

Edge AI-Based Real-Time Physiological Monitoring for Resource-Constrained Wearable Healthcare Systems

Akash L. Malik

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
akash.malik695@buffalo.edu

Ruben Hail

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
rubenhail12@missouri.edu

Abstract

The proliferation of wearable physiological sensors has created new possibilities for continuous health surveillance outside clinical settings. However, cloud-centric processing architectures introduce latency, communication bottlenecks, privacy risks, and energy overheads that are incompatible with real-time intervention. This paper examines the system-level design of edge artificial intelligence for wearable healthcare, focusing on resource-constrained devices that must acquire, process, and interpret photoplethysmography, electrocardiography, electrodermal activity, accelerometry, and related signals. We analyze architectural patterns that distribute inference across sensors, microcontrollers, gateways, and hospital information infrastructures. The discussion emphasizes hardware-software co-optimization, model compression, sensor fusion, context awareness, and the trade-offs among accuracy, latency, battery life, and clinical reliability. Attention-based methods for motion-tolerant optical sensing are considered within a broader framework of multimodal signal restoration. The paper further addresses robustness, fairness, and governance, highlighting the need for regulatory alignment, representative training data, auditability, and transparent deployment. Finally, sustainability and lifecycle concerns are discussed, including device longevity, energy footprint, and responsible update mechanisms. The argument is that real-time edge AI in wearable healthcare is not solely a machine learning problem but a socio-technical infrastructure challenge requiring coordinated advances in architecture, policy, and clinical validation.

Keywords

edge artificial intelligence; wearable healthcare; physiological monitoring; resource constraints; real-time inference; federated learning; governance; sustainability.

1. Introduction

The integration of sensing, computation, and networking into wearable devices has transformed physiological monitoring from episodic clinical measurements into continuous, ambulatory observation. The Internet of Things for healthcare has been surveyed as a broad paradigm in which heterogeneous devices collect health-related data and coordinate through communication infrastructures [1]. Wearable sensors for remote health monitoring now encompass electrocardiography, photoplethysmography, galvanic skin response, temperature,

and accelerometry, enabling longitudinal assessment of cardiovascular, respiratory, and neurological status [2]. These devices promise early detection of deterioration, personalized interventions, and reduced dependence on centralized medical facilities. Yet the volume and velocity of physiological data challenge traditional cloud-centric analytics. Continuous streaming of raw biosignals to remote data centers incurs considerable latency, consumes scarce bandwidth, increases energy demand on battery-powered devices, and exposes sensitive health information during transmission and storage.

Edge artificial intelligence has emerged as a structural response to these constraints. Edge intelligence relocates feature extraction, inference, and decision support toward the data source, allowing wearable systems to operate with reduced dependence on wide-area networks [3]. The convergence of edge computing and deep learning has been examined as a layered architecture in which inference tasks are split among end devices, edge gateways, and cloud services according to latency, privacy, and computational complexity [4]. In resource-constrained wearable healthcare, this convergence is particularly significant because the device itself often lacks the memory, processing capacity, and energy budget required by large neural models. Learning on the edge for the Internet of Things therefore requires careful partitioning of workloads, lightweight model design, and adaptive communication strategies that respect dynamic physiological and environmental contexts [5].

Despite notable progress, the deployment of real-time edge AI in wearable healthcare remains incomplete. Many prototypes demonstrate algorithmic performance in controlled settings but fail to account for motion artifacts, sensor drift, heterogeneous device fleets, intermittent connectivity, and the need for clinical accountability. Systematic reviews of wearable patient monitoring systems identify persistent gaps in interoperability, validation, energy efficiency, and usability [6]. This paper addresses these gaps from a systems perspective. Rather than proposing a single algorithm, it examines the structural trade-offs and architectural choices that determine whether edge AI can be safe, fair, robust, and sustainable in everyday healthcare. The objective is to provide an interdisciplinary synthesis of design principles, infrastructure requirements, and governance considerations for real-time physiological monitoring on constrained wearable platforms.

2. System Architecture and Edge AI Paradigms for Wearable Physiological Monitoring

A practical edge AI wearable system is best understood as a distributed computing hierarchy. At the lowest layer, body-worn nodes acquire multimodal biosignals. Multi-sensor fusion in body sensor networks has been identified as essential for reducing uncertainty and reconstructing physiological state under ambulatory conditions [7]. The clinical perspective on healthcare IoT emphasizes that aggregation of heterogeneous sensor streams must preserve temporal alignment, calibration metadata, and provenance if downstream decisions are to be trusted [8]. An edge gateway, typically a smartphone or dedicated home hub, provides a second layer that can perform more demanding inference, manage device pairing, and hold temporary data when connectivity fails. Fog computing provides a conceptual model for this intermediate layer by placing storage, networking, and computation closer to end users, thereby avoiding the latency and bandwidth penalties of centralized cloud processing [9].

A layered architecture also separates concerns that are often conflated in monolithic cloud systems. On-node processing performs filtering, peak detection, and lightweight anomaly screening; gateway processing executes larger transformer or convolutional models; cloud processing handles population-level analytics, electronic health record integration, and model retraining. This separation supports graceful degradation: when the wide-area network is

unavailable, the gateway can continue local inference and store encrypted summaries for later synchronization. Such designs reflect the broader fog computing principle that latency-sensitive tasks should be handled near the data source while latency-tolerant tasks can be deferred [9]. In wearable healthcare, the clinical value of this distribution depends on semantic interoperability among layers, including consistent units, sampling rates, device identifiers, and alarm thresholds. Without semantic alignment, local decisions may conflict with central policies and erode clinician confidence.

Wearable respiratory monitoring offers a concrete illustration of how modality-specific constraints shape architectural choices. Respiratory signals derived from chest impedance or accelerometry can be used to detect apnea, dyspnea, and activity-related changes, but they are vulnerable to motion and posture [10]. Attention-based methods have been explored to improve opto-physiological monitoring during physical activity by learning which signal segments and channels carry reliable information under motion artifacts [11]. Although algorithmic details vary, the architectural implication is significant: a resource-constrained wearable can learn to weight reliable photoplethysmography segments while suppressing motion-corrupted windows, but only if the attention layer is compressed to fit within tens of kilobytes of memory and computed within millisecond-scale inference budgets. This requirement places model compression and hardware-software co-optimization at the center of edge wearable development.

Beyond computational layering, an effective edge AI architecture must reflect clinical workflows. A gateway may detect a suspected arrhythmia and present a context-rich summary to the user or clinician, but the summary must include confidence, signal quality, and recent physiological context. If this context is stripped away by aggressive compression, downstream decisions become opaque. The architectural challenge is therefore not merely placing a neural network on an embedded processor; it is designing the entire data path so that local inference preserves enough provenance for clinical reasoning. This requirement influences memory allocation, communication protocols, and model update policies.

3. Model Compression and Hardware-Software Co-optimization

The first structural obstacle is the mismatch between state-of-the-art deep learning models and the capabilities of wearable microcontrollers. Embedded machine learning platforms such as TensorFlow Lite Micro have demonstrated that neural inference can run on microcontrollers, but only when model size, operator count, and memory allocation are aggressively controlled [12]. Wearable devices often use ARM Cortex-M class cores with static random-access memory below one megabyte, and audio or multi-channel biosignal inputs can exhaust this budget rapidly. Therefore edge AI design begins with model selection, operator minimization, and memory planning rather than with training accuracy alone.

Quantization has become a central co-optimization technique because many low-power processors lack floating-point units or suffer large energy penalties for floating-point arithmetic [13]. By representing weights and activations with integer types, designers can replace costly multiply-accumulate operations with fixed-point alternatives. However, quantization must be performed with careful calibration to preserve rare physiological events that may have small activation magnitudes. Aggressive quantization can suppress subtle electrocardiographic features, which may be clinically important even when overall classification accuracy remains high.

Model size reduction also depends on pruning redundant connections and applying entropy coding to the remaining parameters [14]. Deep compression illustrates that large models contain substantial redundancy that is not necessary for inference on constrained devices. Yet pruning a model trained for cloud deployment can produce brittle edge behavior because irregular sparsity may not map efficiently to commodity microcontroller memory hierarchies. Structured pruning, knowledge distillation, and neural architecture search are therefore more common in edge wearable design because they produce models with predictable memory access patterns and operator schedules.

These compression choices must be aligned with the broader insight that deep learning capacity is useful only when embedded in a reliable inference infrastructure [15]. The architectural lesson is that hardware-aware model design is not a post-training step; it must influence data sampling, feature extraction, and alarm logic. A wearable system that saves energy by detecting low-quality signal periods and disabling costly inference may provide more reliable real-time monitoring than a more accurate model that drains the battery before the end of a clinical shift.

4. Multimodal Fusion, Context Awareness, and Federated Learning

Physiological measurements are rarely reliable in isolation. Motion, sweat, electrode displacement, and optical path changes can corrupt individual channels. Multimodal fusion therefore improves robustness by combining electrocardiography, photoplethysmography, accelerometry, and galvanic skin response. However, fusion is not free. Synchronizing heterogeneous sampling rates and sensor clocks requires buffering that increases latency and memory consumption. Edge AI architectures must decide whether fusion occurs at the signal level, feature level, or decision level. Signal-level fusion preserves the most information but imposes the highest communication and computation burden. Decision-level fusion is cheaper but may lose subtle cross-modal correlations. These trade-offs shape the placement of fusion nodes across the wearable, gateway, and cloud layers.

Context awareness also changes what should be inferred. A heart rate increase during sleep has different clinical meaning than the same increase during stair climbing. Edge systems can use accelerometry to suppress exercise-induced changes from alarm logic, reducing false alerts. However, context models introduce additional state variables and require transition handling. Probabilistic state machines or lightweight recurrent networks can track activity context, but they must be robust to abrupt changes and sensor dropout. This context layer is often missing from cloud-centric evaluations that assume clean, stationary input.

Because training data are distributed across patients, devices, and institutions, centralizing raw physiological data creates privacy risks and regulatory burdens. Federated learning allows edge devices to compute model updates locally and share only those updates with a coordinating server [16]. This approach fits wearable healthcare because sensors collect sensitive, continuous biological signals that users may not wish to upload. Federated learning also enables personalization by continuing local adaptation on the user's device. However, the future of digital health with federated learning faces practical obstacles including non-identically distributed data, heterogeneous device capabilities, and the need for secure aggregation [17]. Edge AI architectures must support such training loops without interrupting real-time inference.

5. Robustness, Fairness, and Clinical Validation

Real-time edge AI must be robust to distribution shift, sensor drift, and demographic variability. A model that performs well on young healthy volunteers may fail on older patients with darker skin tones or low perfusion. Bias and fairness in machine learning have been extensively surveyed, and the sources of bias are not always obvious because biosignal morphology varies across populations [18]. Wearable optical sensors, for example, may produce weaker photoplethysmographic signals in patients with high melanin concentration or peripheral vascular disease. If training data underrepresent these groups, attention mechanisms may learn to ignore the very signal components that are clinically necessary. Fairness in edge AI therefore requires representative data collection, transparent performance reporting across subgroups, and recalibration on deployment.

The deployment of deep learning in healthcare also requires that models fit within clinical workflows and contribute to actionable decisions [19]. A real-time arrhythmia detector is clinically useful only if it generates alerts that clinicians can interpret, trust, and act upon without excessive alarm fatigue. This implies that edge systems must expose confidence estimates, signal quality indicators, and contextual information. Black-box inference at the sensor is inappropriate when patient safety depends on downstream reasoning. Machine learning in medicine similarly highlights the gap between retrospective accuracy and prospective clinical utility [20]. Edge wearables must be evaluated in independent cohorts under ambulatory conditions rather than solely on curated datasets.

The data-centric view of deep learning reminds system designers that model architecture is often less important than the quality, diversity, and size of training data [21]. Wearable healthcare suffers from fragmented datasets, inconsistent labeling, and narrow demographic coverage. Edge AI platforms that cannot support continuous improvement through local adaptation risk becoming obsolete as sensor hardware and clinical practice evolve. Federated learning offers one route for updating models without centralizing data, but its fairness depends on the demographic composition of participating devices. A federated network dominated by affluent users may produce a model that performs well for those users and poorly for underserved populations.

6. Governance, Infrastructure, and Sustainability

Governance of edge AI in healthcare extends beyond model accuracy. The World Health Organization guidance on ethics and governance of artificial intelligence for health emphasizes transparency, accountability, and avoidance of health inequities [22]. These principles have concrete architectural consequences. An edge system must log which model version made a decision, preserve the input context for audit, and allow overrides by users and clinicians. Device-level inference without upstream visibility can make adverse events difficult to investigate. Regulatory frameworks such as those for medical device software increasingly require postmarket surveillance and change control. Edge AI architectures must therefore include secure update mechanisms, model provenance, and rollback capabilities. Over-the-air updates introduce security risks, but frozen models may fail to address emerging pathologies or sensor drift.

Sustainability is another governing constraint. Millions of wearable devices with embedded inference engines create material, energy, and e-waste burdens. Aggressive model compression reduces energy per inference but may increase the frequency of retraining cycles. Battery life determines whether continuous monitoring is feasible for chronic care, and frequent recharging disproportionately affects low-income users. Edge AI design should consider lifecycle carbon footprint, battery replacement intervals, and device interoperability.

Standardized data formats and network protocols can extend device life by allowing software updates without hardware replacement. The long-term success of edge AI in wearable healthcare thus depends on aligning clinical safety, algorithmic fairness, and environmental responsibility within a coherent governance framework.

7. Conclusion

Real-time physiological monitoring on resource-constrained wearable devices requires more than the migration of cloud models to embedded processors. It demands a coordinated architectural strategy that spans sensing, local inference, gateway processing, federated learning, and clinical integration. The structural trade-offs among accuracy, latency, energy, privacy, and clinical actionability cannot be resolved by algorithmic innovation alone. Attention-based multimodal signal restoration offers a promising route toward motion-tolerant opto-physiological monitoring, but its deployment on constrained wearables depends on model compression, careful quantization, and hardware-aware scheduling. Equally important are fairness, robustness, and governance. Wearable edge AI systems must be evaluated across diverse populations, integrated into clinical workflows, and subjected to transparent postmarket surveillance. Sustainability and update mechanisms must be treated as first-class design criteria rather than afterthoughts. The future of wearable healthcare lies not in any single model or sensor, but in the careful construction of an edge intelligence infrastructure that is technically efficient, clinically meaningful, ethically accountable, and environmentally responsible.

References

1. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 17(4), 2347–2376.
2. Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. *Sensors*, 17(1), 130.
3. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8), 1738–1762.
4. Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(2), 869–904.
5. Murshed, M. G. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. (2021). Machine learning at the network edge: A survey. *ACM Computing Surveys*, 54(8), 1–35.
6. Baig, M. M., Gholamhosseini, H., Moqem, A. A., Mirza, F., & Lindén, M. (2017). A systematic review of wearable patient monitoring systems: Current challenges and opportunities for clinical adoption. *Journal of Medical Systems*, 41(7), 115.
7. Gravina, R., Alinia, P., Ghasemzadeh, H., & Fortino, G. (2017). Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, 35, 68–80.
8. Islam, S. M. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K.-S. (2015). The Internet of Things for health care: A comprehensive survey. *IEEE Access*, 3, 678–708.

9. Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the internet of things. *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, 13–16.
10. Charlton, P. H., Birrenkott, D. A., Bonnici, T., Pimentel, M. A. F., Johnson, A. E. W., Alastruey, J., Tarassenko, L., Watkinson, P. J., Beale, R., & Clifton, D. A. (2018). Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review. *IEEE Reviews in Biomedical Engineering*, 11, 2–20.
11. Zheng, X., Hu, S., Dwyer, V., Barrett, L., & Derakhshani, M. (2026). Joint attention mechanism learning to facilitate opto-physiological monitoring during physical activity. *Biomedical Signal Processing and Control*, 113, 108949.
12. David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., Regev, S., Rhodes, R., Wang, T., & Warden, P. (2021). TensorFlow Lite Micro: Embedded machine learning for TinyML systems. *Proceedings of Machine Learning and Systems*, 3, 800–811.
13. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704–2713.
14. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations*.
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
16. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.
17. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3, 119.
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
19. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.
20. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
21. Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE International Conference on Computer Vision*, 843–852.
22. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.