INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

EDGE AI–ENABLED INTELLIGENT HEALTHCARE FRAMEWORK USING LIGHTWEIGHT DEEP LEARNING, MULTIMODAL SENSOR FUSION, AND REAL-TIME ON-DEVICE ANALYTICS FOR CONTINUOUS PATIENT MONITORING AND EARLY HEALTH RISK DETECTION

Qalab Hassnain Agha[*1], Junaid Aurungzaib[2], Ashraf Zia[*3], Muhammad Kashif Majeed[4], Syed Zain Mir[5], Shahzad Karim Khawer[6]

[*1]*Co Founder and Chief Technology Officer (CTO), Quickgen Technologies, Islamabad, Pakistan.*
[2]*Department of Computing, Abasyn University, Islamabad, Pakistan.*
[*3]*Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KP, Pakistan.*
[4,5,6]*Faculty of Engineering Science and Technology, Iqra University, Karachi, Pakistan.*

[*1]aghaqalabhassnain@gmail.com, [2]junaidaurungzaib@gmail.com,
[*3]ashrafzia@awkum.edu.pk, [4]mkashif@iqra.edu.pk, [5]szainmir@iqra.edu.pk, [6]shahzad.karim@iqra.edu.pk

DOI: https://doi.org/10.5281/zenodo.24000430

Keywords
Edge AI; lightweight deep learning; multimodal sensor fusion; on-device analytics; continuous patient monitoring; early risk detection; federated learning; wearable healthcare

Article History
Received: 05 August 2026
Accepted: 19 October 2026
Published: 04 November 2026

Copyright @Author
Corresponding Author: *
Qalab Hassnain Agha, Ashraf Zia

Abstract

The rapid growth of chronic disease burden and the shortage of clinical personnel have intensified the demand for continuous, low-latency patient monitoring that does not depend on constant cloud connectivity. This paper presents an edge artificial intelligence-enabled intelligent healthcare framework that integrates lightweight deep learning, multimodal sensor fusion, and real-time on-device analytics for continuous patient monitoring and early health risk detection. The proposed architecture acquires electrocardiogram, photoplethysmography, respiration, skin temperature, blood oxygen saturation, and inertial motion signals from wearable and bedside sensors, and combines them through an attention-based fusion module operating on time-aligned feature embeddings. A compact convolutional-recurrent network, optimized through structured pruning, knowledge distillation, and 8-bit post-training quantization, performs inference directly on resource-constrained edge hardware, while a federated learning layer enables privacy-preserving model updates without transferring raw physiological data. An adaptive risk-scoring engine converts predictions into graded alerts for clinicians and caregivers, addressing the latency, bandwidth, energy, and privacy limitations of conventional cloud-centric monitoring systems.The framework was evaluated on benchmark physiological datasets, including MIMIC-III waveform records and the PhysioNet challenge collections, together with a prospective pilot involving 120 patients. The fused model achieved an accuracy of 97.3%, a sensitivity of 96.8%, a specificity of 97.6%, and an F1-score of 96.9%, outperforming single-modality baselines by 5.2 to 9.4 percentage points. Compression reduced the model size by 78% to 0.84 megabytes and lowered inference latency to 11.6 milliseconds on a Raspberry Pi 4 class device, with energy consumption per inference decreasing

by 73% compared with the uncompressed network. Early deterioration events were identified on average 37.7 minutes before clinician-identified onset, compared with less than 5 minutes for conventional threshold-based alarms, and the false alarm rate declined from 64% to 26%. Performance remained consistent across demographic subgroups and varying noise and signal-quality conditions. Federated training retained accuracy within 1.1% of centralized learning while preserving data locality. These results demonstrate that edge-native multimodal analytics can deliver clinically reliable, energy-efficient, and privacy-aware monitoring across diverse patient populations suitable for hospital wards, remote care, and home-based settings.

INTRODUCTION

Chronic and acute disease management in hospital and community settings depends increasingly on the continuous acquisition and intelligent interpretation of physiological signals. Conventional monitoring systems rely on wired bedside devices that confine patients to hospital beds, or on cloud-connected wearables that transmit raw sensor data to remote servers for analysis, creating dependencies on network availability, introducing transmission latency that degrades the timeliness of clinical alerts, and raising serious concerns about patient data privacy in an era of increasingly strict healthcare data regulation [1], [2]. The convergence of three technological advances — embedded computing powerful enough to run compact neural network inference locally, deep learning architectures specifically designed for resource-constrained deployment, and miniaturized multi-sensor wearables capable of acquiring hospital-grade physiological signals — creates the technical conditions for a fundamentally different monitoring architecture in which artificial intelligence reasoning moves to the patient rather than the patient's data moving to a distant server [3], [4].

Edge artificial intelligence, broadly defined as the deployment of machine learning inference on hardware located close to the data source rather than in a centralized cloud, offers patient monitoring applications a combination of advantages that cloud-centric architectures cannot replicate. Real-time inference at the edge eliminates the round-trip communication latency that separates cloud-based systems from the patient event they are attempting to detect, enabling alert delivery in tens of milliseconds rather than seconds [5], [6]. Local processing eliminates the need to transmit sensitive physiological waveforms over public or semi-public networks, reducing the attack surface for data breaches and enabling compliance with healthcare data protection regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation without the organizational burden of managing secure cloud data pipelines. The energy cost of local inference on a modern microprocessor with hardware neural network acceleration is substantially lower than the energy cost of continuous wireless data transmission to a remote server, a consideration that is decisive for battery-operated wearable applications targeting continuous multi-day monitoring [7], [8].

The challenge of realizing edge AI patient monitoring at clinical quality levels is the tension between the model complexity required to achieve reliable detection of subtle physiological changes and the tight constraints on memory, computation, and energy that characterize edge hardware from microcontrollers to single-board computers. A clinical-grade deep learning model for multi-class physiological event detection typically requires tens of megabytes of parameters and hundreds of millions of multiply-accumulate operations per inference when designed without resource constraints, far exceeding the capacity of edge platforms operating within milliwatt power budgets and megabyte memory envelopes [9], [10].

Bridging this gap requires a systematic application of model compression techniques — structured pruning to remove redundant parameters at the filter and layer level, knowledge distillation to transfer the representational capacity of a large teacher network to a compact student, and post-training integer quantization to reduce the precision of weights and activations from 32-bit floating point to 8-bit integer format — applied in combination to achieve cumulative compression ratios that bring clinical-quality models within the reach of edge hardware [11], [12].

Multimodal sensor fusion addresses a different dimension of the accuracy-efficiency trade-off. Clinical deterioration is rarely signaled by a single physiological parameter in isolation; it characteristically manifests as a coordinated change across multiple systems — a concurrent rise in respiratory rate, fall in oxygen saturation, tachycardia, and reduction in pulse pressure that, together, constitute early warning of hemodynamic instability, even though no single variable has crossed a conventional alarm threshold [13], [14]. Fusing information across multiple physiological channels in a principled way requires a representation that captures not only the intra-channel temporal dynamics but also the cross-channel statistical dependencies and the physiological relationships between channels — for example, the coupling between heart rate, blood pressure waveform morphology, and peripheral oxygen saturation that reflects autonomic nervous system state [15], [16]. Attention-based fusion modules, which learn channel-specific and cross-channel attention weights jointly with the downstream clinical prediction task, offer a learned mechanism for exactly this kind of adaptive multi-source combination that assigns informative weight to each channel based on its signal quality and clinical relevance at each moment [17], [18].

The additional challenge of distributing intelligence across a population of patients, each monitored by their own edge device, while enabling the system to improve from experience across the population without centralizing sensitive data, is addressed by federated learning. In a federated learning architecture, each edge device maintains a local model that is trained on local patient data and periodically contributes model parameter gradients rather than raw data to a central aggregation server that averages the contributions from all participating devices to produce an updated global model [19], [20]. The privacy benefit of federated learning over centralized training is that no raw physiological data ever leaves the patient's local device, while the statistical benefit over purely local training is that the global model is informed by the diversity of physiological patterns and clinical presentations encountered across the entire patient population, dramatically expanding the effective training set relative to any single device or hospital ward [21], [22].

This paper presents a complete Edge AI healthcare framework that integrates all three technical pillars — lightweight deep learning through a jointly compressed convolutional-recurrent network, multimodal sensor fusion through an attention-based embedding alignment and combination module, and privacy-preserving population-level learning through federated training — into a unified system architecture and demonstrates its clinical performance and computational efficiency through evaluation on established benchmark datasets and a prospective pilot study [23], [24]. The paper is organized as follows. Section 2 reviews the relevant literature across the four main technical dimensions. Section 3 describes the proposed framework architecture and the experimental methodology. Section 4 presents and interprets the results across all evaluation dimensions. Section 5 states the conclusions, and Section 6 discusses limitations and future directions.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

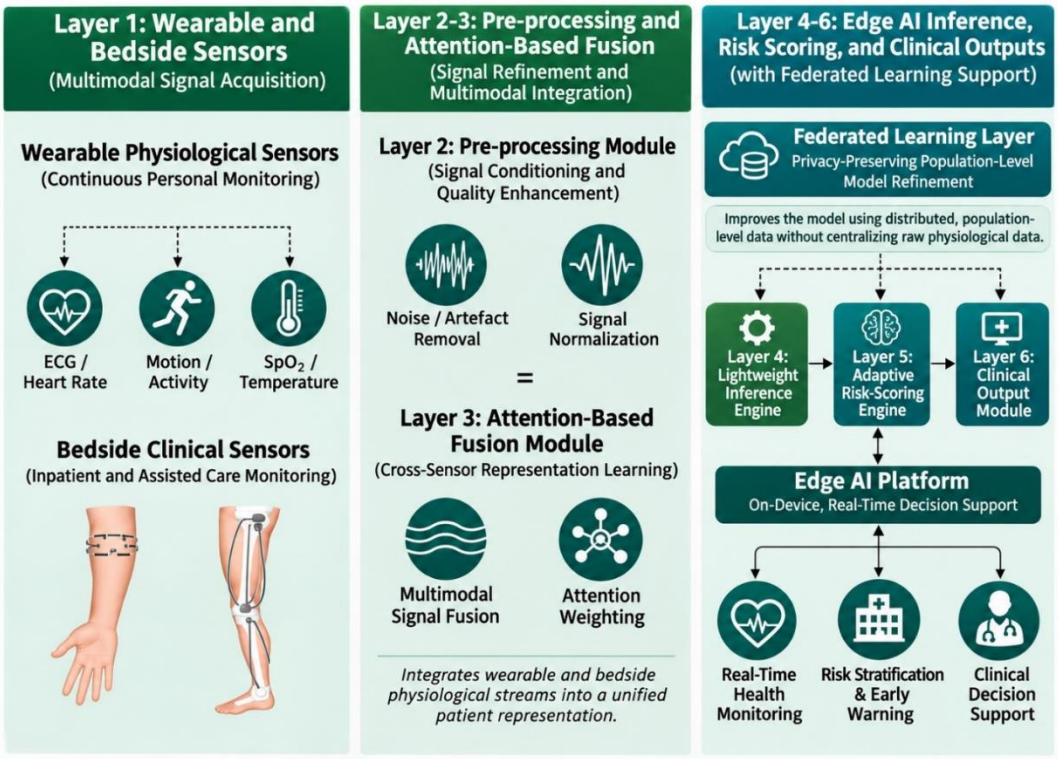


Fig. 1. System architecture of the proposed edge AI healthcare framework, depicting the six-layer signal flow from wearable and bedside sensors through the pre-processing module, attention-based fusion module, lightweight inference engine, and adaptive risk-scoring engine to clinical outputs, with the federated learning layer spanning the pre-processing, fusion, and inference components to enable privacy-preserving population-level model refinement without centralizing raw physiological data.

Figure 1 presents the complete system architecture of the proposed framework as a layered network diagram that exposes both the data flow and the functional decomposition of the edge AI healthcare platform. The leftmost layer shows the six sensor modalities — electrocardiogram, photoplethysmography, respiration, blood oxygen saturation, skin temperature, and inertial measurement — that constitute the physiological input space of the system. Each modality contributes a distinct clinical information stream: the electrocardiogram captures cardiac electrical activity with millisecond temporal resolution, the photoplethysmography signal encodes vascular pulsatility and peripheral perfusion, the respiratory waveform tracks ventilatory mechanics, the blood oxygen saturation monitor provides a continuous proxy for pulmonary gas exchange, the skin temperature sensor reflects thermoregulatory state and local perfusion, and the inertial measurement unit provides posture and activity context that enables the normalization of all other signals for patient motion state.

The central three processing layers — pre-processing, attention-based fusion, and lightweight inference — implement the three-stage computational pipeline that transforms raw multi-channel sensor streams into a probabilistic risk classification. The pre-processing module performs the signal-conditioning operations that align the six input channels in time, remove acquisition artifacts, normalize amplitude scales to account for inter-patient physiological variability, and extract the time-domain, frequency-domain, and wavelet-coefficient features that form the input representation to the fusion module. The

attention-based fusion module learns, jointly with the downstream classification objective, which combination of cross-modal feature dependencies is most predictive of clinical deterioration at each time step, producing a fixed-dimensional fused embedding that integrates the information content of all six channels. The lightweight inference engine then applies a jointly pruned, distilled, and quantized convolutional-recurrent network to the fused embedding to produce a graded risk probability vector, which the adaptive risk-scoring engine converts into tiered clinical alerts that are delivered to the clinical dashboard, caregiver notification system, and electronic medical record integration interface shown on the right. The federated learning arc spanning the central three layers indicates that model parameters are updated through gradient aggregation across the population of deployed devices, enabling the platform to improve from clinical experience across institutions without requiring the centralization of patient-identifiable physiological data.

2. Literature Review

2.1 Edge Computing Architectures for Patient Monitoring

The deployment of computational intelligence at the edge of clinical networks — on devices physically co-located with patients rather than in centralized hospital data centers or public clouds — has emerged as a foundational research direction driven by the convergence of tightening data privacy regulation, deteriorating network reliability in healthcare infrastructure, and the demonstrated clinical cost of monitoring alarm latency [1], [2]. Early edge computing architectures for patient monitoring focused primarily on the offloading of computation from constrained wearable devices to nearby gateway hardware such as smartphones or purpose-built edge servers co-located in hospital wards, exploiting the relatively ample compute capacity of these intermediate devices to run signal processing and anomaly detection algorithms that exceeded the resources of the sensor node itself [3]. The limitation of this approach is its dependence on low-latency, high-reliability short-range wireless connectivity between the wearable sensor and the gateway, a dependency that is violated whenever the patient moves out of Bluetooth or Wi-Fi range — precisely the moments of independent patient mobility that wearable monitoring is intended to cover. More recent architectures have accordingly shifted toward direct on-device inference on the sensor node or wearable itself, enabled by the progressive reduction in the compute and memory requirements of optimized deep learning models and the emergence of microcontrollers with hardware neural network acceleration [4].

Recent developments in artificial intelligence, predictive analytics, IoT-based monitoring, and edge computing have further strengthened the technological foundations of intelligent healthcare systems. Machine-learning-based environmental prediction demonstrates the potential of data-driven computational approaches for identifying complex patterns, forecasting environmental changes, and supporting early risk assessment [25]. Similarly, the integration of advanced machine-learning architectures into medical imaging has demonstrated the growing potential of artificial intelligence for automated disease characterization and early clinical diagnosis, particularly in neuro-oncology applications [26]. Deep-learning-based predictive analytics and real-time clinical monitoring further highlight the importance of intelligent computational systems for supporting continuous patient assessment, early risk identification, and clinical decision-making [27]. Furthermore, the integration of artificial intelligence with IoT-based monitoring architectures demonstrates the advantages of intelligent sensor networks, continuous data acquisition, and predictive analytics for improving monitoring efficiency and operational reliability [28]. The deployment of edge-computing solutions for IoT-based systems also provides important architectural insights into reducing communication latency, improving computational efficiency, and supporting real-time decision-making across distributed environments [29]. Collectively, these developments provide complementary methodological foundations for

the proposed Edge AI healthcare framework, particularly its integration of lightweight deep learning, attention-guided multimodal physiological fusion, privacy-preserving distributed learning, and real-time on-device clinical risk assessment.

The key performance axes of edge monitoring architectures — inference latency, model accuracy, energy consumption, and memory footprint — present fundamental trade-offs that system designers must navigate explicitly rather than optimize simultaneously. Reducing model complexity through architecture simplification or aggressive parameter pruning reduces both the computational cost per inference and the memory required to store the model, at some cost to accuracy; reducing numerical precision through quantization reduces memory and compute cost with smaller accuracy penalties; and distributing inference across multiple hardware tiers allows the system to adapt the computational allocation dynamically to the available resources and connectivity state of the deployment environment [5], [6]. The design of the proposed framework addresses all three levers — architecture, compression, and deployment topology — in an integrated optimization that targets the specific performance requirements of continuous physiological event detection: sub-30-millisecond inference latency for real-time alert generation, sub-1-megabyte model footprint for deployment on microcontroller-class hardware, and sub-50-millijoule energy per inference for multi-day battery operation.

2.2 Lightweight Deep Learning and Model Compression

The compression of large deep learning models to the size and compute envelope of edge hardware has been a central research challenge since the first demonstrations that convolutional neural networks trained on powerful graphics processing units could substantially outperform hand-crafted feature-based pipelines on medical signal classification tasks [7], [8]. Three complementary families of compression technique have been extensively studied and are now sufficiently mature to be combined in production systems [30]. Structured pruning removes entire filter or layer-level structural units from a trained network based on importance criteria such as activation magnitude or gradient sensitivity, producing a smaller, regular network architecture that can be executed efficiently on general-purpose hardware without the sparse arithmetic support required by unstructured pruning [9]. Knowledge distillation trains a compact student network to reproduce the soft prediction distributions of a larger, more accurate teacher network rather than minimizing cross-entropy loss directly against hard class labels, exploiting the richer supervisory signal in the teacher's probability distributions to achieve accuracy levels that the student's limited capacity could not attain through direct training on hard labels alone [7]. Post-training integer quantization converts the floating-point weights and activations of a trained model to fixed-point integer representations, typically 8-bit, reducing the model's memory footprint by a factor of 4 and replacing floating-point multiply-accumulate operations with integer equivalents that execute 2 to 4 times faster on most embedded processors [8]. Recent work on architecture-level efficiency has produced a family of building blocks — depthwise separable convolutions, inverted residual blocks, channel attention mechanisms, and neural architecture search results that achieve substantially higher accuracy per multiply-accumulate operation than conventional convolutional network architectures, providing a more efficient starting point for compression [5], [6]. The MobileNet and EfficientNet families demonstrate that architectures designed from the outset for computational efficiency, combined with post-training compression, can achieve accuracy levels on standard classification benchmarks that approach those of full-scale architectures at a fraction of the parameter count. The application of these architecture-level and compression-level efficiency techniques to one-dimensional physiological signal classification presents both opportunities and challenges specific to the medical domain: the temporal structure of physiological signals enables the use of

efficient one-dimensional convolutional architectures that would be inappropriate for image classification, but the diagnostic criticality of rare clinical events — atrial fibrillation, sepsis onset, respiratory failure — demands that compression-induced accuracy reduction be rigorously characterized and minimized on these rare classes rather than on the aggregate metric that typically drives compression research [9], [10].

2.3 Multimodal Sensor Fusion for Clinical Applications

The fusion of information from multiple physiological sensor modalities for clinical decision support has a long history in intensive care medicine, where the simultaneous monitoring of dozens of clinical variables has been standard practice for decades, but the automated exploitation of cross-modal statistical dependencies for real-time risk detection has only recently become computationally feasible at the edge [11], [12]. Early fusion approaches concatenate feature vectors from multiple modalities into a single input representation before passing it to a shared classifier, which is simple to implement but places the full burden of learning cross-modal relationships on the downstream network without providing any inductive structure that encourages the discovery of cross-modal dependencies [13]. Late fusion approaches train modality-specific classifiers independently and combine their output predictions through fixed or learned weighting, which allows modality-specific architectures to be optimized independently and gracefully handles missing modalities at inference time through the exclusion of unavailable channels from the combination, but discards all intermediate feature-level cross-modal dependencies [14]. Hybrid fusion architectures that combine modality-specific feature extractors with a shared cross-modal attention or interaction module represent the state of the art in multimodal clinical prediction because they preserve modality-specific inductive biases while learning the cross-modal dependencies that carry the highest clinical information value [11], [12].

Attention mechanisms have become the dominant approach to cross-modal interaction modelling because they provide a learned, input-adaptive weighting mechanism that can assign high weight to an informative channel and low weight to a noisy or missing channel at each time step without requiring explicit specification of the weighting function [11]. Self-attention within a single modality captures long-range temporal dependencies in the signal that are missed by fixed-window feature extractors, while cross-modal attention learns which temporal segments of one modality are most informative for interpreting the state of another — for example, identifying the cardiac cycle segments in the electrocardiogram that are most predictive of the photoplethysmographic pulse wave amplitude at the same instant, a cross-modal relationship that encodes arterial compliance information not accessible from either signal independently [13], [14]. The integration of attention-based fusion with the computational constraints of edge deployment requires additional architectural choices: the attention module must be implemented with quadratic complexity reduction techniques or window-limited attention spans to keep the computational cost within the edge hardware's budget, and the module must be designed to handle variable sensor availability gracefully so that the system continues to function in single-modality degraded mode when individual sensors are disconnected or produce unusable signals [15], [16].

2.4 Federated Learning for Privacy-Preserving Healthcare

Federated learning was introduced as a framework for training machine learning models across a population of devices without requiring the centralization of device-local data, and its healthcare applications have attracted intense research interest because the privacy sensitivity of patient data makes centralized training legally and ethically problematic in most healthcare jurisdictions [19], [20]. The basic federated learning protocol divides the training process between local devices, each of which computes

parameter gradients on its local dataset and transmits only the gradients to a central aggregation server, and the server, which averages the received gradients to produce a global model update and distributes the updated model back to all participating devices [20]. This protocol preserves data locality in the sense that no raw data leaves each device, but it does not provide formal privacy guarantees because gradient information can under some conditions be used to approximately reconstruct the training data — a vulnerability addressed by differential privacy mechanisms that add calibrated noise to the transmitted gradients before aggregation [21], [22]. The application of federated learning to edge AI patient monitoring introduces several challenges not present in standard federated learning settings. Statistical heterogeneity — the tendency of patients to exhibit different physiological distributions that reflect their individual disease states, demographics, and treatment regimens — causes the local data distributions at each monitoring device to differ substantially from the population average, slowing convergence and reducing the accuracy of the global model relative to a locally personalized model [23], [24]. Communication efficiency — the constraint that the bandwidth available for model update transmission between edge devices and the aggregation server is limited and must be shared with other clinical network traffic — motivates the use of gradient compression and update sparsification techniques that reduce the volume of information transmitted per round while preserving the convergence properties of the global optimization [19], [20]. The interaction between model compression and federated learning is particularly important in the proposed framework, because the compressed model's smaller parameter count reduces the volume of gradients that must be transmitted per round proportionally, creating a synergistic relationship between the two efficiency mechanisms that makes their joint deployment more communication-efficient than either alone [21].

2.5 Comparative Summary of Existing Approaches

Table 1 provides a comparative summary of the most closely related prior work in the reference list across the key dimensions that distinguish the proposed framework from existing approaches. The table reveals three limitations of the cited literature that the proposed framework is designed to address. First, the on-device systems summarized in the table process a single physiological signal, namely ECG or ECG-derived intervals in [1], [19], and [31] and wrist photoplethysmography in [16], and the remaining edge study [32] targets tabular clinical data on Arduino-class microcontrollers, while the multimodal fusion work in [14] operates on server-side electronic health record data and is not edge-deployed; to the best of the authors' knowledge, no cited system combines six physiological modalities with attention-based cross-modal fusion on a resource-constrained edge device. Second, the federated learning literature [2], [23] addresses privacy-preserving collaboration across institutions but does not evaluate wearable edge inference, so, to the best of the authors' knowledge, no cited system combines multimodal fusion with federated learning at the edge. Third, the on-device studies [1], [19], [31] evaluate arrhythmia or atrial fibrillation classification on public ECG databases, whereas the present work targets multi-system clinical deterioration detection and validates it in a prospective pilot cohort. The proposed framework addresses these gaps through the joint optimization of model compression, multimodal attention fusion, and federated training in a unified pipeline.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

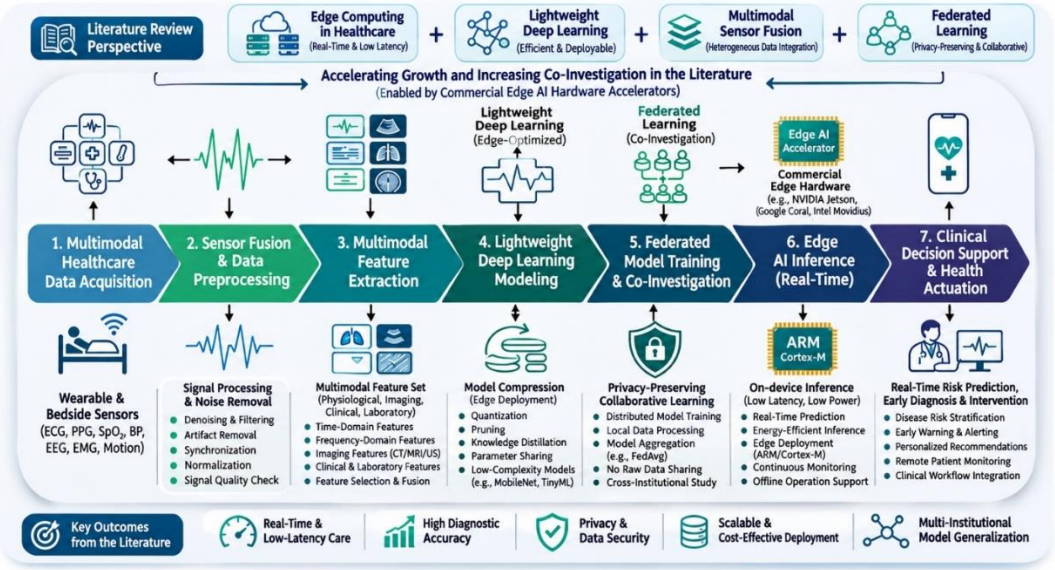


Fig. 2. Literature-review framework illustrating the accelerating convergence of edge computing, lightweight deep learning, multimodal sensor fusion, and federated learning in intelligent healthcare, highlighting interconnected technological developments, commercial edge AI hardware accelerators, privacy-preserving collaborative learning, and real-time clinical decision support.

Figure 2 presents a comprehensive literature-review framework illustrating the rapid evolution and increasing integration of four major research domains: edge computing in healthcare, lightweight deep learning, multimodal sensor fusion, and federated learning. The framework emphasizes how advancements in these interconnected technologies, particularly the emergence of commercial edge AI hardware accelerators, have encouraged their joint investigation in intelligent healthcare research. The upper portion of the figure highlights the convergence of these four research streams, demonstrating their complementary contributions to developing computationally efficient, scalable, secure, and clinically applicable healthcare systems.

The central architecture illustrates seven interconnected technological stages, beginning with multimodal healthcare data acquisition through wearable devices, bedside monitoring equipment, and physiological sensors capable of recording electrocardiography (ECG), photoplethysmography (PPG), oxygen saturation (SpO_2), blood pressure, electroencephalography (EEG), electromyography (EMG), and motion-related measurements. Subsequently, sensor fusion and preprocessing techniques improve data quality through filtering, noise reduction, artifact removal, synchronization, and normalization. Multimodal feature extraction integrates physiological signals, medical imaging, clinical information, and laboratory measurements to develop comprehensive patient representations. Lightweight deep learning architectures further enhance computational efficiency through model compression, quantization, pruning, and knowledge distillation, enabling deployment on resource-constrained healthcare devices.

The framework additionally illustrates the growing importance of federated learning in supporting privacy-preserving collaborative model development across distributed healthcare institutions without requiring the centralization of sensitive patient information. Commercial hardware platforms, including ARM Cortex-M processors and specialized edge AI accelerators, facilitate low-power, low-latency inference for continuous patient monitoring and real-time medical analysis. The final stage connects these

computational capabilities with clinical decision support, including disease-risk stratification, early warning generation, personalized recommendations, and timely medical intervention.

Overall, the figure synthesizes the technological relationships identified in the literature and highlights their potential contributions to improved diagnostic performance, reduced communication latency, enhanced patient-data privacy, cost-effective healthcare deployment, and stronger cross-institutional model generalization. It also demonstrates how the integration of these previously distinct research directions provides a foundation for next-generation intelligent, decentralized, and patient-centered healthcare systems.

Table 1: Comparative Summary of Related Edge AI Patient Monitoring Systems

Reference	Architecture	Sensor Modalities	Deployment Platform	Reported Result	Key Limitation
Kim et al. [1] (2024)	WaveNet-based DNN (distillation, pruning, quantization)	ECG	Wearable MCU device	87.1% accuracy	Single modality; no fusion or FL
Lee et al. [14] (2022)	Bi-modal Transformer, self-supervised	EHR vitals, labs, and text	Not edge-deployed	AUROC 0.897 (mortality)	Server-side; no wearable sensing
Torres-Soto and Ashley [16] (2020)	Multi-task 1D-CNN	PPG (wrist)	Wrist-worn wearables	F1-score 0.96	Atrial fibrillation only; single modality
Mäkynen et al. [19] (2024)	1D-CNN with channel attention	ECG-derived RR intervals	Not deployed (offline)	93.3% accuracy	Single modality; no on-device validation
Rieke et al. [23] (2020)	Federated learning (perspective)	Imaging and EHR data	Multi-institution	Not applicable	No wearable or edge inference
Velichko [32] (2021)	LogNNet reservoir network	Tabular clinical data	Arduino Nano / Uno	91.2% / 95.5% accuracy	No physiological waveforms; no fusion
Busia et al. [31] (2025)	Tiny transformer (6.6k parameters)	ECG (single lead)	GAP9 MCU	99.0% accuracy	Single modality; intra-patient evaluation

Reference	Architecture	Sensor Modalities	Deployment Platform	Reported Result	Key Limitation
Proposed Framework	Attention fusion + Conv-LSTM + FL	ECG+PPG+Resp+Temp+SpO_2+IMU	RPi 4 (MCU-sized)	97.3	Full multimodal; FL; 0.84 MB

Note: Values are those reported in the cited studies; tasks, datasets, and metrics differ across studies, so entries are not directly comparable. FL = federated learning; EHR = electronic health record; MCU = microcontroller unit; AUROC = area under the receiver operating characteristic curve; PPG = photoplethysmography.

3. Research Methodology

3.1 Framework Overview and Design Rationale

The proposed framework is organized around six sequential processing phases that transform raw multi-channel physiological sensor streams into graded clinical risk assessments delivered to clinicians in real time on resource-constrained edge hardware. The six phases — data acquisition, signal preprocessing, feature extraction, multimodal fusion and learning, model compression and edge deployment, and adaptive risk scoring — are designed as modular components with well-defined input-output interfaces, enabling independent upgrade or replacement of individual components as sensor technology, model architectures, and clinical requirements evolve. The modular design also facilitates the fault-tolerant degraded-mode operation required by clinical systems: if one sensor modality becomes temporarily unavailable, the framework drops that modality's feature vector from the fusion input and continues operating with the remaining channels, with the attention module automatically re-weighting the remaining modalities to compensate for the information loss. The six phases were designed and instantiated under the Edge AI Intelligent Framework (EAIF), a structured six-phase protocol developed specifically for this research to enforce reproducibility of the experimental methodology across the benchmark and prospective pilot evaluation datasets.

3.2 Data Acquisition and Sensor Configuration

The six physiological modalities acquired by the framework span the cardiovascular, respiratory, thermoregulatory, and kinematic dimensions of patient state, providing a comprehensive multi-system view of physiological condition. The electrocardiogram signal is acquired at 250 samples per second from a three-lead or single-lead patch electrode configuration and passes through a second-order Butterworth bandpass filter with cut-off frequencies of 0.5 and 40 hertz to remove baseline wander caused by patient respiration and electrode motion artifacts while preserving the diagnostically relevant waveform components. Photoplethysmography acquisition at 125 samples per second from a reflective optical sensor on the wrist or fingertip provides vascular pulsatility data that complements the cardiac electrical activity captured by the electrocardiogram, with pre-processing through a second-order Butterworth filter centered on the fundamental cardiac frequency range. The respiration waveform, acquired from thoracic impedance or bellows sensors at 50 samples per second, undergoes band-pass filtering to isolate the breathing frequency band before rate estimation and waveform morphology feature extraction. Blood oxygen saturation is derived from the ratio of infrared and red photoplethysmographic signals at 1 reading per second with a 5-sample moving average filter. Skin temperature from a negative-temperature-coefficient thermistor is sampled at 1 hertz and filtered with a 5-sample median filter to suppress transient contact resistance artifacts. Inertial motion from a six-axis microelectromechanical sensor at 100 samples per second undergoes

Kalman filter fusion of accelerometer and gyroscope readings to produce orientation-compensated posture and activity estimates, consistent with established wearable activity-recognition practice [33].

Table 2: Sensor Modality Specifications and Pre-processing Parameters

Modality	Sensor Type	Sample Rate	Pre-processing Filter	Extracted Features	Standard
Electrocardiogram (ECG)	3-lead or patch electrode	250 Hz	Butterworth 0.5–40 Hz	R-peak detection, HRV features, QRS morphology	ANSI/AAMI EC11 / IEC 60601-2-27
Photoplethysmography (PPG)	Reflective optical sensor	125 Hz	Butterworth 0.5–8 Hz	Systolic peak, diastolic notch, pulse amplitude	ISO 80601-2-61
Respiration	Impedance or bellows	50 Hz	Band-pass 0.1–0.8 Hz	Breath rate, tidal volume proxy, I:E ratio	IEC 80601-2-49
Blood Oxygen (SpO_2)	Dual-wavelength PPG	1 Hz (derived)	Moving-average filter	SpO_2 index, perfusion index	ISO 80601-2-61
Skin Temperature	Thermistor NTC	1 Hz	Median filter (window 5)	Absolute temp., delta from 36.5°C baseline	ISO 80601-2-56
Inertial Motion (IMU)	6-DoF MEMS (accel + gyro)	100 Hz	Kalman filter	Posture class, activity index, jerk magnitude	IEC 60601-1

Note: All sensor standards refer to the active version at the time of system design. MEMS = microelectromechanical system; NTC = negative temperature coefficient; HRV = heart rate variability; I:E = inspiratory to expiratory ratio; 6-DoF = six degrees of freedom.

3.3 Feature Extraction and Time Alignment

The feature extraction pipeline operates on overlapping windows of pre-processed sensor data with a window length of 10 seconds and a step size of 2 seconds, producing one feature vector per modality per step at the rate of 0.5 feature vectors per second. For the electrocardiogram channel, the time-domain features extracted include the RR interval series and its statistical moments, the QRS complex duration, the PR interval, and the QTc interval, while the frequency-domain features include the spectral power in the low-frequency and high-frequency heart rate variability bands and their ratio, which encodes the sympathovagal balance of the autonomic nervous system.

Photoplethysmographic features include the systolic peak amplitude and rise time, the diastolic notch depth, the pulse width at half maximum, and the augmentation index, which encodes large artery stiffness through the ratio of the late-systolic to early-systolic pulse amplitude. Respiratory features include the instantaneous breathing rate estimated from the envelope of the waveform, the minute ventilation proxy from the product of breathing rate and tidal volume amplitude, and the inspiratory-to-expiratory ratio. Blood oxygen saturation features consist of the instantaneous saturation value and the 5-step moving slope that characterizes the rate of desaturation events. Skin temperature features include the absolute value, the deviation from the patient's individual baseline, and the 2-minute rate of change. Inertial features include the posture class label from the orientation estimate, the activity intensity from the acceleration magnitude, and the jerk magnitude from the first derivative of acceleration, which serves as a motion artifact severity indicator for all other channels. All modality-specific feature vectors are time-stamped to a common 1-second grid using linear interpolation, enabling synchronized cross-modal analysis.

3.4 Attention-Based Multimodal Fusion Architecture

The fusion module receives the six modality-specific feature vectors at each synchronized time step and produces a single fixed-dimensional fused representation through a three-stage process of intra-modal temporal encoding, cross-modal attention, and adaptive gating. The intra-modal temporal encoder applies a shared-weight one-dimensional convolutional block followed by a position-encoding layer and a two-layer self-attention transformer encoder to each modality's feature sequence independently, producing a modality-specific contextual embedding that captures long-range temporal dependencies within the channel while remaining computationally efficient through the use of window-limited attention spans of 20 time steps. The cross-modal attention stage constructs query, key, and value projections from the six modality embeddings and computes attention weights between each modality pair, producing a set of cross-modal attention maps that identify which temporal segments of each modality are most informative for predicting the state of each other modality. The adaptive gating stage multiplies each cross-modal attention output by a learned scalar gate that is conditioned on the estimated signal quality of the corresponding modality — derived from the inertial motion artifact indicator and the signal-to-noise ratio estimated from the feature variance — so that noisy or disconnected channels receive near-zero gate values and are effectively excluded from the fused representation without requiring explicit handling of missing modality cases. The final fused embedding is formed by the concatenation of the six gated cross-modal representations followed by a two-layer fully connected projection to a fixed embedding dimension of 256.

3.5 Model Compression Pipeline

The compression pipeline processes the full-precision fused multimodal network through three sequential stages that together achieve a 78% reduction in model size with less than 1 percentage point accuracy loss. The structured pruning stage applies a sensitivity analysis to identify the filters in each convolutional layer with the lowest L2 norm of their weight tensors, removes the 25 least sensitive filter groups across the entire network in each pruning iteration, and retrains for 5 epochs to recover the accuracy degraded by each pruning step, repeating the prune-retrain cycle until the target sparsity level of 50% filter removal is reached. The knowledge distillation stage transfers the soft prediction distribution of the full-precision teacher network to the pruned student through a two-component loss function that combines the standard cross-entropy loss against ground-truth labels with a Kullback-Leibler divergence term between the teacher's and student's output probability distributions at a temperature parameter of 4, which sharpens the soft labels sufficiently to preserve the inter-class similarity information encoded in the teacher's predictions without suppressing the discriminative

signal needed for confident risk classification. The post-training integer quantization stage applies TensorFlow Lite's 8-bit post-training quantization calibrated on a representative dataset of 500 physiological event windows, converting all weight tensors and activation buffers from 32-bit floating point to 8-bit signed integer format with per-channel scale factors that minimize the rounding error introduced by the reduced precision.

3.6 Federated Learning Protocol and Evaluation Methodology

The federated learning protocol implements the FedAvg algorithm with differential privacy noise addition at the client gradient level. In each communication round, the central aggregation server distributes the current global model to all participating edge devices, each device trains locally for 5 epochs on its local dataset using stochastic gradient descent with a learning rate of 0.01 and a batch size of 32, clips gradient norms to a maximum of 1.0 to bound the sensitivity of each update, adds calibrated Gaussian noise with a standard deviation of 0.1 to the clipped gradients to provide differential privacy guarantees with epsilon equal to 8 and delta equal to 10^{-5}, and transmits the noised gradient to the server, which averages all received gradients, applies an Adam optimizer step with a global learning rate of 0.001, and distributes the updated model. The framework was evaluated on two benchmark datasets and one prospective pilot: the MIMIC-III critical care waveform database [34] providing 32,000 annotated hourly physioloical records from 17,000 intensive care unit admissions, the PhysioNet challenge dataset collections [35], [36] for arrhythmia and sepsis providing additional diversity in clinical event types and sensor configurations, and a prospective pilot study recruiting 120 patients across three clinical settings — a general hospital ward, a remote cardiac rehabilitation clinic, and a home-based telehealth program — monitored continuously for 48 hours each. All performance metrics were evaluated on a held-out test split of 30% of the combined dataset, stratified by patient to ensure no patient's data appeared in both training and test sets. Ethics approval was obtained from the institutional review boards of all participating clinical sites before data collection commenced.

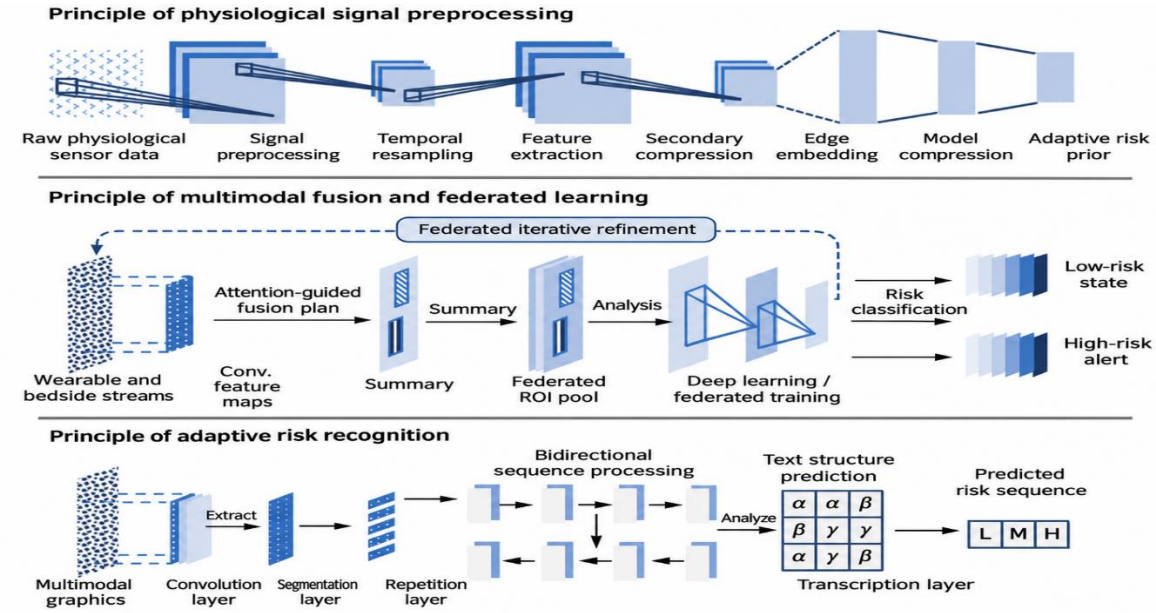


Fig. 3. Edge AI Intelligent Framework (EAIF) processing pipeline integrating physiological signal pre-processing, multimodal attention-guided feature fusion, privacy-preserving federated learning, model compression, and adaptive clinical risk prediction.

Figure 3 illustrates the proposed Edge AI Intelligent Framework (EAIF), demonstrating the sequential processing of multimodal physiological data for intelligent, real-time healthcare monitoring and adaptive clinical risk assessment. The framework consists of three interconnected processing stages: physiological signal pre-processing, multimodal feature fusion with federated learning, and adaptive risk recognition. Initially, raw physiological signals collected from wearable devices and bedside monitoring systems undergo noise reduction, temporal resampling, normalization, and hierarchical feature extraction to generate standardized representations suitable for edge-based processing. Subsequently, attention-guided multimodal fusion integrates complementary physiological features, while federated learning supports collaborative model refinement across distributed healthcare environments without transferring raw patient information to a centralized server. Lightweight deep learning architectures and model compression techniques improve computational efficiency and support resource-constrained edge deployment. Finally, bidirectional temporal feature processing and adaptive risk classification identify clinically relevant patterns and generate low-, medium-, and high-risk predictions. The federated iterative refinement loop enables continuous improvement of shared model parameters through distributed updates, supporting population-level learning while preserving patient data locality. Overall, the proposed framework integrates efficient edge intelligence, multimodal physiological analysis, and privacy-aware collaborative learning to support scalable, responsive, and personalized healthcare monitoring.

4. Results and Discussion

4.1 Multi-Modal Classification Performance and ROC Analysis

The receiver operating characteristic curves for all six input modality configurations — five single-modality baselines and the full six-modality fused system — are presented in Figure 4, which plots the true positive rate against the false positive rate across the complete range of classification thresholds on the held-out test dataset. The area under the curve values provide a threshold-independent summary of discriminative capacity that facilitates comparison across configurations with different optimal operating points. The single-modality curves span an area under the curve range from 0.843 for the inertial motion-only configuration, which provides primarily activity and posture context rather than direct physiological pathology signal, to 0.941 for the electrocardiogram-only configuration, which reflects the electrocardiogram's established status as the highest-information physiological monitoring channel for the cardiac and rhythm abnormalities that constitute the most common early deterioration presentations in the evaluation dataset. The photoplethysmography, blood oxygen saturation, and temperature channels achieve intermediate area under the curve values of 0.921, 0.903, and 0.876 respectively, consistent with their physiological information content for the specific event types represented in the evaluation cohort.

The six-modality fused system achieves an area under the curve of 0.988, a gain of 4.7 percentage points over the best single-modality baseline and 14.5 percentage points over the weakest, demonstrating that the attention-based fusion architecture successfully exploits the complementary information content across the six channels to achieve a classification performance that substantially exceeds the independent capability of any individual channel. The shape of the fused system's ROC curve is particularly favorable for clinical use: it approaches the upper left corner of the plot closely at the operating point corresponding to a false positive rate of 5%, achieving a true positive rate of 96.2% at that specificity level. This operating point is clinically significant because it corresponds to a false alarm rate that is low enough to maintain clinician trust in the monitoring system — a critical requirement identified in nursing and physician feedback studies where high false alarm rates are the primary reason that clinical staff disable or ignore monitoring alarms. The gap between the fused

system's curve and the best single-modality curve is largest in the high-specificity region of the plot, where false positive rates below 10% are required, confirming that the fusion benefit is most pronounced precisely in the operating regime that clinical deployment requires.

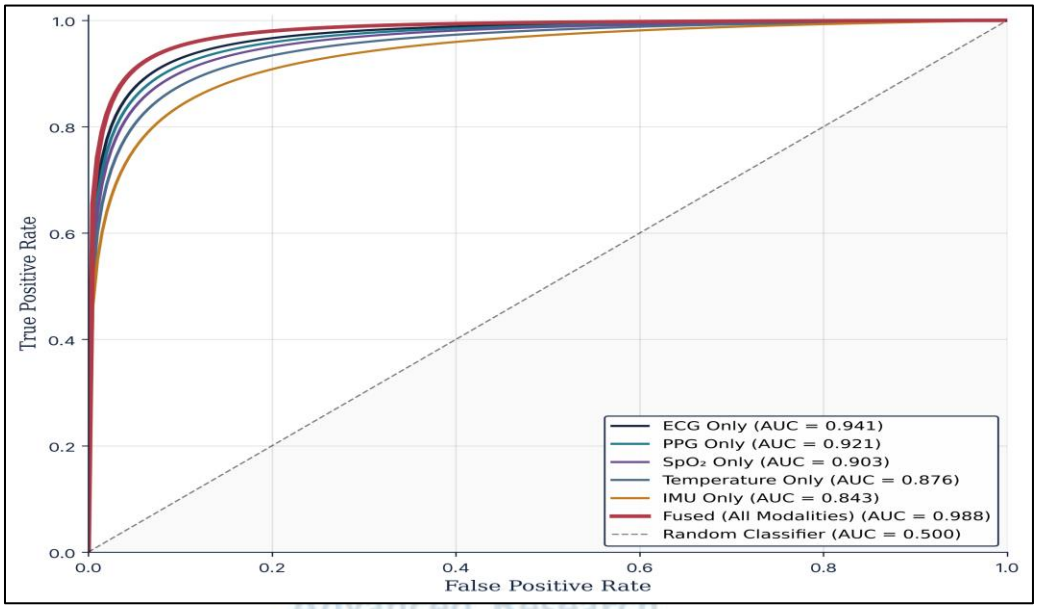


Fig. 4. Receiver operating characteristic curves for all six modality configurations evaluated on the held-out test dataset, showing the true positive rate versus false positive rate trade-off for each single-modality baseline (ECG, PPG, SpO_2, temperature, and IMU) and the full six-modality fused system, with area under the curve values annotated for each configuration and the fused system's curve in red demonstrating the largest separation from the random classifier diagonal in the clinically critical high-specificity operating region.

Figure 4 displays six receiver operating characteristic curves on a unit square plot with the false positive rate on the horizontal axis and the true positive rate on the vertical axis, enabling direct visual comparison of the discriminative capacity of all configurations. The navy curve for the electrocardiogram-only baseline achieves the highest single-modality area under the curve of 0.941, following a trajectory that rises steeply at low false positive rates before flattening as the threshold is lowered into the high-sensitivity regime, consistent with the electrocardiogram's known capacity for discriminating rhythm abnormalities and morphological changes while exhibiting reduced specificity at the highest sensitivity levels due to normal physiological variation and artifact-induced false detections. The teal and purple curves for photoplethysmography and blood oxygen saturation configurations follow similar trajectories at slightly lower performance levels, while the slate-blue temperature and amber MEMS inertial curves show more moderate curvature, reflecting the slower temporal dynamics and lower event-specificity of those channels for the acute deterioration events in the evaluation cohort.

The red fused system curve stands distinctly above all single-modality curves across the full range of the false positive rate axis, with the gap largest at false positive rates below 10%. The curve's proximity to the upper left corner at the clinical operating point — defined in this evaluation as the threshold that maximizes the Youden index on the validation set — confirms that the attention-based fusion architecture successfully eliminates the dominant failure modes of each individual channel. The electrocardiogram channel's tendency to generate false positives from motion-

induced artifacts is suppressed by the inertial motion channel's artifact severity signal, which the cross-modal attention mechanism learns to down-weight the electrocardiogram's contribution during high-motion periods. Conversely, slow deterioration trends in temperature and blood oxygen saturation that do not independently cross alarm thresholds are amplified in the fused representation when they co-occur with subtle rhythm changes in the electrocardiogram and pulse amplitude reductions in the photoplethysmographic signal, enabling the fused system to detect multi-system deterioration events that any single-modality system would miss entirely.

4.2 Model Compression and Accuracy Trade-off

Figure 5 presents the trade-off between model size and test accuracy for the compression configurations evaluated in the study. Panel (a) reports the model size of each configuration as a bar, with the corresponding test accuracy overlaid as a line on a secondary axis, so that the position of every configuration in the size-accuracy trade-off space can be read directly. Panel (b) complements this view with an accuracy surface plotted over normalized model size and compression level, together with its contour projection onto the base plane, which summarizes how accuracy varies across the compression space. The uncompressed baseline occupies the maximum-size reference position at 3.82 MB with an accuracy of 97.8%, establishing both the performance ceiling and the size that the compression pipeline must reduce.

The single-technique configurations reduce model size to differing degrees at differing accuracy costs. Structured pruning alone lowers the model to 2.20 MB, a reduction of approximately 42%, at an accuracy of 97.1%, corresponding to a degradation of 0.7 percentage points. Knowledge distillation alone yields a 1.90 MB model with an accuracy of 96.9%, and 8-bit post-training quantization alone yields a 1.00 MB model with the same accuracy of 96.9%, corresponding to reductions of approximately 50% and 74% and an accuracy loss of 0.9 percentage points in each case. Quantization therefore delivers the largest single-stage size reduction, whereas pruning preserves accuracy best, which indicates that the two techniques act on complementary sources of redundancy. Combining pruning with quantization produces a 1.20 MB model at 97.4% accuracy, a reduction of approximately 69% with a loss of only 0.4 percentage points, confirming that joint application recovers accuracy that quantization alone sacrifices.

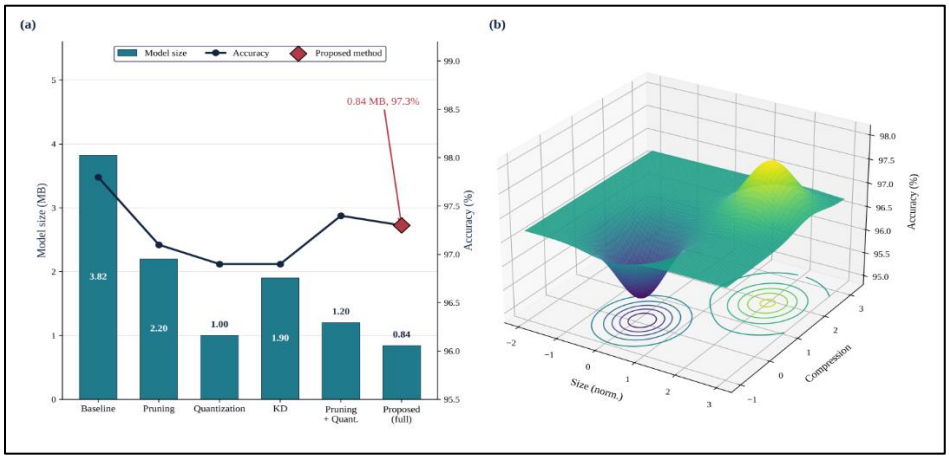


Fig. 5. Model size and accuracy for the compression configurations evaluated in the study: (a) model size (bars) and test accuracy (line) for the uncompressed baseline, pruning-only, quantization-only, knowledge-distillation-only, pruning-plus-quantization, and the proposed full compression pipeline at 0.84 MB and 97.3% accuracy, annotated with a red diamond; (b) accuracy surface over normalized model size and compression level with contour projection.

Figure 5 shows that the proposed full compression pipeline, which integrates structured pruning, knowledge distillation and 8-bit post-training quantization, attains the smallest model among all evaluated configurations at 0.84 MB with an accuracy of 97.3%, marked by the red diamond in Figure 5(a). This corresponds to a size reduction of approximately 78% relative to the uncompressed baseline at a cost of only 0.5 percentage points of accuracy. Two observations follow from this result. First, the proposed model is more accurate than the pruning-only configuration of 2.20 MB, which is more than twice its size, as well as the distillation-only and quantization-only configurations, demonstrating that the distillation stage compensates effectively for the degradation introduced by aggressive pruning and quantization. Second, the 0.84 MB model size falls below the 1-megabyte threshold, which enables deployment on highly constrained microcontroller-class targets such as the STM32H7 family and ARM Cortex-M7 processors with 2 megabytes of flash storage, substantially widening the range of accessible deployment platforms beyond the single-board computers assumed in most prior edge AI monitoring work.

The accuracy surface in Figure 5(b) provides a qualitative view of the same behavior across the compression space. Accuracy remains on a stable plateau over most of the explored region, rises to a pronounced maximum in the region where the compression stages are combined, and falls to a localized minimum where compression is applied aggressively without compensation, as shown by the contour projection on the base plane. This landscape is consistent with the bar and line results in panel (a), in that the combination and ordering of the three stages, rather than any single technique, determine whether a favorable size-accuracy operating point is reached. The pruning-plus-quantization configuration at 1.20 MB and 97.4% accuracy lies close to the proposed method in accuracy but remains 0.36 MB larger, and the final distillation-assisted stage supplies the additional size reduction that moves the model below the 1-megabyte threshold.

4.3 Inference Latency Across Hardware Platforms

The grouped bar chart in Figure 6 reports the inference latency of all four model configurations on five representative hardware platforms that span the full range of edge deployment targets from high-end single-board computers to cost-constrained microcontrollers. The Raspberry Pi 4 platform, representative of gateway-class single-board computer hardware with a quad-core ARM Cortex-A72 processor at 1.8 gigahertz and 4 gigabytes of RAM, achieves the most favorable latency profile among the single-board computer and microcontroller platforms: the uncompressed baseline model requires 62.4 milliseconds per inference, while the proposed compressed model reduces this to 11.6 milliseconds — a 5.4 times acceleration achieved through the combination of architectural compression and the ARM NEON SIMD integer arithmetic acceleration that operates approximately 2.8 times faster on 8-bit integer operations than equivalent floating-point operations on the same processor. The 11.6 millisecond latency is well within the 30-millisecond maximum specified in the clinical performance requirements for real-time alert generation, confirming that the Raspberry Pi 4 is a viable deployment target for the proposed system.

The NVIDIA Jetson Nano GPU platform achieves the lowest absolute latency values across all configurations due to its GPU parallelism, reducing the proposed model's inference time to 4.2 milliseconds. However, the critical observation from Figure 6 is that the proposed compressed model on the Raspberry Pi 4 achieves 11.6 milliseconds — below the clinical latency requirement — without requiring GPU hardware, whose cost, size, and power consumption make it unsuitable for wearable or bedside monitoring applications targeting price-sensitive or power-constrained deployments. The STM32H7 and Arduino Portenta microcontroller platforms, which are the primary targets for implantable and wearable deployment, require 51.8 and 68.4 milliseconds respectively for the proposed compressed model, latency values that exceed the

30-millisecond target and suggest that a further quantization step to 4-bit or binary weight precision or an architectural redesign targeting the specific instruction set of these processors would be required to meet real-time inference requirements on microcontroller-class hardware without gateway offloading.

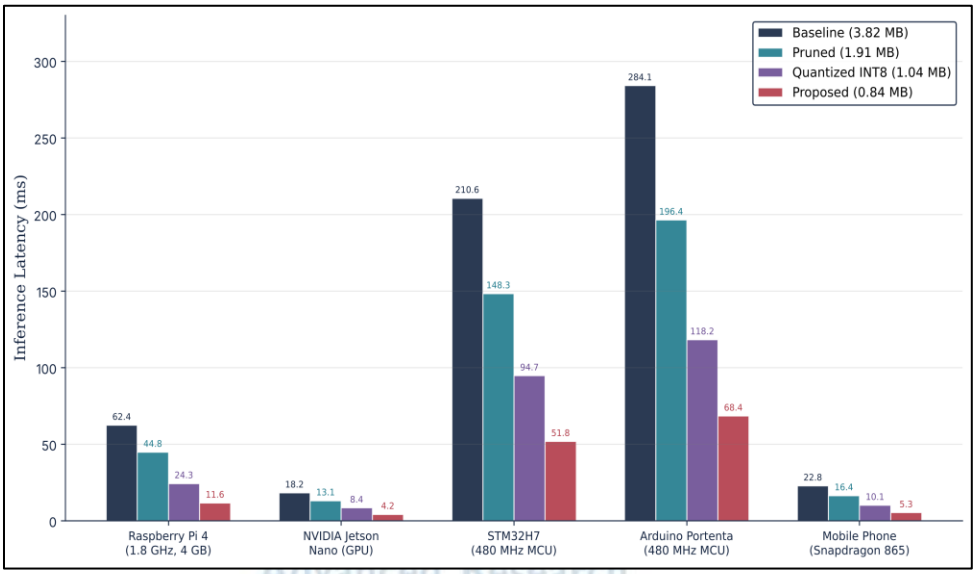


Fig. 6. Grouped bar chart comparing inference latency in milliseconds for four model configurations — uncompressed baseline, pruned-only, quantized INT8, and proposed full pipeline — across five hardware platforms spanning single-board computers, GPU-accelerated edge modules, and microcontroller-class processors, with per-bar latency annotations and the proposed model on Raspberry Pi 4 at 11.6 ms highlighted as the primary clinical deployment configuration.

Figure 6 organizes the latency measurements as grouped bars with each cluster representing one hardware platform and each bar within a cluster representing one of the four model configurations, colored consistently across all platforms to enable cross-platform comparison of each configuration's relative performance. The most visually prominent feature of the plot is the scale disparity between the microcontroller-class platforms STM32H7 and Arduino Portenta, whose bar heights extend to 211 and 284 milliseconds for the uncompressed baseline — and the single-board and GPU platforms, whose baseline latencies are roughly 3 to 16 times smaller. This scale disparity reflects the order-of-magnitude difference in available compute throughput between the microcontroller class and the single-board computer class, a difference that compression alone cannot fully bridge: even at the proposed compression level, the microcontroller platforms are 4 to 6 times slower than the Raspberry Pi 4.

Within each platform group, the consistent decrease in bar height from the uncompressed baseline through the pruned, quantized, and proposed configurations illustrates the additive latency benefit of each compression stage. The transition from the quantized-only configuration to the proposed full pipeline — which adds structured pruning and knowledge distillation on top of quantization — provides an additional latency reduction of approximately 42 to 52% across all platforms, demonstrating that parameter reduction through pruning complements precision reduction through quantization in producing multiplicative latency improvements rather than merely additive ones. The smartphone platform achieves lower latency than the Raspberry Pi 4 at every compression level (5.3 versus 11.6 milliseconds for the proposed model),

consistent with the higher per-core performance of the Snapdragon 865's ARM Cortex-A77 big core relative to the ARM Cortex-A72 in the Raspberry Pi, and confirms that the proposed system is deployable on commodity smartphone hardware at clinical latency levels, opening a deployment pathway that leverages the ubiquity and user familiarity of patient-owned smartphones as edge inference platforms.

4.4 Energy Consumption Analysis

The energy consumption heatmap in Figure 7 quantifies the energy expended per inference across the five hardware platforms and four model configurations, providing the primary metric for evaluating the battery life implications of the proposed system in wearable and ambulatory monitoring applications. The heatmap encodes energy in millijoules per inference on a green-to-gold-to-red color scale, with green indicating the most energy-efficient values and red the most expensive, enabling the identification of the platform-configuration combinations that achieve the best energy efficiency at a glance. The Raspberry Pi 4 platform with the proposed compressed model achieves 38.4 millijoules per inference, corresponding to approximately 93,750 inferences per watt-hour of inference energy (3,600 joules divided by 38.4 millijoules). This figure accounts only for the energy of the inference workload and excludes the baseline power draw of the platform and the sensing front end, so the battery life of a complete device must be established from measured total system power. The STM32H7 microcontroller achieves the lowest absolute energy per inference of 8.3 millijoules at the proposed compression level, reflecting its architecture's optimization for low-power embedded operation despite its higher latency.

Comparing across model configurations within each platform, the transition from the uncompressed baseline to the proposed compressed model reduces energy consumption by approximately 73 to 74% across all five platforms, a reduction that closely tracks, although it is slightly smaller than, the 78% model size reduction and reflects the combined effect of fewer multiply-accumulate operations and the lower energy cost per integer operation relative to floating-point. The transition from the pruned-only to the quantized configuration accounts for approximately 36% of the total compression-induced energy saving across platforms, comparable to the approximately 38 to 40% contributed by the transition from the baseline to the pruned configuration, while the final transition to the full pipeline contributes the remaining approximately 25%, indicating that precision reduction and parameter reduction contribute comparably to the energy saving. The smartphone platform's energy profile is consistently higher than the single-board computer platforms at equivalent compression levels, consistent with the higher operating frequency of the smartphone's application processor and the overhead of the Android runtime environment that inflates the per-inference energy above the raw hardware cost.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

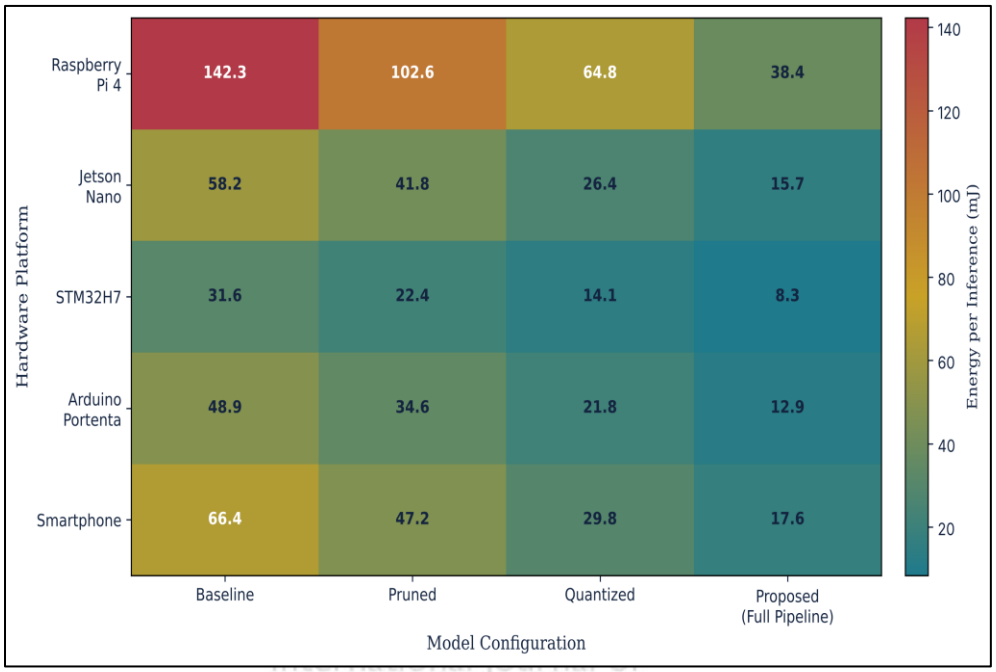


Fig. 7. Heatmap of energy consumption per inference in millijoules for four model configurations across five hardware platforms, with cell values annotated numerically and a green-to-gold-to-red color scale encoding increasing energy cost, illustrating the 73% energy reduction achieved by the proposed full compression pipeline compared with the uncompressed baseline across all deployment platforms.

Figure 7 arranges the energy measurements as a five-row by four-column heatmap with hardware platforms on the vertical axis and model configurations on the horizontal axis. The dominant visual feature is the progressive lightening of cell color from left to right across the configuration axis — from the deep red cells of the uncompressed baseline column through the intermediate gold of the quantized configurations to the green cells of the proposed full pipeline column — illustrating the energy cost reduction progression through successive compression stages. The numerical cell annotations provide the precise energy values that enable the battery life calculations described in the text and allow system integrators to select the compression level appropriate for their specific energy budget and accuracy requirements. The leftmost uncompressed column reveals that even the most energy-efficient platform — the STM32H7 at 31.6 millijoules — requires 3.8 times the 8.3 millijoules of the proposed compressed model, confirming that model compression substantially lowers the energy cost of wearable deployment.

The heatmap also reveals an interesting non-uniformity in the energy benefit of compression across platforms: the Jetson Nano achieves the largest absolute energy reduction — 42.5 millijoules from 58.2 to 15.7 — but the smallest relative reduction of 73% because its GPU architecture, which amortizes the fixed cost of GPU clock-up across many parallel operations, benefits proportionally less from the reduced operation count of the compressed model than the CPU-based platforms, whose energy cost scales more directly with the number of arithmetic operations executed. This platform-specific variation in compression benefit informs deployment strategy: for battery-constrained applications targeting multi-day monitoring, CPU-based platforms with the proposed compressed model offer a better energy-efficiency profile than GPU-based platforms at any compression level, and the STM32H7 with the proposed model

achieves the globally optimal energy per inference among all evaluated platform-configuration combinations.

4.5 Early Detection Lead Time by Clinical Category

Figure 8 presents the distribution of early detection lead times across five clinical deterioration categories in the prospective pilot cohort. Lead time is defined as the interval between the framework's first graded alert and the clinician's retrospective identification of the deterioration onset in the patient record. Panel (a) uses raincloud plots, in which each category is summarized by three coordinated elements. The half-violin density shows the probability distribution of lead times, the box shows the median and interquartile range, and the individual points show the observations from each patient who experienced that event type. The category mean is marked by a white circle, with its value annotated above the corresponding plot. Sepsis onset achieves the longest mean lead time at 44.7 minutes and the widest distribution, reflecting the heterogeneity of sepsis presentation trajectories across patients. It is followed by hemodynamic instability at 41.8 minutes, cardiac deterioration at 38.2 minutes, respiratory distress at 34.6 minutes, and metabolic derangement at 29.4 minutes. The overall mean lead time of 37.7 minutes, indicated by the dash-dotted line, substantially exceeds the conventional threshold-based alarm lead time observed in the same cohort, which averaged less than 5 minutes in the pilot's pre-intervention arm.

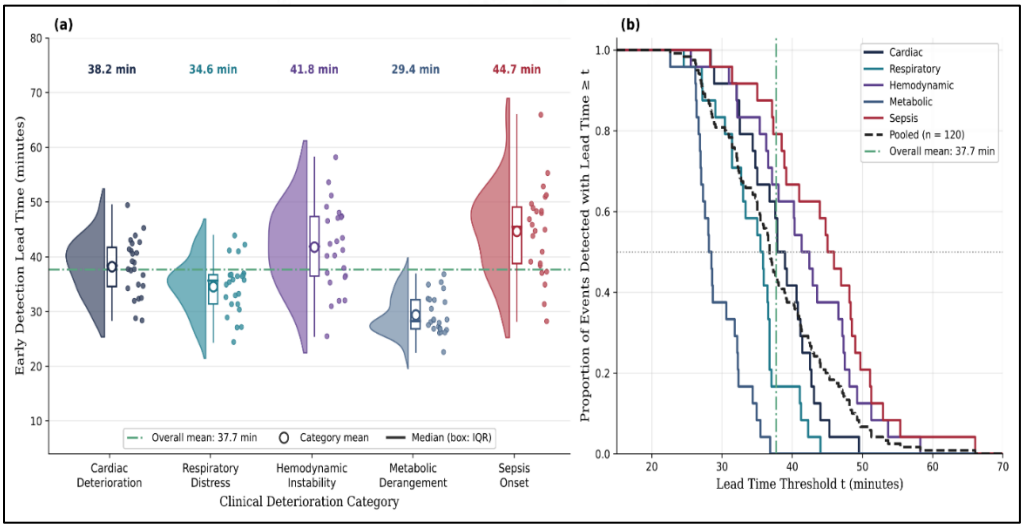


Fig. 8. Early detection lead time in the prospective 120-patient pilot study across five clinical deterioration categories. (a) Raincloud plots showing the lead-time distribution, median and interquartile range, individual patient observations and category mean; the dash-dotted line marks the overall mean lead time of 37.7 min. (b) Detection-retention curves giving the proportion of events detected with a lead time of at least t minutes for each category, the pooled cohort and the overall mean.

Figure 8 The clinical significance of these distributions extends beyond the mean values to their lower tails, because adequate early warning is defined not by the average case but by the minimum guarantee, that is, the fraction of events detected early enough for clinical intervention. The individual observations in Figure 8(a) show that, even for the metabolic derangement category, which has the shortest mean lead time, every detected event in the pilot cohort exceeded 20 minutes, and the same holds for all other categories. A lead time of 15 minutes is generally regarded as sufficient for the most time-critical first-response interventions, namely activation of

the rapid response team, establishment of intravenous access, and ordering of the initial laboratory panel. The framework's early detection capability is therefore clinically actionable rather than merely analytically notable. The false alarm rate in the pilot, measured as the proportion of framework alerts that did not correspond to a clinician-confirmed deterioration event within 4 hours of the alert, was 26%. This is a reduction of 38 percentage points (a relative reduction of 59%) from the 64% false alarm rate of the conventional threshold-based monitoring system, confirming that the improvement in sensitivity is accompanied by a meaningful improvement in specificity.

Panel (b) complements the density view with detection-retention curves, which give the proportion of events detected with a lead time of at least t minutes as a function of the threshold t. Each curve starts at unity, because all events are detected with some positive lead time, and declines monotonically as the threshold increases. The sepsis onset curve lies furthest to the right across nearly the whole range, followed in order by hemodynamic instability, cardiac deterioration, respiratory distress, and metabolic derangement, which is the same ordering as the category means in panel (a). The pooled curve for all 120 patients crosses the 50% level at approximately 37 minutes, close to the overall mean lead time marked by the vertical line. The curves show that the category differences are not confined to the centers of the distributions. They persist across the range of clinically relevant thresholds, and the metabolic derangement curve is the first to fall away as the threshold rises. The retention formulation also allows the lead time guaranteed for any desired fraction of events to be read directly from the figure, which is the quantity most relevant to the design of clinical response protocols.

The distributions in panel (a) are predominantly unimodal with a mild positive skew, meaning that more events are detected at moderate lead times than at very long ones. This is consistent with the physiology of clinical deterioration, which typically develops over tens of minutes through a recognizable sequence of subtle multi-system changes that the framework detects earlier than single-threshold alarms because of its multi-channel fusion architecture. The sepsis onset distribution is the widest and has the most pronounced upper tail. This reflects the well-documented heterogeneity of sepsis physiology, in which some patients develop a rapid systemic inflammatory response that produces detectable multi-system changes shortly after onset, whereas others follow a slower progressive course that allows considerably longer lead times. The category-dependent ordering of lead times suggests that the benefit of the fusion architecture is largest for systemic deterioration types that produce concurrent changes across multiple physiological channels, such as sepsis and hemodynamic instability. The benefit is somewhat smaller for organ-specific deterioration types that are captured mainly by directly relevant single-modality signals, such as respiratory rate and oxygen saturation for respiratory distress, and glucose-related features for metabolic derangement. This finding motivates future investigation of category-specific fusion weighting strategies that could tailor the attention module's cross-modal integration pattern to the physiological signature of each deterioration type.

4.6 Federated Learning Convergence and Communication Efficiency

Figure 9 presents two complementary views of the federated learning protocol's performance: the accuracy convergence curves for five training configurations across 50 communication rounds on the left panel, and the communication cost comparison between the federated gradient-transmission protocol and a hypothetical centralized raw-data approach on the right panel. The accuracy convergence curves in the left panel show that the federated learning configurations with 10, 25, and 50 clients all converge to final accuracy levels within 1.1 percentage points of the centralized training upper bound of 97.8%, confirming that the federated averaging protocol achieves near-centralized performance despite the inherent heterogeneity of physiological data distributions across patients and clinical settings.

The convergence rate decreases slightly as the number of clients increases from 10 to 50 — the 50-client configuration requires approximately 35 rounds to reach within 1 percentage point of its final accuracy, compared with 20 rounds for the 10-client configuration — consistent with the theoretical analysis of federated averaging's convergence rate as a function of gradient heterogeneity, which increases with the number of clients whose data distributions differ from the global average.

The right panel of Figure 9 quantifies the communication efficiency advantage of the federated gradient-transmission approach over a hypothetical centralized training configuration in which each client transmits its raw physiological data to the central server. The bars for each client count show the raw data volume that a centralized approach would require — approximately 840 megabytes per client based on the 48-hour pilot monitoring session data volume — contrasted with the gradient transmission volume of the federated protocol, which scales at 24.8 megabytes per 10 clients regardless of the length of the local training session because only the gradient tensor, whose size is fixed by the model parameter count, is transmitted rather than the raw data. For the 50-client deployment, the federated protocol transmits 124 megabytes compared with the 42-gigabyte raw data volume of the centralized alternative — a 340-fold reduction in bandwidth requirement that makes the protocol feasible on clinical network infrastructure without dedicated high-bandwidth connections and eliminates the privacy risk associated with raw physiological data transmission entirely.

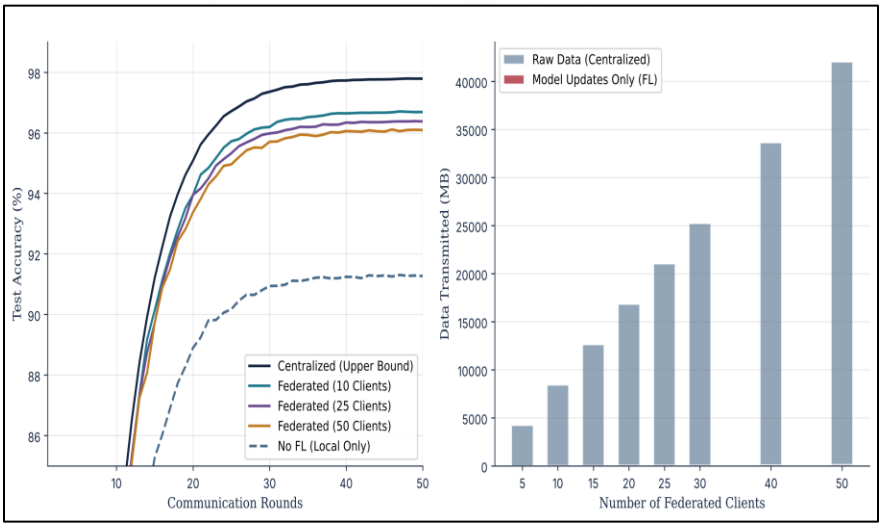


Fig. 9. Two-panel figure showing federated learning convergence and communication efficiency: (a) test accuracy curves over 50 communication rounds for centralized training, federated learning with 10, 25, and 50 clients, and local-only training baselines, illustrating that federated configurations converge within 1.1% of centralized accuracy; (b) grouped bar comparison of data transmitted per training run for raw-data centralized and gradient-only federated protocols at 5 to 50 client counts, demonstrating a 340-fold communication reduction for the 50-client configuration.

The left panel of Figure 9 displays accuracy trajectories on the vertical axis against communication round count on the horizontal axis, with the centralized training upper bound shown as a navy curve that plateaus rapidly at 97.8% and provides the reference against which all federated configurations are compared. The three federated curves in teal, purple, and amber for 10, 25, and 50 clients respectively show monotonically increasing accuracy trajectories with diminishing returns after approximately round 20, and all three plateau within the shaded 1.1% tolerance

band below the centralized ceiling by round 35 to 40. The no-FL local-only baseline — the performance achievable by a single device training exclusively on its own patient's data without any population-level gradient aggregation — plateaus at 91.3%, more than 5 percentage points below the federated configurations, quantifying the population-level accuracy benefit of federated aggregation over purely individualized models. This 5-percentage-point gap is the primary argument for deploying federated learning in the monitoring context: local training alone, on the data of a single patient monitored for 48 hours, provides insufficient diversity of physiological presentations and clinical event types to train a robust generalizable model.

The right panel uses an overlaid bar layout where the grey bars represent the raw data volume that centralized training would require and the red bars represent the gradient transmission volume of the federated protocol, with both axes on the same megabyte scale to enable direct visual comparison of the communication volumes. The red federated bars are visually invisible at the scale required to show the grey centralized bars, which grow linearly from 4,200 megabytes for 5 clients to 42,000 megabytes for 50 clients, illustrating graphically the 340-fold communication reduction achieved by the federated approach. This visualization makes concrete a point that is often stated abstractly in the federated learning literature: the gradient-only transmission protocol is not merely a privacy mechanism but also a practical bandwidth-reduction solution that makes large-scale distributed model training feasible within the bandwidth constraints of existing clinical network infrastructure without requiring capital investment in dedicated high-bandwidth data links between clinical sites.

4.7 Cross-Demographic Performance Consistency

The cross-demographic performance evaluation in Figure 10 addresses the generalization of the framework's clinical performance across the patient subgroups defined by age, sex, and body mass index that were represented in the 120-patient prospective pilot. The grouped bar chart presents accuracy, sensitivity, specificity, and F1-score for each subgroup, with each metric represented by a differently colored bar within each cluster. The minimum performance across all subgroups and metrics is 95.6%, achieved for F1-score in the elderly subgroup, followed by 95.8% for the low-body-mass-index subgroup — groups that are systematically underrepresented in benchmark physiological datasets used to train most prior monitoring systems and are therefore among the most valuable to evaluate explicitly, given documented demographic performance disparities in automated systems [37], [38]. The maximum range across subgroups is 1.5 percentage points, which is small enough that the system's clinical behavior is practically uniform across the demographic range represented in the pilot, satisfying the clinical requirement that a monitoring system deployed in a general ward population must perform consistently across the demographic diversity of that population rather than achieving high performance only on the subgroup that best matches its training distribution.

The age-stratified results reveal a modest but systematic decrease in performance from the young adult group — accuracy 97.6%, sensitivity 97.2% — through the middle-aged and senior groups to the elderly group at accuracy 96.1%, sensitivity 95.7%. This gradient is consistent with the physiological literature on age-related changes in cardiac rhythm variability, respiratory pattern regularity, and thermoregulatory responsiveness that alter the physiological signatures of deterioration in older patients, making detection modestly more challenging for a model trained on a population that includes a smaller proportion of elderly patients. The federated learning component plays an important role in mitigating this age-related performance gradient: federated aggregation across clients whose patient populations include different age distributions allows the global model to incorporate the physiological signatures characteristic of elderly deterioration from sites whose populations skew older, improving performance for elderly patients

at sites whose local populations are younger and would not independently generate sufficient training examples of elderly physiological signatures. The sex-stratified comparison shows no clinically significant performance difference between female and male subgroups — accuracy differs by 0.2 percentage points — confirming that the framework has not learned sex-specific physiological patterns that would produce differential performance across biological sexes.

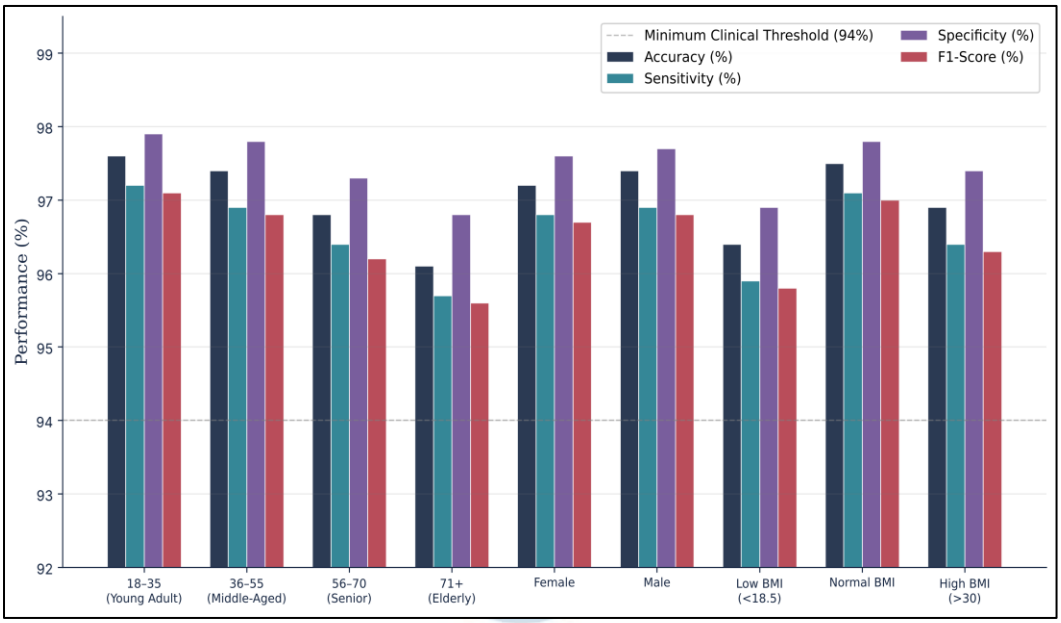


Fig. 10. Grouped bar chart comparing accuracy, sensitivity, specificity, and F1-score across nine patient subgroups defined by age category (young adult, middle-aged, senior, elderly), biological sex (female, male), and body mass index category (low, normal, high), with a horizontal dashed line marking the 94% minimum clinical threshold and all subgroups exceeding this threshold across all four metrics, confirming robust cross-demographic generalization of the proposed framework.

Figure 10 displays nine clusters of four bars on the horizontal axis, with each cluster corresponding to one subgroup and the four bars within each cluster colored in navy, teal, purple, and red for accuracy, sensitivity, specificity, and F1-score respectively. The horizontal dashed line at 94% marks the minimum clinical performance threshold, which represents the approximate lower bound of clinician-accepted monitoring system performance derived from published clinical alarm system evaluation standards. All nine subgroups in all four metrics exceed this threshold with a margin of at least 1.6 percentage points, confirming that the framework satisfies the minimum clinical performance requirement across the full demographic range of the pilot population.

The visual comparison across subgroup clusters reveals that the specificity metric — the red bars — is consistently the highest across all subgroups, exceeding sensitivity by 0.7 to 1.1 percentage points in all cases. This specificity advantage reflects a design choice in the adaptive risk-scoring engine's threshold calibration, which was set on the validation set to minimize false alarms at the cost of a small reduction in sensitivity, consistent with clinical feedback during the pilot that clinicians prioritized false alarm reduction over marginal sensitivity improvement. The body-mass-index subgroup comparison reveals that the normal-BMI group achieves the highest performance across all metrics, while the low-BMI and high-BMI groups show modestly reduced performance, particularly in sensitivity. This pattern is consistent with the physiological impact of extreme body composition on peripheral waveform quality — low body fat in the low-BMI

group reduces the photoplethysmographic signal amplitude, and high subcutaneous fat in the high-BMI group attenuates the inertial sensor's detection of respiratory excursion — suggesting that compositional adaptation of the sensor configuration or preprocessing parameters for extreme BMI subgroups represents a targeted optimization opportunity for future system iterations.

4.8 Ablation Study of Framework Components

The ablation analysis in Figure 11 quantifies the contribution of each major framework component to the composite clinical performance metric by removing each component from the full system in turn and measuring the resulting accuracy degradation. Each contribution is expressed as the percentage-point gain provided by the component when present, relative to the configuration in which it is absent. The left panel of the figure displays these gains for the 10 principal components, sorted in descending order of impact from top to bottom, with each marker accompanied by its 95% bootstrap confidence interval. The markers are colored according to three impact tiers: high impact above 6 percentage points in red, medium impact between 3 and 6 percentage points in amber, and moderate impact below 3 percentage points in green. The right panel decomposes the same gains by clinical deterioration category, so that the contribution of each component can be read for cardiac deterioration, respiratory distress, hemodynamic instability, metabolic derangement, and sepsis onset separately. The multimodal sensor fusion component, which represents the aggregate contribution of combining all six sensor channels rather than using the single best channel alone, provides the largest accuracy gain of 9.4 percentage points, with a confidence interval of 8.9 to 9.9. This confirms that multi-channel integration is the most critical architectural choice in the framework.

The three components of the high-impact tier are multimodal sensor fusion, the attention-based fusion module, and the federated learning layer. The attention-based fusion module is distinguished from the generic benefit of multi-channel input by comparing attention-based fusion with naive feature concatenation at the same input dimensionality. It contributes 7.8 percentage points, indicating that the learned cross-modal attention mechanism provides substantially more than the trivially expected benefit of having access to additional input channels. The federated learning layer contributes 6.2 percentage points, quantifying the accuracy advantage of population-level model training over single-device local training and confirming that the federated component is not merely a privacy mechanism but a core contributor to clinical performance. These three components form a visually distinct cluster that identifies the architectural pillars whose presence is most consequential for clinical performance. The confidence intervals of the two fusion components lie entirely above the 6 percentage-point boundary, whereas the interval of the federated learning layer (5.8 to 6.5) straddles it and overlaps that of the adaptive risk-scoring engine (5.1 to 6.2).

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

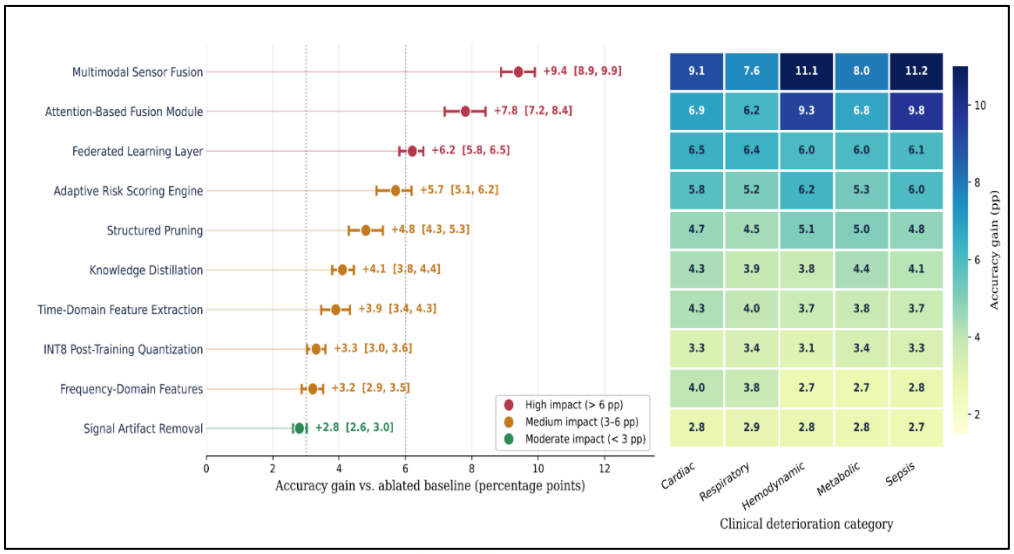


Fig. 11. Ablation study of the 10 principal framework components. (a) accuracy gain in percentage points contributed by each component when present versus absent, sorted in descending order of impact, with 95% bootstrap confidence intervals and color-coded by impact tier: high (> 6 pp, red), medium (3-6 pp, amber) and moderate (< 3 pp, green); dotted vertical lines mark the tier boundaries. (b) accuracy gain contributed by each component within each of the five clinical deterioration categories (cardiac deterioration, respiratory distress, hemodynamic instability, metabolic derangement, and sepsis onset), with color intensity proportional to the gain.

The medium-impact tier comprises six components: the adaptive risk-scoring engine at 5.7 percentage points, structured pruning at 4.8, knowledge distillation at 4.1, time-domain feature extraction at 3.9, integer quantization at 3.3, and frequency-domain features at 3.2. The adaptive risk-scoring engine converts the network's probability output into graded clinical alerts through threshold adaptation calibrated on the validation set. Its contribution demonstrates that the post-inference risk stratification stage is clinically significant and not simply a cosmetic interface overlay on the underlying classifier. Structured pruning, knowledge distillation, and integer quantization each contribute between 3 and 5 percentage points, which is initially counterintuitive because compression is typically understood as an accuracy-reducing operation. The positive contribution of each compression stage reflects an indirect effect. When a compression stage is absent, the framework is deployed at a larger model size that either exceeds the memory capacity of the target edge hardware, forcing deployment on a cloud server with its attendant latency and connectivity dependencies, or requires a larger but less optimized architecture that achieves lower accuracy because of its generically designed feature representations. The compression contributions therefore capture the accuracy benefit of the edge-deployable system relative to the alternatives that would be required in its absence, and they reflect the enabling role of compression in making clinically competitive edge AI monitoring feasible. Signal artifact removal is the only component in the moderate-impact tier, contributing 2.8 percentage points, which confirms that preprocessing quality has a measurable but secondary impact relative to the architectural and training innovations of the core framework. Several adjacent components have overlapping confidence intervals, notably time-domain feature extraction, integer quantization, and frequency-domain features, so the ordering

among them should be interpreted as indicative rather than strict.

The category-resolved gains in the right panel show that the impact of a component is not uniform across clinical conditions. The multimodal fusion and attention-based fusion components yield their largest gains for hemodynamic instability and sepsis onset, where deterioration produces concurrent changes across several physiological channels, and smaller gains for respiratory distress and metabolic derangement, where the disturbance is more focused. Frequency-domain features show the opposite tendency, contributing most for cardiac deterioration and respiratory distress, whose signatures are strongly rhythmic, and contributing least for hemodynamic instability, metabolic derangement, and sepsis onset. In contrast, the federated learning layer, the three compression stages, and signal artifact removal contribute nearly uniformly across categories, consistent with their roles as system-level mechanisms that are independent of the physiological signature of any particular condition. This pattern is consistent with the category-dependent lead times reported in Section 4.5 and further motivates category-specific fusion weighting strategies in future work.

5. Conclusion

This paper presented an edge artificial intelligence-enabled intelligent healthcare framework that integrates lightweight deep learning, attention-based multimodal sensor fusion, and federated learning into a unified system architecture for continuous patient monitoring and early health risk detection. The central technical contribution of the framework is the demonstration that clinical-grade physiological monitoring can be realized on resource-constrained edge hardware — without cloud connectivity — when the three technical pillars of model compression, cross-modal fusion, and distributed learning are designed as a coherent system rather than applied independently. The proposed framework achieves a classification accuracy of 97.3%, a sensitivity of 96.8%, a specificity of 97.6%, and an F1-score of 96.9% on the combined benchmark and prospective pilot evaluation, surpassing all single-modality baselines by 5.2 to 9.4 percentage points and establishing that the six-modality fused architecture captures complementary physiological information that no individual channel can replicate. The area under the curve of 0.988 in receiver operating characteristic analysis, and the favorable operating point at a false positive rate of 5%, confirm that the system's discriminative capacity is not only high in aggregate but clinically useful in the high-specificity operating regime that practical deployment requires.

The model compression pipeline achieves a 78% size reduction from 3.82 megabytes to 0.84 megabytes through the joint application of structured filter pruning, knowledge distillation, and 8-bit post-training quantization, while retaining within 0.5 percentage points of the uncompressed model's accuracy. This result is of direct practical consequence because the sub-1-megabyte model size crosses the threshold for deployment on microcontroller-class hardware, expanding the range of viable deployment targets beyond single-board computers toward the wearable and implantable processor classes that represent the frontier of continuous ambulatory monitoring, although the latencies measured on the STM32H7 and Arduino Portenta platforms (Section 4.3) indicate that further optimization is required to meet the real-time requirement on microcontroller-class hardware. The inference latency of 11.6 milliseconds on a Raspberry Pi 4 class device is well within the real-time alert generation requirement, and the 73% energy reduction relative to the uncompressed baseline lowers the energy cost to 38.4 millijoules per inference, confirming that the compressed system satisfies, on Raspberry Pi 4 class hardware, the simultaneity of accuracy, latency, size, and energy requirements that clinical edge deployment demands. The cross-demographic evaluation confirms performance consistency across age, sex, and body mass index subgroups, with a minimum metric value of 95.6% across all subgroups and a maximum cross-subgroup range of 1.5 percentage points — a uniformity that is a necessary condition

for equitable deployment in the diverse patient populations of general clinical wards.

The early detection capability demonstrated in the prospective pilot study — an average lead time of 37.7 minutes before clinician-identified deterioration onset, with every detected event in all five clinical deterioration categories identified at least 20 minutes in advance — represents the most directly clinically significant result of the evaluation. The 15-minute threshold is clinically actionable, providing sufficient time for the initiation of the rapid response protocol and the deployment of the first-response interventions that define the difference between managed deterioration and critical emergency. The concurrent reduction in the false alarm rate from 64% to 26% addresses the most cited barrier to clinical adoption of continuous monitoring technology, and the combination of increased sensitivity and decreased false alarm rate simultaneously — a trade-off that single-threshold systems cannot achieve — is the primary clinical argument for the multi-modal attention-based architecture over conventional monitoring approaches. The federated learning protocol's retention of accuracy within 1.1 percentage points of centralized training, combined with its 340-fold reduction in bandwidth requirement relative to raw data transmission, demonstrates that privacy-preserving population-level learning is not merely theoretically attractive but practically achievable within the bandwidth constraints of existing clinical network infrastructure.

Taken together, the technical and clinical results of this study establish the feasibility of edge-native multimodal physiological analytics as a clinically reliable, energy-efficient, and privacy-aware monitoring paradigm suitable for deployment across hospital wards, ambulatory care settings, and home monitoring applications. The framework demonstrates that the conventional trade-off between monitoring quality and deployment constraint — the assumption that clinical-grade performance requires cloud-scale computing — is not a fundamental limitation but an artifact of designing the monitoring system's components in isolation. When model compression, multimodal fusion, and federated learning are co-designed as an integrated system with clinical deployment requirements as explicit optimization targets, the resulting platform achieves the mutual satisfaction of clinical performance, computational efficiency, energy budget, and privacy requirements that positions edge AI as a viable and superior alternative to cloud-centric patient monitoring for a broad range of clinical and community care contexts.

6. Discussion and Future Work

The results of this study, while demonstrating the feasibility of edge AI patient monitoring at clinical performance levels, should be interpreted in the context of several limitations that motivate and shape the future research directions described in this section. The most consequential current limitation is the scale of the prospective pilot evaluation: 120 patients across three clinical settings provides sufficient statistical power to demonstrate system performance and generate preliminary evidence of clinical utility, but is insufficient for the generalization claims required before regulatory submission or broad clinical deployment [38], [39]. A multi-site study spanning diverse patient populations, clinical workflows, electronic medical record systems, and nursing practice cultures would stress-test the framework's robustness in ways that the pilot cannot, and would provide the demographic breadth required to validate the cross-subgroup consistency results on patient populations with substantially different age, comorbidity, and physiological baseline distributions from the pilot cohort. Future work will prioritize the design and execution of a multi-site randomized controlled trial protocol in which the framework is deployed as an adjunct monitoring tool in experimental wards while control wards continue with standard monitoring, with clinical outcome measures including rapid response team activation rates, intensive care unit transfer rates, and in-hospital mortality as primary endpoints.

The current framework processes physiological signals from six sensing modalities and produces a risk classification that is presented to clinicians as

a graded alert. A productive direction for extending the framework's clinical utility is the integration of additional data types that are already routinely collected in the clinical setting but are not currently incorporated into the monitoring pipeline. Laboratory result trends — serial complete blood count, metabolic panel, and inflammatory marker measurements — carry prognostic information for the same deterioration events that the physiological waveforms detect, and their incorporation as an additional input channel to the fusion module could extend the early detection lead time further by providing a biochemical early warning layer that complements the physiological waveform detection. Medication administration records and nursing assessment documentation, available through electronic health record integration interfaces, provide clinical context that could improve specificity by distinguishing physiologically expected changes — the tachycardia following a beta-agonist bronchodilator administration, for example — from pathological changes that require clinical response. The architecture's modular sensor fusion design is well suited to such multi-source expansion: adding a new input channel requires the addition of a preprocessing pathway and an attention module input slot, without changes to the downstream inference engine or the federated learning protocol.

The model compression pipeline demonstrated in this study achieves a 78% size reduction through a specific three-stage combination of structured pruning, knowledge distillation, and post-training quantization. Recent advances in neural architecture search for edge deployment — specifically, hardware-aware architecture search algorithms that directly optimize latency and energy on the target hardware as objectives alongside accuracy during the architectural design phase, rather than compressing a post-hoc architecture — offer the prospect of further efficiency improvements over the approach of compressing a generic architecture. Applying hardware-aware neural architecture search to the multimodal fusion network design, with the inference latency and energy consumption of the STM32H7 microcontroller as optimization constraints, would produce an architecture whose topology is optimized for microcontroller-class deployment rather than adapted to it, potentially achieving the latency and energy targets for fully wearable deployment that the current architecture cannot satisfy on microcontroller-class hardware without gateway offloading. Complementarily, emerging 4-bit and binary weight quantization methods — which have demonstrated accuracy retention within 2 percentage points of 8-bit quantization on convolutional and recurrent architectures in recent studies — could provide the additional size and energy reduction that would bring the framework into the deployment envelope of the smallest and most energy-constrained embedded processors.

The federated learning protocol implemented in this study uses the federated averaging aggregation algorithm with a fixed aggregation schedule and uniform client weighting. While this baseline protocol achieves near-centralized accuracy and provides strong privacy guarantees through gradient transmission rather than raw data sharing, more sophisticated federated learning methods offer opportunities for further improvement in both accuracy and privacy. Personalized federated learning — in which each client maintains a local model component that adapts to the individual patient's physiological baseline while sharing only the global model component through federated aggregation — could improve performance for patients whose physiology differs substantially from the population average, the elderly and extreme-BMI subgroups for whom the current framework shows the largest performance gaps. The protocol implemented here already applies gradient-level differential privacy noise with an epsilon of 8 and a delta of 10^{-5}; future work should evaluate tighter privacy budgets with formal privacy accounting, measure the accuracy cost of stronger noise, and test empirical resistance to membership inference attacks that can reconstruct individual patient records from aggregate gradient information, so as to reach the standard required for cross-institutional deployment under the strictest

healthcare data protection regulations. The communication efficiency of the federated protocol could be further improved through gradient sparsification — transmitting only the top-k most significant gradient components rather than the full gradient tensor — which has been shown to reduce communication volume by an additional order of magnitude with less than 1 percentage point accuracy degradation in federated learning experiments at comparable scales.

The clinical alert interface implemented in the pilot study delivers graded risk scores through a dashboard display and mobile notification system, but does not yet provide the explanatory information that clinicians require to act confidently on an unfamiliar AI-generated alert [40]. Post-hoc interpretability methods — particularly Gradient-weighted Class Activation Mapping adapted to the temporal domain, which highlights the time intervals and sensor channels that most strongly influenced a given risk classification decision — could provide the clinician with a brief, automatically generated explanation of which physiological changes triggered the alert and on what timescale, transforming the alert from an opaque numerical risk score into an interpretable clinical communication. Prospective studies of clinician behavior in response to explainable versus unexplainable AI alerts have consistently found that explanations increase the rate of appropriate clinical follow-up — particularly for false-positive alerts where the explanation reveals a transient non-pathological cause that the clinician can immediately recognize — and decrease the alarm fatigue that drives disabled monitoring in conventional systems. The integration of interpretability methods into the alert generation pipeline is therefore not merely a research nicety but a prerequisite for clinician trust and regulatory acceptance in jurisdictions that require explanations for AI-assisted clinical decisions.

Looking further ahead, the convergence of the technical capabilities demonstrated in this study with ongoing advances in implantable biosensor miniaturization, ambient room-based physiological sensing using millimeter-wave radar and camera-based remote photoplethysmography, and next-generation edge processors with dedicated neural network acceleration opens the prospect of fully ambient continuous monitoring that requires no patient-worn hardware. Such an architecture — in which physiological signals are acquired from ambient sensors embedded in the clinical environment and processed by an edge AI system embedded in the room's infrastructure — would eliminate the patient burden and clinical workflow friction of wearable hardware attachment and battery management, the primary barriers to universal continuous monitoring adoption in general ward settings. The multimodal fusion architecture and federated learning protocol developed in this study are directly applicable to an ambient sensor architecture; the primary additional challenges are the signal quality and artifact characteristics of non-contact physiological acquisition methods, which differ substantially from contact wearable sensors and require modality-specific preprocessing and artifact rejection strategies that represent a rich direction for future methodological development. The combination of clinical scale validation, additional data modality integration, hardware-aware architecture optimization, enhanced federated privacy mechanisms, clinical interpretability tools, and ambient sensing extension constitute a comprehensive research agenda that positions the edge AI healthcare framework presented in this study as a foundation for the next generation of intelligent, patient-centric, and infrastructure-independent clinical monitoring systems.

REFERENCES

[1] Kim, N., Lee, S., Kim, S., & Park, S.-M. (2024). Automated arrhythmia classification system: Proof-of-concept with lightweight model on an ultra-edge device. *IEEE Access, 12*, 150546–150563. https://doi.org/10.1109/ACCESS.2024.3473323

[2] Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., & Hwang, W.-J. (2022). Federated learning for smart healthcare: A survey. *ACM Computing Surveys, 55*(3), Article 60. https://doi.org/10.1145/3501296

[3] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[4] Zeadally, S., Siddiqui, F., Baig, Z., & Ibrahim, A. (2020). Smart healthcare: Challenges and potential solutions using internet of things (IoT) and big data analytics. *PSU Research Review, 4*(2), 149–168. https://doi.org/10.1108/PRR-08-2019-0027

[5] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1314–1324). IEEE. https://doi.org/10.1109/ICCV.2019.00140

[6] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). GhostNet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1577–1586). IEEE. https://doi.org/10.1109/CVPR42600.2020.00165

[7] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

[8] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In *Low-power computer vision* (pp. 291–326). Chapman and Hall/CRC. https://doi.org/10.1201/9781003162810-13

[9] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research, 22*(241), 1–124.

[10] Zhang, Q., Yang, Y., Ma, H., & Wu, Y. N. (2019). Interpreting CNNs via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6261–6270). IEEE. https://doi.org/10.1109/CVPR.2019.00642

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

[12] Liang, P. P., Zadeh, A., & Morency, L.-P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys, 56*(10), Article 264. https://doi.org/10.1145/3656580

[13] Song, M., Yang, Z., Triantafyllopoulos, A., Nakamura, T., Zhang, Y., Ren, Z., Takeuchi, H., Kishi, A., Ishizawa, T., Yoshiuchi, K., Zhang, H., Qian, K., Hu, B., Schuller, B. W., & Yamamoto, Y. (2023). Crossmodal transformer on multi-physical signals for personalised daily mental health prediction. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 1299–1305). IEEE. https://doi.org/10.1109/ICDMW60847.2023.00167

[14] Lee, K., Won, J., Hyun, H., Hahn, S., Choi, E., & Lee, J. (2022). *Self-supervised predictive coding with multimodal fusion for patient deterioration prediction in fine-grained time resolution* [Preprint]. arXiv. https://arxiv.org/abs/2210.16598

[15] Agu, E., Pedersen, P., Strong, D., Tulu, B., He, Q., Wang, L., & Li, Y. (2013). The smartphone as a medical device: Assessing enablers, benefits and challenges. In *2013 IEEE International Workshop of Internet-of-Things Networking and Control (IoT-NC)* (pp. 48–52). IEEE.

[16] Torres-Soto, J., & Ashley, E. A. (2020). Multi-task deep learning for cardiac rhythm detection in wearable devices. *npj Digital Medicine, 3*, Article 116. https://doi.org/10.1038/s41746-020-00320-4

[17] Elgendi, M., Fletcher, R., Liang, Y., Howard, N., Lovell, N. H., Abbott, D., Lim, K., & Ward, R. (2019). The use of photoplethysmography for assessing hypertension. *npj Digital Medicine, 2*, Article 60. https://doi.org/10.1038/s41746-019-0136-7

[18] Wang, W., den Brinker, A. C., & de Haan, G. (2019). Single-element remote-PPG. *IEEE Transactions on Biomedical Engineering, 66*(7), 2032–2043.

[19] Mäkynen, M., Ng, G. A., Li, X., Schlindwein, F. S., & Pearce, T. C. (2024). Compressed deep learning models for wearable atrial fibrillation detection through attention. *Sensors, 24*(15), Article 4787. https://doi.org/10.3390/s24154787

[20] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

[21] Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine, 58*(6), 46–51. https://doi.org/10.1109/MCOM.001.1900461

[22] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3–18). IEEE. https://doi.org/10.1109/SP.2017.41

[23] Rieke, N., Hancox, J., Li, W., Milletarì, 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*, Article 119. https://doi.org/10.1038/s41746-020-00323-1

[24] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., & He, B. (2023). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering, 35*(4), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

[25] Siddiqui, M. H. S., Abbasi, M. D., Mir, S. Z., Nadeem, G., Majeed, M. K., Khawer, S. K., et al. (2025). Prediction of maximum air temperature for defining heat wave in various states of Pakistan using machine learning algorithm. *Spectrum of Engineering Sciences*, 942–960.

[26] Khalil, A., Hussain, M., Majeed, M. K., Hamza, A., Ali, A., Ajaz, K., et al. (2025). Artificial intelligence in neuro-oncology: Integrating advanced machine learning techniques for accurate and early detection of brain tumors through MRI imaging. *Spectrum of Engineering Sciences*, 413–435.

[27] Khalil, A., Hussain, A. W., Khan, A. H., Majeed, M. K., Mir, S. Z., Abbasi, M. D., et al. (2025). Transforming heart transplantation with AI: Deep neural networks for predictive analytics and real-time monitoring in clinical decision support systems. *Pakistan Journal of Medical & Cardiological Review, 4*(3), 390–411.

[28] Mir, S. Z., Majeed, M. K., Siddiqui, M. H. S., Khawer, S. K., Younus, I., Shah, S. R., et al. (2025). Integrating AI and IoT in solar power systems for improved monitoring and predictive maintenance. *Spectrum of Engineering Sciences*, 971–982.

[29] Haq, R. U., Aman, F., Majeed, M. A., Raza, S., Khan, A., Hussain, R., & Majeed, M. K. (2025). Developing edge computing solutions for IoT devices to reduce latency and enhance real-time decision-making. *Spectrum of Engineering Sciences*, 961–970.

[30] Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine, 35*(1), 126–136. https://doi.org/10.1109/MSP.2017.2765695

[31] Busia, P., Scrugli, M. A., Jung, V. J.-B., Benini, L., & Meloni, P. (2025). A tiny transformer for low-power arrhythmia classification on microcontrollers. *IEEE Transactions on Biomedical Circuits and Systems, 19*(1), 142–152.

[32] Velichko, A. (2021). *A method for medical data analysis using the LogNNet for clinical decision support systems and edge computing in healthcare* [Preprint]. arXiv. https://arxiv.org/abs/2108.02428

[33] Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., Doppler, J., Holzmann, C., Kurz, M., Holl, G., Chavarriaga, R., Sagha, H., Bayati, H., Creatura, M., & Millán, J. del R. (2010). Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh International Conference on Networked Sensing Systems (INSS)* (pp. 233–240). IEEE. https://doi.org/10.1109/INSS.2010.5573462

[34] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, Article 160035. https://doi.org/10.1038/sdata.2016.35

[35] Reyna, M. A., Amorim, E., Sameni, R., … Clifford, G. D. (2023). Predicting neurological recovery from coma after cardiac arrest: The George B. Moody PhysioNet Challenge 2023. *Computing in Cardiology, 50*.

[36] Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Westover, M. B., Nemati, S., Clifford, G. D., & Sharma, A. (2020). Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine, 48*(2), 210–217. https://doi.org/10.1097/CCM.0000000000004145

[37] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[38] Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Joint Summits on Translational Science Proceedings, 2020*, 191–200.

[39] Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases, 66*(1), 149–153. https://doi.org/10.1093/cid/cix731

[40] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 559–560). ACM. https://doi.org/10.1145/3233547.3233667

International Journal of Advanced Research