapanagiotou et al., 2026); (iii) alarm budget personalization and cognitive load co-optimization, building on quantified alarm burdens and tiered designs (Rosario et al., 2026; Wu et al., 2026); (iv) distillation of foundation models to bedside budgets (Gu et al., 2026); and (v) sequential validation silent trial, then prospective multicenter deployment with subgroup reporting coupling real-ward usability evidence (Pan et al., 2026) to modern multimodal accuracy (Abuhamad et al., 2026; Sadanandan, 2026). The next advance will come not from isolated gains but from an integrated architecture validated under the conditions where monitoring actually happens.

References

1. Abuhamad, H., Zainudin, S., & Bakar, A. A. (2026). “Integrative multimodal hybrid data fusion for mortality prediction”. Scientific Reports, 16(1), 5803–5803. https://doi.org/10.1038/s41598-026-36296-6
2. Gu, X., Tang, W., Liu, Sangha, V., Liu, F., Gowda, S. N., Ribeiro, A. L. P., Schwab, P., Branson, K., Liu, Z., Ribeiro, A. L. P., Liu, Z., & Clifton, D. A. (2026). “Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals”. Nature Machine Intelligence, 8(2), 220–233. https://doi.org/10.1038/s42256-026-01180-5
3. Liu, C., Fan, H., Kim, M., Zhou, T., Yang, P., Zhao, L., Wang, Y. M., Che, Z., Liu, C., Li, B., & Zhu, Y. (2026). “Multimodal Wearable Biosensing Meets Multidomain AI: A Pathway to Decentralized Healthcare”. https://doi.org/10.1002/advs.202522900
4. Liu, W., Wang, J., Ksibi, A., & Yin, N. (2026). “Multi-Modal Fusion of IoT and EHR Data Using Transformer Networks for Holistic Patient in Consumer-Centric Digital Twins”. https://doi.org/10.1109/tce.2026.3672115
5. Pan, J., Dowding, D., Wong, D., Scott, A., & Zhao, Q. (2026). “The Usability of Continuous Monitoring Devices With Deterioration Alerting Systems in Noncritical Care Units: Scoping Review”. https://doi.org/10.2196/75713
6. Papapanagiotou, I., Vrettou, C. S., Theodorakopoulou, M., Mastora, Z., Giannopoulou, V., Kampouropoulou, O., Karalis, A., Pratikaki, M., Golemati, S., Poupouzas, G., Issaris, V., Karkoulias, K.,

Pouriki, S., Keskinidou, C., Lotsios, N. S., Kotanidou, A., Vassiliou, A. G., Papapanagiotou, S., & Dimopoulou, I. (2026). “Prospective Multicenter Validation of Machine Learning Models for Mortality Prediction in Adult Critically Ill Patients using Transfer Learning”. In Research Square. https://doi.org/10.21203/rs.3.rs-8872055/v1

7. Pastore, F., Anwar, R. W., Jabeur, N. H., & Ali, S. (2026). “Intelligent Fusion: A Resilient Anomaly Detection Framework for IoMT Health Devices”. https://doi.org/10.3390/info17020117
8. Rosario, N., Mitchell, H., Zhang, S., Selvaraj, N., Zhang, X., Hernández, C. B., Lipsitz, S. R., & Levine, D. M. (2026). “The Clinical Utility of Traditional and Machine Learning Alarms during the Care of Acutely Ill Patients”. Applied Clinical Informatics, 17(1), 107–117. https://doi.org/10.1055/a-2815-1912

9. Sadanandan, B. (2026). “Multimodal Deep Learning for Early Prediction of Patient Deterioration in the ICU: Integrating Time-Series EHR Data with Clinical Notes”. In arXiv (Cornell University). Cornell University. https://doi.org/10.48550/arxiv.2603.14719

10. Wu, L., Mai, L., Wang, H., Huang, J., He, X., Zhan, X., Khalemsky, A., Kumar, V. A., Paxton, J. H., Tsilimingras, D., Hachimi-Idrissi, S., Liu, S. W., Savioli, G., Rathlev, N. K., Tazarourte, K., Slagman, A., Christ, M., Qureshi, M., Hariri, H., … Bellou, A. (2026). “Artificial Intelligence-powered tiered early warning framework addressing high false alarm rates for in-hospital mortality prediction”. https://doi.org/10.1038/s41746-026-02522-8