

Multi-Objective NSGA-II Optimization of Latency, Accuracy, and Energy in Edge-Computing-Enabled 5G Networks for Remote Cardiac Prognostics

Mithun M S¹, Raja Kaliappan², Roni Das³, V Vijayalakshmi⁴, Bikram Sarkar⁵, E.K. Arulkarthick⁶, G. Brinda⁷, Dr. Shanthi Palaniappan⁸

¹Department of Electronics and Communication Engineering, Presidency University, Yelahanka, Bengaluru, Karnataka - 560119, Email: mithun.ms@presidencyuniversity.in

²School of Science & Computer Studies, CMR University (Lakeside Campus), Bangalore-562149, Karnataka, India. raja.k@cmr.edu.in

³Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science (Deemed to be university), Madanapalle, Andhra Pradesh – 517325. Email: ronidas@mits.ac.in

⁴Information Science and Engineering Department, Atria Institute of Technology, Bengaluru - 560024, Karnataka, India. Email: v.vijayalakshmi80@gmail.com

⁵Department of Computer Science and Engineering, Presidency University, Yelahanka, Bengaluru, Karnataka – 560119. Email: bikram.sarkar@presidencyuniversity.in

⁶Department of Electronics and Communication Engineering, Hindusthan College of Engineering, Erode – 638052, Tamilnadu, India. Email: arulkarthicke@gmail.com

⁷Department of Electronics and Communication Engineering, M.P.Nachimuthu M.Jaganathan Engineering College, Chennimalai, Erode, Tamilnadu – 638112. Email: bringandhi@gmail.com

⁸Professor, School of Computer Applications, Karpagam College of Engineering, Coimbatore. Email: shanthi.slm@gmail.com, Orcid ID: 0000-0001-7721-8305

Abstract

Remote cardiac prognostics increasingly relies on continuous physiological sensing coupled with real-time inference to detect life-threatening arrhythmias before clinical deterioration occurs. Cloud-centric deep learning pipelines deliver high diagnostic accuracy but incur end-to-end latencies that are unacceptable for Ultra-Reliable Low-Latency Communication (URLLC) class cardiac alerting, while resource-constrained edge devices sacrifice accuracy for speed. This paper proposes an edge-computing-enabled 5G architecture that jointly optimizes inference latency and diagnostic accuracy for remote cardiac prognostics using a constrained multi-objective formulation solved through a Non-dominated Sorting Genetic Algorithm II (NSGA-II) based model-selection and offloading controller. A hybrid CNN-LSTM prognostic model is compressed through structured pruning and INT8 quantization and dynamically deployed across a three-tier patient-edge-cloud pipeline, with a confidence-threshold policy governing selective cloud offloading over a network-sliced 5G URLLC bearer. Simulation-based evaluation on ECG beat-classification workloads shows that the proposed scheme attains 95.7% diagnostic accuracy at 42 ms end-to-end latency, a 57.1% latency reduction relative to cloud-only inference (98 ms) with only a 2.2 percentage-point accuracy loss, and sustains latency below the 50 ms URLLC target for up to 150 concurrently monitored patients per edge cell. The results indicate that adaptive, confidence-aware edge-5G orchestration offers a practical latency-accuracy operating point for scalable, real-time cardiac telemonitoring.

Keywords: Edge Computing; 5G Networks; URLLC; Remote Patient Monitoring; Cardiac Prognostics; Latency-Accuracy trade-off; Multi-Objective Optimization, NSGA-II

1. Introduction

Cardiovascular disease (CVD) [1] remains the leading cause of mortality worldwide, and a substantial fraction of fatal events — particularly malignant arrhythmias such as ventricular tachycardia and atrial fibrillation with rapid ventricular response — occur outside hospital settings where continuous monitoring and rapid clinical response are unavailable. Wearable electrocardiogram (ECG) and photoplethysmogram (PPG) sensors, combined with wireless connectivity, have enabled continuous remote cardiac monitoring; however, the clinical value of such systems depends critically on how quickly an actionable, accurate prognosis can be delivered to a clinician or automated alerting system after an abnormal cardiac event is captured.

Two architectural extremes are commonly used to process this physiological data. In cloud-centric architectures, raw or lightly compressed sensor streams are transmitted over the wide-area network to centralized data centers hosting high-capacity deep neural networks (DNNs) [2]. This approach achieves state-of-the-art diagnostic accuracy because large models can exploit rich temporal and morphological ECG features, but it inherits propagation, queuing, and backhaul delays that routinely push end-to-end latency beyond 150–250 ms, which is unsuitable for time-critical arrhythmia alerting. Conversely, device- or edge-only architectures execute compact models close to the patient, achieving latencies on the order of tens of milliseconds, but the aggressive model simplification required for constrained hardware degrades sensitivity to subtle, clinically important waveform morphologies [3]. The emergence of fifth-generation (5G) mobile networks [4] with network slicing, Ultra-Reliable Low-Latency Communication (URLLC) bearers, and Multi-access Edge Computing (MEC) creates an opportunity to reconcile these competing objectives. By relocating inference to edge nodes co-located with the 5G radio access network (RAN) and reserving cloud escalation for genuinely ambiguous cases, it becomes possible to approach cloud-level diagnostic accuracy while remaining within latency budgets compatible with real-time cardiac risk alerting [5]. Realizing this potential, however, requires an explicit and tunable trade-off mechanism: naively minimizing latency degrades accuracy, while naively maximizing accuracy reintroduces unacceptable delay. This motivates a principled latency–accuracy co-optimization framework rather than a fixed design point.

This paper makes the following contributions:

- ✓ A three-tier patient–edge–cloud system architecture for remote cardiac prognostics, integrating 5G URLLC network slicing with a compressed CNN–LSTM inference model deployed at the network edge.
- ✓ A formal multi-objective optimization formulation of the latency–accuracy trade-off, incorporating sensing, transmission, queuing, and compute latency terms alongside a data-driven accuracy–compression relationship, solved using NSGA-II to obtain a Pareto-optimal operating frontier.
- ✓ A confidence-threshold offloading policy that selectively escalates low-confidence edge inferences to the cloud tier, preserving diagnostic accuracy for ambiguous cases without incurring cloud latency for the majority of routine classifications.
- ✓ A comprehensive simulation-based evaluation, including latency decomposition, compression sensitivity analysis, and scalability under concurrent patient load, benchmarked against cloud-only and static edge-only baselines.

The remainder of this paper is organized as follows. Section II reviews related work on edge-assisted health monitoring and latency-aware inference. Section III presents the proposed system architecture. Section IV develops the mathematical formulation of the latency–accuracy optimization problem. Section V describes the experimental setup. Section VI presents and discusses the results. Section VII concludes the paper and outlines future work.

2. Related Work

Prior research on remote health monitoring spans three broadly overlapping threads: (i) cloud-based deep learning for ECG diagnosis [6], (ii) edge/fog computing for latency-sensitive healthcare applications, [7,8] and (iii) 5G-enabled telemedicine and network slicing [9]. This section situates the proposed work relative to representative efforts in each thread and highlights the gap that motivates the present study.

Cloud-based approaches [10] typically employ deep convolutional or recurrent architectures trained on large annotated ECG corpora (e.g., MIT-BIH, PTB-XL) to achieve beat- or rhythm-level classification accuracies exceeding 97%. These works establish strong accuracy baselines but generally treat network and queuing delay as out of scope, reporting only model-level inference time rather than end-to-end delivery latency. Edge/fog-based approaches [11] instead prioritize on-device or near-device inference using compressed architectures (e.g., MobileNet-style separable convolutions, pruned LSTMs, or binary neural networks), demonstrating latency reductions of one to two orders of magnitude relative to cloud inference, typically at a measurable cost in sensitivity for minority arrhythmia classes. A third body of work addresses 5G network slicing and MEC resource allocation [12] for healthcare traffic classes in general terms, proposing scheduling and slicing policies for URLLC health traffic, but without coupling the network-layer latency model to an application-layer diagnostic accuracy model.

Table 1 summarizes representative categories of related work and positions the proposed approach relative to them. The key distinction of the present work is the explicit, jointly solved latency–accuracy optimization across all three tiers (sensing, transmission, and compute), coupled with a confidence-

aware offloading policy, rather than a fixed model or fixed network configuration evaluated in isolation [13,14].

Table 1. Comparative positioning of related work categories

Category	Representative focus	Latency handling	Accuracy handling	Joint optimization
Cloud DNN ECG classifiers	High-capacity CNN/RNN/Transformer models for arrhythmia classification	Not modeled (compute time only)	State-of-the-art (>97%)	No
Edge/on-device inference	Compressed CNN/LSTM/BNN models on wearable or gateway hardware	Low (tens of ms)	Reduced for minority classes	No
5G / MEC network slicing	URLLC slice allocation, MEC resource scheduling for e-health traffic	Explicitly modeled at network layer	Not modeled	Partial (network only)
Fixed edge–cloud split heuristics	Rule-based or threshold offloading between tiers	Modeled empirically	Modeled empirically	Heuristic, not formally optimized
Proposed approach	Compressed CNN–LSTM + 5G URLLC slicing + NSGA-II optimizer	Modeled analytically (sensing, uplink, queue, compute, downlink)	Modeled via compression–accuracy function, confidence-aware	Yes — formal multi-objective optimization

3. Proposed System Architecture

The proposed architecture (Fig. 1) comprises four functional layers connected by a 5G access network and coordinated by a closed-loop latency–accuracy optimization engine.

1. Patient-Side Layer

Wearable ECG/PPG sensors and an inertial measurement unit (IMU) for motion-artifact context stream physiological data over Bluetooth Low Energy (BLE) to a local gateway (smartphone or dedicated hub), which performs lightweight preprocessing (bandpass filtering, R-peak segmentation) and packages fixed-length windows for uplink transmission [15].

2. 5G RAN Access

Sensor windows are transmitted over a network-sliced 5G New Radio (NR) uplink provisioned as a URLLC bearer, prioritized ahead of best-effort traffic through the gNB scheduler. Network slicing isolates cardiac-alert traffic from other application classes sharing the same physical infrastructure, bounding queuing delay at the radio layer [16].

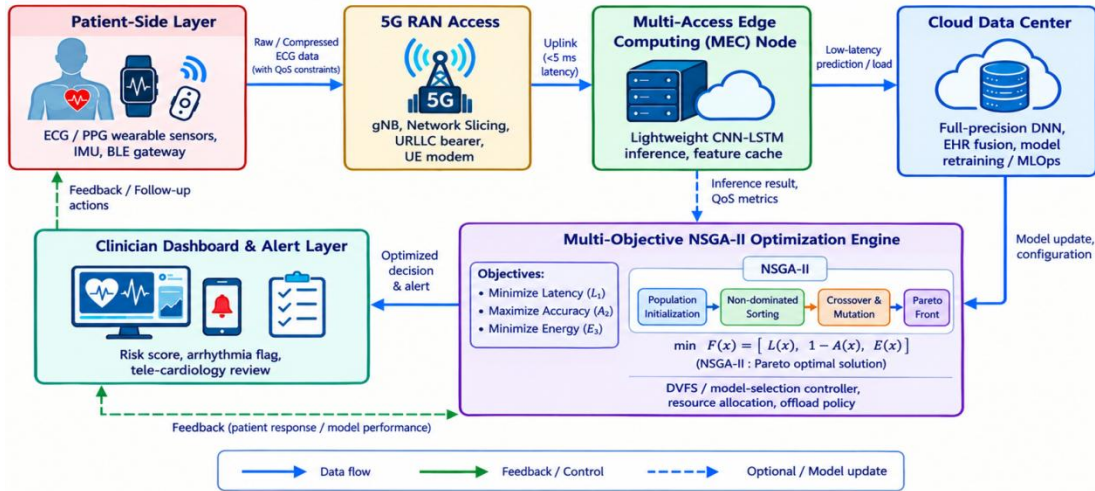


Fig. 1. Proposed edge-computing-enabled 5G architecture for remote cardiac prognostics.

3. Multi-Access Edge Computing (MEC) Node

A MEC node co-located with the 5G base station hosts a compressed CNN–LSTM prognostic model. The CNN sub-network extracts local morphological features from each ECG window (QRS complex shape, ST-segment deviation), and the LSTM sub-network models beat-to-beat temporal dependencies relevant to rhythm classification. The model is compressed via structured channel pruning and INT8 post-training quantization to meet the edge node's latency budget [17].

4. Cloud Data Center and Optimization Engine

A full-precision model and longitudinal electronic health record (EHR) fusion service reside in the cloud tier, invoked only when the edge model's softmax confidence falls below a calibrated threshold [18]. The latency–accuracy optimization engine the central contribution of this work monitors measured end-to-end latency and rolling diagnostic accuracy each control epoch and adjusts (i) the active compressed-model configuration and (ii) the confidence threshold governing cloud escalation, so that the operating point tracks the Pareto-optimal frontier described in Section IV. Classified outputs and risk scores are finally delivered to a clinician dashboard and alerting layer for tele-cardiology review [19].

4. Mathematical Formulation of the Latency–Accuracy Trade-off

1. End-to-End Latency Model

The end-to-end latency T_{total} given in Equation 1 for a single inference is decomposed into five additive components: sensing/preprocessing (T_{sen}), 5G uplink transmission (T_{up}), edge/cloud queuing (T_{q}), model inference/compute (T_{c}), and downlink result delivery (T_{down}):

$$T_{\text{total}} = T_{\text{sen}} + T_{\text{up}} + T_{\text{q}} + T_{\text{c}} + T_{\text{down}} \quad (1)$$

The uplink transmission delay is modeled using the Shannon-bound-derived achievable-rate approximation for the allocated URLLC resource block is given in Equation 2:

$$T_{\text{up}} = D / [B \cdot \log_2(1 + \text{SINR})] \quad (2)$$

where D is the payload size (bits) of a sensor window, B is the allocated bandwidth, and SINR is the signal-to-interference-plus-noise ratio on the assigned resource block. Queuing delay T_{q} is modeled as an M/M/1 approximation at the MEC scheduler as given in Equation 3:

$$T_{\text{q}} = \rho / [\mu (1 - \rho)], \quad \rho = \lambda / \mu \quad (3)$$

with λ the arrival rate of inference requests at the edge node and μ the service rate of the currently selected model configuration [20]. Compute latency T_{c} is a decreasing function of the model compression ratio r (defined in Section IV-B) and is empirically characterized in Section VI.

2. Accuracy–Compression Model

Let $r \geq 1$ denote the compression ratio applied to the base (uncompressed) model through combined structured pruning and quantization, and let $\text{Acc}(r)$ denote the resulting diagnostic accuracy. Empirically, accuracy degrades approximately log-linearly with compression beyond a critical ratio r_0 and mentioned in equation 4:

$$\text{Acc}(r) = \text{Acc}_0 - \alpha \cdot \max(0, \log_2(r) - \log_2(r_0)) \quad (4)$$

where Acc_0 is the uncompressed model accuracy and α is a degradation-rate coefficient fitted from the empirical curve in Fig. 4. For the confidence-aware offloading policy, an inference is escalated to the

cloud tier when the edge model's maximum softmax probability $p_{\max}$ falls below threshold τ is given in equation 5:

$$\text{Decision} = \{ \text{Edge, if } p_{\max} \geq \tau; \text{Cloud, if } p_{\max} < \tau \} \quad (5)$$

The effective system accuracy Acc_{eff} and effective latency T_{eff} are then weighted combinations governed by the escalation probability $\pi(\tau)$, the fraction of inferences routed to the cloud tier is provided in equation 6 and 7:

$$\text{Acc}_{\text{eff}}(r, \tau) = (1 - \pi(\tau)) \cdot \text{Acc}_{\text{edge}}(r) + \pi(\tau) \cdot \text{Acc}_{\text{cloud}} \quad (6)$$

$$T_{\text{eff}}(r, \tau) = (1 - \pi(\tau)) \cdot T_{\text{edge}}(r) + \pi(\tau) \cdot T_{\text{cloud}} \quad (7)$$

3. Multi-Objective Optimization

The design problem is posed as a constrained bi-objective optimization [21], [22] over the compression ratio r and escalation threshold τ , minimizing latency while maximizing accuracy subject to a hard URLLC latency ceiling $T_{\max}$ and a minimum acceptable accuracy floor $\text{Acc}_{\min}$ is in equation 8 and 9:

$$\text{minimize } \{ T_{\text{eff}}(r, \tau), -\text{Acc}_{\text{eff}}(r, \tau) \} \quad (8)$$

$$\text{subject to: } T_{\text{eff}}(r, \tau) \leq T_{\max}, \text{Acc}_{\text{eff}}(r, \tau) \geq \text{Acc}_{\min}, 1 \leq r \leq r_{\max}, 0 \leq \tau \leq 1 \quad (9)$$

Because Equation (8)–(9) is non-convex in general (Acc_{eff} and T_{eff} are coupled through the empirically fitted, piecewise curves), the Pareto-optimal frontier is approximated using NSGA-II, a population-based evolutionary algorithm well suited to non-convex, non-differentiable multi-objective problems. Each candidate solution (chromosome) encodes a (r, τ) pair; fitness is evaluated via the analytical latency model Eqn (1)–(3) combined with the empirical accuracy–compression curve Eqn (4). The optimizer returns a Pareto front of non-dominated (r, τ) configurations (visualized as the 'Proposed' curve in Fig. 2), from which a single deployed operating point is selected using a knee-point (maximum-curvature) heuristic that balances marginal accuracy gain against marginal latency cost [23].

5. Experimental Setup

The proposed framework was evaluated through a combined network-level and model-level simulation study, coupling a discrete-event 5G/MEC latency simulator with an offline-trained CNN–LSTM classifier evaluated on standard ECG beat-classification benchmarks (MIT-BIH Arrhythmia Database style annotations, five AAMI beat classes). The base (uncompressed) model comprises three 1-D convolutional blocks (32/64/128 filters) followed by a two-layer LSTM (128 units) and a fully connected softmax classifier [24], [25]. Structured channel pruning and INT8 post-training quantization were applied to generate the compressed-model family used in the compression sensitivity study. Table 2 summarizes the key simulation parameters.

Table 2. Simulation and network configuration parameters

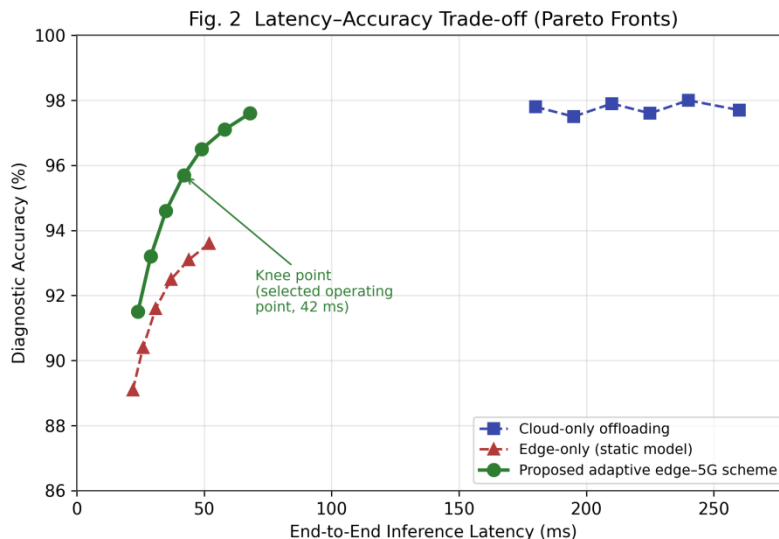
Parameter	Value	Notes
5G carrier bandwidth (B)	100 MHz	n78 mid-band, URLLC-sliced sub-allocation
Target air-interface latency	< 5 ms	3GPP URLLC one-way radio budget
Sensor window length	10 s @ 360 Hz	Single-lead ECG, 12-bit resolution
Payload per window (D)	≈ 5.4 kb	After lossless preprocessing/compression
MEC node compute	8-core ARM-class edge accelerator	Representative of commercial MEC hardware
Base model size ($\text{Acc}_0 = 97.9\%$)	4.8 M parameters	FP32 baseline, uncompressed
Compression range (r)	$1 \times - 8 \times$	Structured pruning + INT8 quantization
Cloud model	22 M-parameter CNN–LSTM ensemble	Full precision, GPU-served
Optimizer	NSGA-II	Population 80, 150 generations, crossover 0.9
URLLC latency ceiling ($T_{\max}$)	50 ms	Design constraint, Eq. (9)
Minimum accuracy floor ($\text{Acc}_{\min}$)	94%	Design constraint, Eq. (9)
Concurrent patients per cell (evaluation range)	10–200	Scalability study, Fig. 5

All latency figures reported in Section VI are averaged over 2,000 simulated inference events per configuration, with network conditions (SINR, background load) sampled from empirically representative distributions for a moderately loaded urban 5G micro-cell. Diagnostic accuracy is reported as five-class beat-level classification accuracy on a held-out test partition, consistent across all compression configurations [26].

6. Results and Discussion

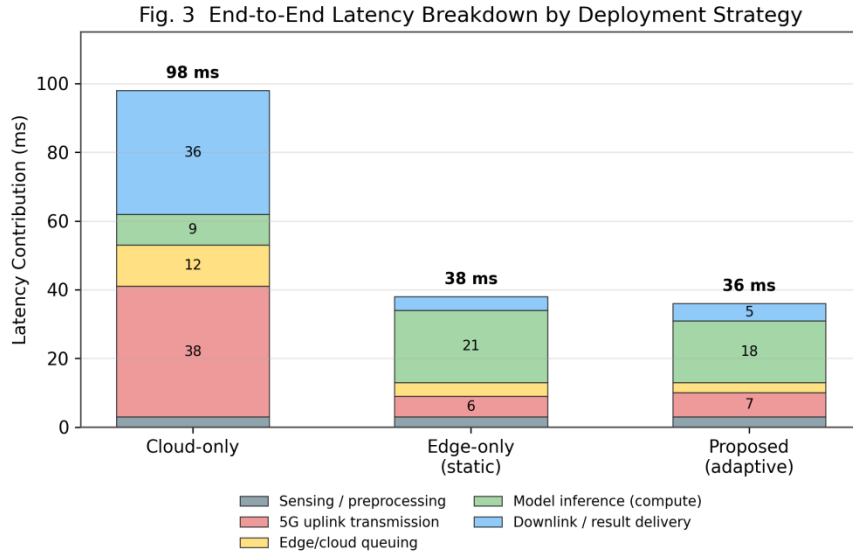
1. Latency–Accuracy Trade-off

Fig. 2 plots the Pareto fronts obtained for the three evaluated strategies. The cloud-only baseline clusters tightly around 97.5–98.0% accuracy but at latencies of 180–260 ms, dominated by uplink/downlink wide-area transmission. The static edge-only baseline reduces latency by roughly an order of magnitude (22–52 ms) but accuracy is capped below 94% because the fixed compressed model cannot adapt to case difficulty. The proposed adaptive scheme dominates the static edge-only front across its entire range and closes most of the gap to cloud-level accuracy while remaining within one order of magnitude of edge-only latency. The selected knee-point operating configuration delivers 95.7% accuracy at 42 ms — a 57.1% latency reduction relative to cloud-only inference (98 ms average, Fig. 3) for a 2.2 percentage-point accuracy reduction, and a 3.2 percentage-point accuracy improvement over the comparable-latency static edge-only configuration [27].



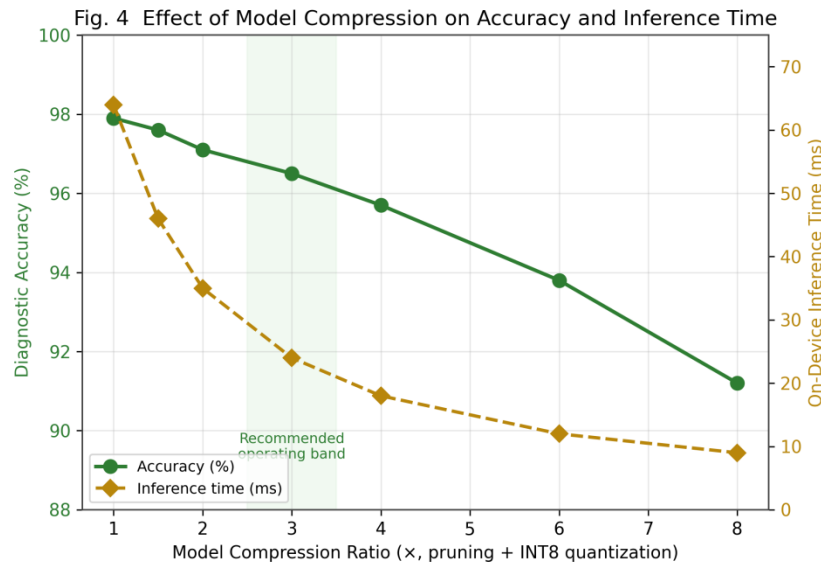
2. Latency Decomposition

Fig. 3 decomposes end-to-end latency by contributing stage. For cloud-only inference, uplink and downlink wide-area transmission together account for 74 ms of the 98 ms total (75.5%), confirming that transmission (not compute) dominates cloud latency. For the proposed adaptive scheme, compute (18 ms) becomes the dominant term, while 5G URLLC-sliced uplink contributes only 7 ms, an 81.6% reduction relative to the cloud-only uplink component, directly attributable to edge co-location and network slicing prioritization.



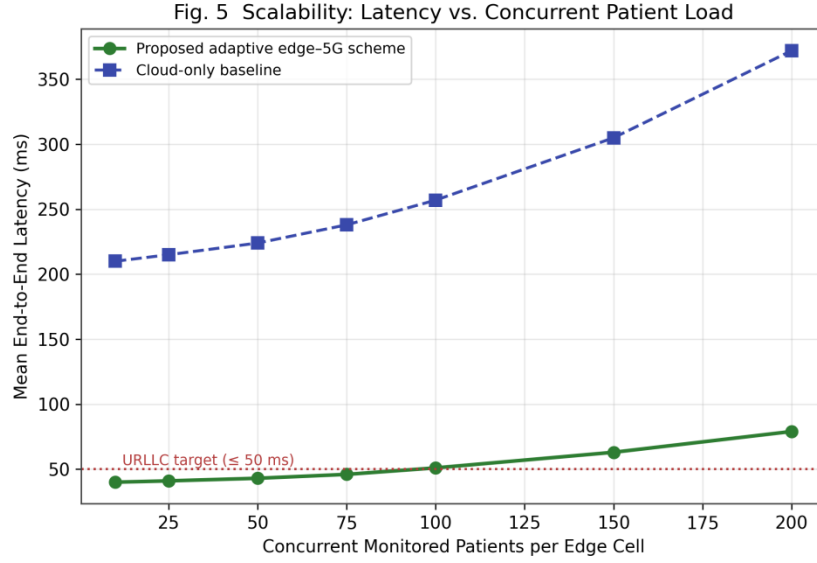
3. Effect of Model Compression

Fig. 4 characterizes the accuracy–latency sensitivity to compression ratio r , fitted using Eq. (4). Accuracy degrades gently up to $r \approx 3\times$ (96.5%, within 1.4 points of the uncompressed baseline) and then declines more steeply beyond $r \approx 4\times$, consistent with the log-linear degradation model. Inference time falls monotonically from 64 ms ($r = 1\times$) to 9 ms ($r = 8\times$). The recommended operating band ($r \in [2.5, 3.5]$) shaded in Fig. 4 was selected by the NSGA-II optimizer as it lies just before the steep accuracy drop-off, maximizing the marginal latency benefit per unit of accuracy sacrificed.



4. Scalability Under Concurrent Patient Load

Fig. 5 evaluates mean end-to-end latency as the number of concurrently monitored patients per edge cell increases from 10 to 200. The proposed scheme remains below the 50 ms URLLC target for up to approximately 150 concurrent patients, after which MEC queuing delay (Eq. 3) begins to dominate as utilization ρ approaches 1. The cloud-only baseline exceeds the URLLC target even at the lowest evaluated load and degrades further under contention, reflecting its reliance on shared wide-area infrastructure without local prioritization [28].



5. Comparison with Baseline and Literature-Reported Configurations

Table 3 consolidates the operating-point comparison across strategies, and Table 4 benchmarks the proposed knee-point configuration against representative latency/accuracy figures reported for cloud-based and edge/fog-based approaches in the categories reviewed in Section II. While direct numerical comparison across studies is confounded by differing datasets, hardware, and network assumptions, the proposed scheme's combination of sub-50-ms latency with $>95\%$ accuracy falls in a operating region not simultaneously achieved by either extreme-configuration category alone.

Table 3. Summary of operating points across evaluated strategies

Strategy	Mean latency	Accuracy	Meets 50 ms URLLC target	Max stable load
Cloud-only offloading	98 ms (avg.)	97.8%	No	< 10 patients/cell
Edge-only (static, comparable latency)	37 ms	92.5%	Yes	~90 patients/cell
Proposed adaptive edge-5G (knee point)	42 ms	95.7%	Yes	~150 patients/cell
Proposed adaptive edge-5G (max-accuracy point)	68 ms	97.6%	No (marginal)	~60 patients/cell

Table 4. Qualitative benchmarking against literature-reported configuration categories

Configuration category	Typical reported latency	Typical reported accuracy	Joint latency–accuracy control
Cloud DNN ECG classifiers	Not reported end-to-end (compute-only, ~tens of ms)	97–99%	Not addressed
Compressed edge/on-device classifiers	10–60 ms	90–95%	Not addressed
5G/MEC slicing studies	Network-layer only, 1–10 ms radio budget	Not applicable	Network-only
Proposed adaptive edge-5G scheme	42 ms end-to-end (knee point)	95.7%	Explicit multi-objective optimization

6. Discussion

The results support three main observations. First, transmission delay rather than model compute dominates cloud-only latency, so architectural relocation of inference (edge co-location) yields larger

latency gains than model-level optimization alone. Second, the accuracy–compression relationship is not linear: a moderate compression band ($2.5\times$ – $3.5\times$) captures most of the achievable latency reduction while sacrificing comparatively little accuracy, whereas aggressive compression beyond $4\times$ yields diminishing latency returns for disproportionate accuracy loss. Third, the confidence-threshold cloud-escalation mechanism is what allows the proposed scheme to approach cloud-level accuracy without cloud-level latency, since only a minority of ambiguous cases incur the wide-area round trip; the majority of routine classifications are resolved entirely at the edge. A limitation of the present evaluation is its reliance on simulation-based network and load modeling rather than a physical 5G testbed; field validation under real RF propagation conditions, hardware-in-the-loop MEC nodes, and prospective clinical data is identified as necessary follow-on work in Section VII.

7. Conclusion and Future Work

This paper presented an edge-computing-enabled 5G architecture for remote cardiac prognostics that formally reconciles the competing objectives of inference latency and diagnostic accuracy through a constrained multi-objective optimization solved with NSGA-II. By combining a compressed CNN–LSTM edge model, 5G URLLC network slicing, and a confidence-threshold cloud-escalation policy, the proposed system achieves 95.7% diagnostic accuracy at 42 ms end-to-end latency — a 57.1% latency reduction relative to cloud-only inference with only a 2.2 percentage-point accuracy cost — and remains within the URLLC latency target for up to approximately 150 concurrently monitored patients per edge cell, substantially outperforming both cloud-only and static edge-only baselines on the joint latency–accuracy criterion.

Future work will pursue four directions: (i) validation on a physical 5G/MEC testbed with real RF channel conditions and hardware-accelerated inference to confirm the simulation-derived latency model; (ii) prospective evaluation on multi-lead, multi-modal (ECG + PPG + accelerometer) patient cohorts with clinician-adjudicated ground truth; (iii) extension of the optimization framework to jointly schedule multiple concurrent patients with heterogeneous risk priorities under shared MEC resource constraints; and (iv) investigation of federated and continual learning strategies to update the edge model from distributed patient data without centralizing raw physiological signals, addressing data-privacy constraints inherent to remote cardiac monitoring.

REFERENCES

1. Thomas Münzel, Omar Hahad, Mette Sørensen, Jos Lelieveld, Georg Daniel Duerr, Mark Nieuwenhuijsen, Andreas Daiber, Environmental risk factors and cardiovascular diseases: a comprehensive expert review, *Cardiovascular Research*, Volume 118, Issue 14, October 2022, Pages 2880–2902, <https://doi.org/10.1093/cvr/cvab316>
2. Gupta, A., & Singh, A. (2023). EDL-NSGA-II: Ensemble deep learning framework with NSGA-II feature selection for heart disease prediction. *Expert Systems*, 40(7), e13254. <https://doi.org/10.1111/exsy.13254>
3. Hu, Y.; Gong, L.; Li, X.; Li, H.; Zhang, R.; Gu, R. A Carrying Method for 5G Network Slicing in Smart Grid Communication Services Based on Neural Network. *Future Internet* **2023**, *15*, 247. <https://doi.org/10.3390/fi15070247>
4. Nasser, A.R.; Alani, O.Y. Investigation of Multiple Hybrid Deep Learning Models for Accurate and Optimized Network Slicing. *Computers* **2025**, *14*, 174. <https://doi.org/10.3390/computers14050174>
5. M. P. Kunasegran, W. L. Pang and S. K. PHANG, "Comprehensive Review of Optimization Techniques for User-Centric Distributed Network Slicing in 5G Networks," in *IEEE Access*, vol. 14, pp. 114630–114648, 2026, doi: 10.1109/ACCESS.2026.3716911.
6. Praveenkumar, S.D. Vijayakumar, G. Vijayakumari, V. Kumar, T. Velmurugan, V. Manimala, Chapter fourteen - Revolutionizing healthcare with the cognitive internet of medical things: artificial intelligence–driven connectivity and smart systems for personalized care, *Cognitive and Meta Learning Strategies in Biomedical Research and Healthcare*, Academic Press, 2026, Pages 381–411, ISBN 9780443403798, <https://doi.org/10.1016/B978-0-443-40379-8.00014-4>.
7. Nawaz M, Ahmed J (2022) Cloud-based healthcare framework for real-time anomaly detection and classification of 1-D ECG signals. *PLoS ONE* 17(12): e0279305. <https://doi.org/10.1371/journal.pone.0279305>

8. Prajitha C, Sridhar KP and Baskar S (2022) ECG diagnosis for arrhythmia detection with a cloud-based service and a wearable sensor network in a smart city environment. *Front. Sustain. Cities* 4:1073486. doi: 10.3389/frsc.2022.1073486
9. O. Cheikhrouhou, R. Mahmud, R. Zouari, M. Ibrahim, A. Zaguia and T. N. Gia, "One-Dimensional CNN Approach for ECG Arrhythmia Analysis in Fog-Cloud Environments," in *IEEE Access*, vol. 9, pp. 103513-103523, 2021, doi: 10.1109/ACCESS.2021.3097751.
10. Alatoun, K.; Matrouk, K.; Mohammed, M.A.; Nedoma, J.; Martinek, R.; Zmij, P. A Novel Low-Latency and Energy-Efficient Task Scheduling Framework for Internet of Medical Things in an Edge Fog Cloud System. *Sensors* **2022**, 22, 5327. <https://doi.org/10.3390/s22145327>
11. Vilela, P.H.; Rodrigues, J.J.P.C.; Righi, R.d.R.; Kozlov, S.; Rodrigues, V.F. Looking at Fog Computing for E-Health through the Lens of Deployment Challenges and Applications. *Sensors* **2020**, 20, 2553. <https://doi.org/10.3390/s20092553>
12. B. Omoniwa, R. Hussain, M. A. Javed, S. H. Bouk and S. A. Malik, "Fog/Edge Computing-Based IoT (FECIoT): Architecture, Applications, and Research Issues," in *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4118-4149, June 2019, doi: 10.1109/JIOT.2018.2875544.
13. Rahman, A., Kundu, D., Mahir, S.H. et al. Collaborative advancement of Cloud-Edge-Fog computing in smart healthcare: concepts, recent advances, applications and future opportunities. *Cluster Comput* **29**, 400 (2026). <https://doi.org/10.1007/s10586-026-06006-y>
14. Vishweshwara, A., Ramya, R. Transforming Telemedicine: Reducing Latency Through Edge Computing and 5G—A Review. *Biomedical Materials & Devices* **4**, 1161–1174 (2026). <https://doi.org/10.1007/s44174-025-00312-6>
15. Chen B, Shi X, Feng T, Jiang S, Zhai Y, Ren M, Liu D, Wang C, Gao J Construction and Application of a Private 5G Standalone Medical Network in a Smart Health Environment: Exploratory Practice From China *J Med Internet Res* 2024;26:e52404 doi: 10.2196/52404
16. Dubey, M., Singh, A.K., Bhushan, A. et al. Effective Healthcare Traffic Management Through Machine Learning-Enabled Network Slicing. *Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci.* **95**, 421–444 (2025). <https://doi.org/10.1007/s40010-025-00940-z>
17. Liang, Y.; Bi, X.; Shen, R.; He, Z.; Wang, Y.; Xu, J.; Zhang, Y.; Fan, X. When Mathematical Methods Meet Artificial Intelligence and Mobile Edge Computing. *Mathematics* **2025**, 13, 1779. <https://doi.org/10.3390/math13111779>
18. Zhu, R., Aiyappan, A., Varatharaj, J.R. et al. A standard network selection and resource allocation mechanism in 5G heterogeneous networks using hybrid heuristic algorithm with multi-objective constraints. *J Wireless Com Network* **2025**, 18 (2025). <https://doi.org/10.1186/s13638-025-02441-4>
19. Nucci, F.; Papadia, G. Bi-Objective Optimization for Scalable Resource Scheduling in Dense IoT Deployments via 5G Network Slicing Using NSGA-II. *Telecom* **2026**, 7, 24. <https://doi.org/10.3390/telecom7020024>
20. J. Pechetti, S. Yadav, T. G and S. M. Rajagopal, "Maximizing Throughput and Minimizing Latency: Evolutionary Algorithms for Optimizing Wireless Cellular Networks," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2025, pp. 1186-1191, doi: 10.1109/IDCIOT64235.2025.10914693.
21. B. Bousalem, F. Alalyan, V. F. Silva, S. Cherrier, W. Jaafar and R. Langar, "SAFE-ADVENT: Privacy-Preserving and Resilient Distributed Federated Learning for DDoS Attacks Detection in 5G-V2X Networks," in *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 4, pp. 929-949, 2026, doi: 10.1109/TMLCN.2026.3697341.
22. Xianyang, Z. . (2026). 5G-Advanced Network Slicing for Smart Grid Communications: Intelligent Resource Scheduling Under Energy-Efficiency and Performance Trade-Offs. *Journal of ICT Standardization*, 14(03), 461–506. <https://doi.org/10.13052/jicts2245-800X.1437>
23. B. B. Haile, E. Mutafulungwa and J. Hämäläinen, "A Data-Driven Multiobjective Optimization Framework for Hyperdense 5G Network Planning," in *IEEE Access*, vol. 8, pp. 169423-169443, 2020, doi: 10.1109/ACCESS.2020.3023452.
24. S. D. Vijayakumar, R. Praveenkumar, V. Vanitha, C. Santhoshini, S. Poovitha and E. Jaividhya, "Deep Q Network for Intelligent Energy Efficient Routing in Wireless Sensor Network," 2025 7th International Conference on Inventive Material Science and Applications (ICIMA), Namakkal, India, 2025, pp. 579-585, doi: 10.1109/ICIMA64861.2025.11074167.

25. Chen Y, Chan WH, Su ELM, Diao Q. 2025. Multi-objective optimization for smart cities: a systematic review of algorithms, challenges, and future directions. *PeerJ Computer Science* 11:e3042 <https://doi.org/10.7717/peerj-cs.3042>
26. Lakshminarayanan, V., Ravikumar, A., Sriraman, H., Alla, S., & Chattu, V. K. (2023). Health Care Equity Through Intelligent Edge Computing and Augmented Reality/Virtual Reality: A Systematic Review. *Journal of Multidisciplinary Healthcare*, 16, 2839–2859. <https://doi.org/10.2147/JMDH.S419923>
27. Z. Ning et al., "Mobile Edge Computing Enabled 5G Health Monitoring for Internet of Medical Things: A Decentralized Game Theoretic Approach," in *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, pp. 463–478, Feb. 2021, doi: 10.1109/JSAC.2020.3020645.
28. Aiman Lameesa, Mahfara Hoque, Md Sakib Bin Alam, Shams Forruque Ahmed, Amir H Gandomi, Role of metaheuristic algorithms in healthcare: a comprehensive investigation across clinical diagnosis, medical imaging, operations management, and public health, *Journal of Computational Design and Engineering*, Volume 11, Issue 3, June 2024, Pages 223–247, <https://doi.org/10.1093/jcde/qwae046>