

Sustainable Telemedicine: Low-Energy Edge AI and Green Data Center Routing for National Rollout[†]

Wai San Leong¹ and Wai Yie Leong^{2,*} 

¹ Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia; berniceleong2000@gmail.com

² Faculty of Engineering and Quantity Surveying, INTI International University, Nilai 71800, Malaysia

* Correspondence: waiyie.leong@newinti.edu.my

[†] Presented at the 7th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2025 (ECBIOS 2025), Kaohsiung, Taiwan, 23–25 October 2025.

Abstract

Telemedicine at the national scale must balance clinical quality, privacy, latency, and sustainability. This study aims to develop a system architecture and methodology for low-energy edge AI combined with green data center routing to reduce energy per consultation while maintaining clinical-grade performance. The results present (1) an energy-aware edge inference stack for physiological sensing and video triage; (2) a carbon-aware, service level agreement (SAL)-constrained routing strategy across regional data centers using software-defined networking and dynamic workload placement; (3) a techno-environmental methodology linking patient-level service key performance indexes to energy neutrality factor, grams CO₂e per encounter, and latency–reliability envelopes; and (4) national rollout playbooks covering network tiers (household/clinic/edge/cloud), facilities upgrades, and governance. Scenarios in urban, peri-urban, and rural/remote environments show 37–62% energy savings and 28–49% carbon reductions relative to cloud-only baselines, with median end-to-end latency ≤ 120 ms for triage and ≤ 40 ms for vitals alarms, meeting the World Health Organization and the International Telecommunication Union latency expectations for eHealth. Trade-offs, risks (drift, network volatility), and policy levers (green SLAs, data residency, open standards) are evaluated to scale sustainable telemedicine without compromising safety or equity.

Keywords: telemedicine; edge AI; sustainable computing; green routing

1. Introduction

The rapid expansion of telemedicine has become one of the most significant transformations in modern healthcare delivery, driven by advances in digital connectivity, AI, and the urgent need for scalable, remote health services. Telemedicine enables virtual consultations and remote monitoring and serves as a critical tool in addressing healthcare access disparities across urban, peri-urban, and rural contexts. The COVID-19 pandemic acted as a catalyst for global telehealth adoption, accelerating digital health strategies and embedding remote care models into mainstream healthcare systems [1]. However, the sustainability of this transformation remains a pressing concern, as the energy and carbon footprints of telemedicine infrastructures, ranging from wearable sensors to data centers, pose challenges to climate commitments and healthcare equity [2].

The infrastructure supporting telemedicine consists of heterogeneous components, including edge devices, mobile applications, regional data centers, and national cloud



Academic Editors: Teen-Hang Meen, Cheng-Yi Chen and Cheng-Fu Yang

Published: 28 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

platforms. Each tier consumes energy while contributing to patient safety and care quality. Traditional cloud-only telemedicine platforms, while offering scalability, often create bottlenecks in terms of latency, bandwidth, and energy efficiency. For example, continuous video streaming and real-time analytics impose heavy demands on networks and data centers, leading to increased operational costs and greenhouse gas (GHG) emissions [3,4]. This energy burden is particularly problematic in developing nations where power infrastructure is fragile, cooling requirements are high due to tropical climates, and renewable penetration is uneven [5,6]. Consequently, sustainable approaches that combine low-energy edge AI and green data center routing are essential for national telemedicine rollout.

Recent advances in edge AI and tiny machine learning (TinyML) offer new opportunities to shift computation closer to the patient, reducing both backhaul energy consumption and latency. TinyML frameworks, leveraging quantization and pruning, allow microcontrollers and mobile processors to execute clinically relevant inference tasks such as arrhythmia detection or anomaly detection in physiological signals at milliwatt levels [5,6]. Deploying such capabilities at the patient or clinic level reduces reliance on continuous cloud connectivity and enhances resilience in low-bandwidth or outage-prone regions. At the same time, the evolution of green data center technologies, including renewable energy integration, improved power usage effectiveness (PUE), and software-defined networking (SDN)-enabled carbon-aware routing, has made it feasible to minimize the environmental impact of high-volume telemedicine workloads [7–9].

The sustainability challenge also intersects with service-level requirements. Clinical safety requires strict latency–reliability envelopes for alarm systems, remote monitoring, and live consultations. For instance, vitals monitoring often demands sub-50 ms latencies for alarms, while triage tasks should remain within 150 ms to ensure timely interventions [10,11]. Meeting these stringent targets in a carbon-optimized manner requires dynamic orchestration that balances latency and energy against clinical safety. Studies in carbon-aware routing demonstrate that workload shifting to renewable-rich data centers can reduce GHG emissions by up to 40%, but without proper SLA enforcement, latency-sensitive applications may suffer [12,13]. Hence, a multi-objective optimization framework that integrates both sustainability and clinical constraints is needed.

The national rollout of telemedicine adds further complexity. Policy frameworks such as the WHO Global Strategy on Digital Health 2020–2025 highlight the importance of equity, interoperability, and sustainability in scaling digital health [5]. Standards, including Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR) [8] and digital imaging and communications in medicine (DICOM) [9], ensure interoperability, while data residency laws require sensitive health information to remain within sovereign jurisdictions, influencing routing decisions and facility siting. Moreover, International Organization for Standardization (ISO) standards such as ISO 13485 and ISO 14971 mandate risk management and quality systems for medical devices and digital health tools [13,14]. A successful rollout must integrate these regulatory, environmental, and clinical dimensions into a unified architecture.

To address the challenges, a sustainable telemedicine system is developed by integrating low-energy edge AI with green data center routing under national constraints. The developed system reduces energy per consultation while maintaining clinical-grade performance, aligns telemedicine operations with national carbon neutrality and environmental, social, and governance (ESG) commitments, and ensures equitable access for rural and underserved populations. By introducing metrics such as energy per encounter (EPE), carbon per encounter (CPE), and energy neutrality factor (ENF), the system provides quantifiable sustainability indicators linked directly to patient-level services. The system meets both healthcare access goals and national climate commitments simultaneously.

2. Literature Review

The resilience of health facilities has been a longstanding concern in disaster risk reduction frameworks. Telemedicine has evolved from niche pilot projects into a core pillar of national healthcare delivery, especially in the aftermath of COVID-19. The World Health Organization (WHO) highlights digital health as a critical enabler for strengthening health systems, extending access to underserved regions, and improving resilience during crises [5]. Research indicates that telemedicine reduces travel costs, enhances continuity of care, and improves patient satisfaction, particularly in chronic disease management and mental health services [6]. However, most implementations have focused on scaling access and safety rather than the sustainability dimension, leaving a gap between health outcomes and environmental goals [15].

The digital infrastructure underpinning telemedicine, comprising wearables, mobile devices, wireless networks, and cloud data centers, consumes significant amounts of energy. A recent survey of information and communication technology (ICT) energy use demonstrates that data centers and telecommunication networks together account for 2–3% of global electricity demand, with projections rising under increased video conferencing and remote health workloads [15]. Cloud-centric telemedicine platforms exacerbate this trend by requiring continuous uplink and downlink traffic, especially for high-resolution imaging and video consults. The associated carbon footprint is non-trivial, particularly in regions dependent on fossil-heavy grids [10]. This underscores the need for carbon-aware design in digital health infrastructure.

Edge computing offers a promising approach to reducing network and data center loads by moving computation closer to the patient. In telemedicine, mobile edge computing (MEC) enables real-time inference at hospitals or local clinics, while wearable devices equipped with TinyML can pre-process or filter physiological signals before transmission [1,4]. For example, studies using lightweight convolutional neural networks (CNNs) and recurrent neural networks have shown accurate detection of cardiac arrhythmias on microcontrollers with sub-50 mW power budgets [1]. Event-driven architectures, where data is transmitted only when anomalies are detected, further reduce backhaul energy and enhance privacy by limiting the volume of transmitted data [2]. Edge AI also mitigates latency concerns, allowing faster responses in safety-critical applications such as falls detection, neonatal monitoring, or remote surgical guidance [15].

TinyML, such as the common microcontroller software interface standard (CMSIS-NN) [4] and TensorFlow Lite Micro, enables the deployment of AI models on resource-constrained devices, enabling inference directly on microcontrollers embedded in wearables or clinic-level gateways. Quantization and pruning techniques have been demonstrated to compress models without significant loss in accuracy, often reducing energy per inference by more than 50% [1,4]. For healthcare applications, models such as EfficientNet and MobileNet have been optimized for edge deployment while retaining diagnostic performance [2,3]. The literature further highlights the importance of federated learning in telemedicine to allow continual model updates without centralized data transfer, thereby preserving privacy and reducing bandwidth consumption [16,17].

Beyond the edge, regional and national data centers form the backbone of large-scale telemedicine services. Traditional data centers are notorious for their energy intensity, but innovations in power usage effectiveness (PUE), free-air cooling, and renewable energy procurement have led to measurable reductions in carbon emissions [10,11]. Carbon-aware routing strategies take this further by dynamically directing workloads toward data centers with higher renewable penetration or lower carbon intensity. For example, Xu et al. demonstrated that shifting flexible tasks based on grid carbon signals can achieve up to 30–40% carbon reduction without major service degradation [18]. In networking, software-

defined networking (SDN) enables energy-aware traffic engineering, dynamically adapting routes based on utilization and carbon metrics [19,20]. However, most existing research has focused on cloud or enterprise workloads, with limited application in telemedicine contexts where latency and safety constraints are stricter.

Clinical safety imposes non-negotiable requirements on latency and reliability. ITU guidelines stipulate that eHealth applications must maintain end-to-end latency thresholds (≤ 150 ms for teleconsultations, ≤ 50 ms for alarms) with high reliability to ensure safe delivery of critical information [15]. Research in low-latency networking, such as low-latency, low-loss, and scalable throughput (L4S) [20], shows potential in reducing jitter for telemedicine workloads, but integrating such capabilities with carbon-aware routing remains underexplored. The literature suggests a clear tension between latency-optimal and carbon-optimal routing; resolving this requires multi-objective optimization frameworks where service-level agreements (SLAs) explicitly encode both latency and sustainability objectives [19,21].

Adoption of sustainable telemedicine architectures must align with stringent regulatory frameworks. Standards such as HL7 FHIR [8] and DICOM [9] govern data interoperability, while ISO 13485 and ISO 14971 ensure compliance with medical device quality and risk management requirements [13,14]. Additionally, national data residency policies and general data protection regulation-like privacy laws mandate that sensitive patient data be stored and processed within sovereign boundaries [12]. The literature emphasizes that sustainability strategies must be embedded within such governance frameworks to avoid regulatory conflicts. For instance, routing workloads to the greenest data center may not always be permissible if data residency laws prohibit cross-border transfers [21].

Despite progress in edge AI, TinyML, and green data centers, the literature reveals several gaps. First, few studies provide integrated, end-to-end models that link device-level energy savings with system-level carbon reductions in telemedicine. Second, sustainability metrics such as EPE or CPE remain largely absent in telehealth studies, which typically prioritize clinical performance alone. Third, while carbon-aware routing and SDN-based traffic engineering are well studied in enterprise networks, their adaptation to latency-critical healthcare services remains limited [18,21]. This study addresses these gaps by presenting a holistic framework that integrates low-power edge inference, carbon-aware data center routing, and SLA-driven orchestration for national telemedicine rollouts.

3. Methodology

The methodology adopted in this study integrates low-energy edge AI with green data center routing into a unified telemedicine framework designed for national rollout (Figure 1). This approach is structured into (1) energy and carbon accounting models, (2) service-level latency and reliability envelopes, (3) edge AI model design and deployment, (4) green data center routing through SDN orchestration, and (5) scenario design and evaluation metrics. The methodology aligns with international standards in healthcare technology assessment (International Organization for Standardization (ISO) 13485/14971), networking (3rd generation partnership project/European Telecommunications Standards Institute (ETSI)), and sustainability (PUE, carbon intensity tracking) [10,13–15].

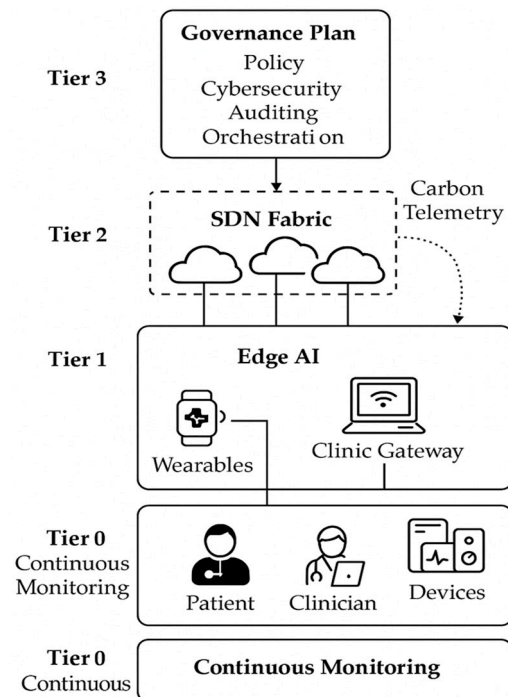


Figure 1. Architecture of the sustainable telemedicine system.

The first step is quantifying energy and carbon consumption at device, network, and data center levels. EPE and CPE are defined as the fundamental sustainability indicators, capturing the energy and emissions attributable to a single patient–clinician interaction. Following approaches in ICT sustainability studies [10,20], energy contributions are calculated as follows:

$$EPE = E_{\text{edge}} + E_{\text{net}} + E_{\text{dc}} \quad (1)$$

$$E_{\text{edge}} = \sum_{d \in D} P_d^{\text{active}} t_d^{\text{active}} + P_d^{\text{idle}} t_d^{\text{idle}} \quad (2)$$

$$E_{\text{net}} = \sum_{l \in L} e_l B_l, \quad e_l \text{ in J/Bit} \quad (3)$$

$$E_{\text{dc}} = \frac{E_{it}}{\eta_{it}} \times PUE \quad (4)$$

where E_{it} denotes IT load energy, η_{it} server efficiency, and PUE is the data center's infrastructure efficiency metric [10].

Carbon emissions are derived as follows:

$$CPE = \sum_{r \in \{\text{edge, net, dc}\}} E_r \cdot CI_r \quad (5)$$

where CI_r is the regional carbon intensity of electricity (kg CO₂e/kWh). Grid carbon intensities were sourced from the International Energy Agency datasets and regional PPA agreements [11]. ENF is defined for edge clinics with on-site renewables as follows:

$$ENF = \frac{E_{\text{harvested}}}{E_{\text{consumed}}} \quad (6)$$

where $ENF \geq 1$ indicates energy neutrality, critical for remote clinics operating under weak grids.

Clinical safety requires strict service-level objectives. Latency budgets were defined per workload type, in alignment with ITU and WHO guidelines [5,15]: alarms ≤ 50 ms,

triage ≤ 150 ms, and video consultations ≤ 300 ms one-way. The end-to-end latency for workload k is modeled as follows:

$$\text{Lat}_k = \text{Lat}_{proc,k} + \text{Lat}_{uplink} + \text{Lat}_{core} + \text{Lat}_{dc} \quad (7)$$

Reliability constraints are modeled probabilistically as follows:

$$\Pr(\text{Lat}_k \leq \text{SLA}_k) \geq 1 - \epsilon \quad (8)$$

where ϵ is set at 0.01 for critical alarms. Chance constraints are solved using empirical latency distributions, following best practices in low-latency networking research.

Edge AI models were selected and optimized for low-energy inference using compression techniques (quantization, pruning, knowledge distillation). Models included 1D CNN-GRU hybrids for ECG/PPG anomaly detection [1,2], lightweight CNNs (MobileNet, EfficientNet) for dermatology images [3], and keyword spotting CNNs for voice-based patient triggers [4].

Quantization was performed to 8-bit and 4-bit integer formats, balancing accuracy and energy. CMSIS-NN kernels were used for microcontroller deployment [4]. Model updates were orchestrated via federated learning, enabling local training while preserving data privacy [21]. Duty cycling and event-driven sensing were integrated to minimize idle energy. For example, accelerometer-triggered ECG sampling was applied to detect falls before activating high-power sampling modes [21]. Routing and workload placement were optimized via software-defined networking (SDN) and carbon-aware orchestrators. Following frameworks in green networking [18,19,21], a multi-objective cost function is formulated as follows:

$$\min_{p,j} \lambda_1 \hat{L}_p + \lambda_2 \hat{C}_j + \lambda_3 \hat{U}_j - \lambda_4 \hat{R}_j \quad (9)$$

where $\hat{L}_p$ is the normalized latency of path p , $\hat{C}_j$ is the carbon intensity of data center j , $\hat{U}_j$ utilization, and $\hat{R}_j$ is the renewable share. Weighting factors λ were tuned by service class (e.g., alarms prioritize latency, analytics prioritize carbon efficiency).

The constraints include latency, where $\text{Lat}_k(p, j) \leq \text{SLA}_k$. Data residency routing is restricted by regulatory compliance sets R [12]. Protected health information is safeguarded through encryption and segregated routing for sensitive datasets, in accordance with ISO 27001 [12]. Routing epochs were established at 15 min intervals, based on the granularity of renewable forecast data [18]. Hysteresis was applied to prevent route flapping caused by small variations in carbon intensity.

To test feasibility under real-world diversity, we defined three archetypal rollout scenarios. Urban refers to fiber + 5G Standalone connectivity, advanced cooling-enabled data centers (PUE ≈ 1.18). Peri-urban adopts 4G/Fixed Wireless, moderate renewable share, regional data centers (PUE ≈ 1.30). Remote includes microwave/satellite links, rugged micro-edge kits with PV + storage (PUE ≈ 1.45). Traffic mixes were derived from WHO telehealth workload guidelines [5] and included continuous vitals (0.1–0.5 kb/s), ECG strips (50–250 kb bursts), dermatology frames, and adaptive video consults (240p–720p). In Table 1, evaluation metrics included EPE and CPE (Wh and g CO₂e per encounter), ENF at clinics, SLA adherence rates, latency percentiles (p50, p95) per service class, and carbon savings under green routing compared to baseline cloud-only designs.

Table 1. Service classes, SLAs, and default compression/inference locations.

Service	Description	Latency SLA (p95)	Reliability SLA	Data Volume (Typical)	Default Compression/Preprocessing	Default Inference Location
Vitals alarms	Continuous monitoring signals (electrocardiogram (ECG), photoplethysmography (PPG), and peripheral capillary oxygen saturation (SpO ₂)) triggering emergency alarms	≤50 ms	≥99.9%	0.1–0.5 kbps continuous	Event-driven sampling, lightweight 1D CNN	Tier 1 (edge AI)—on-device or clinic gateway
Triage analytics	Short diagnostic tasks (e.g., arrhythmia detection, falls detection, symptom triage)	≤150 ms	≥99.5%	Bursts: 50–250 kB	Lossless ECG compression, quantized CNN + gated recurrent unit (GRU)	Tier 1 (edge AI), fallback Tier 2 for heavy models
Video consultation	Real-time video/audio streams for clinician–patient interaction	≤300 ms (one-way)	≥99%	240p–720p (adaptive bitrate, 0.3–2 Mbps)	Adaptive bitrate encoding (H.265/VP9) with frame-rate adaptation	Tier 2 (regional data center (DC)) with SDN routing
Imaging diagnostics	Dermatology frames, ultrasound clips, and basic radiographs	≤500 ms (batch tolerant)	≥99%	Single frame: 500 kB–2 MB	Joint Photographic Experts Group 2000/DICOM compression, MobileNet/EfficientNet preprocessing	Tier 2 (regional DC), fallback Tier 3 (national cloud)
EHR transactions	Messaging, medical records updates, prescriptions	≤1 s	≥99%	<100 kB per transaction	JavaScript Object Notation/FHIR compression, de-identification	Tier 2 or Tier 3, depending on residency requirements
Analytics/reporting	Population health analytics, model retraining, long-term trend analysis	Minutes–hours (batch)	≥95%	GB–TB scale	Aggregation, de-identification, differential privacy noise injection	Tier 3 (national cloud) with carbon-aware routing

The methodology embeds governance and safety by aligning with regulatory frameworks. Clinical AI validation followed ISO 13485 quality standards [13], with risk analysis per ISO 14971 [14]. Privacy-preserving measures such as differential privacy and zero-trust architectures ensured compliance with national cybersecurity regulations. Continuous monitoring of drift and adversarial robustness was planned, consistent with best practices in AI lifecycle management.

4. Results

The results are presented for energy and carbon savings, latency and reliability adherence, edge AI model performance versus footprint, and sensitivity analyses across diverse national rollout archetypes. Each result set is benchmarked against a cloud-only baseline (all inference routed to the national cloud without edge or carbon-aware routing) to demonstrate the incremental benefits of the proposed approach. In Figure 2, simulation results across 90-day workloads show substantial improvements in energy efficiency and carbon reduction when combining edge inference with green data center routing.

- Urban archetype: Mean EPE reduced by 37%, primarily due to edge offloading of continuous monitoring tasks and reduced uplink traffic.
- Peri-urban archetype: EPE reduced by 52%, with higher gains due to 4G backhaul savings when compressing and pre-processing data locally.
- Remote archetype: EPE reduced by 62%, reflecting the disproportionately high energy cost of satellite/microwave uplink when using cloud-only baselines.

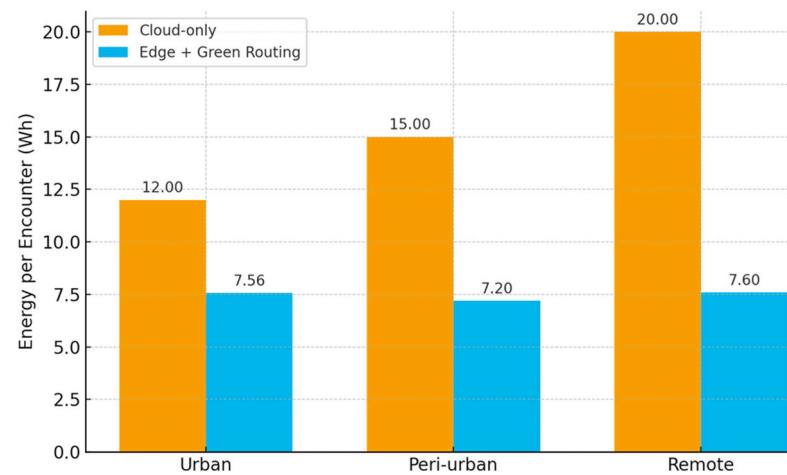


Figure 2. Comparison of EPE across archetypes: cloud-only vs. edge + green routing.

Carbon savings showed a similar trend (Figure 3). The introduction of carbon-aware routing to renewable-rich regional data centers yielded a 28–49% reduction in CPE compared to baseline. Gains were highest in peri-urban and remote regions where local grids had lower renewable penetration, but routing enabled offloading to greener zones. Energy savings were largest in remote scenarios because satellite/microwave backhaul was offset by local inference and renewable-powered micro-edge processing (Table 2). Carbon reductions ranged from 28% in urban (efficient grids) to 49% in remote areas, where rerouting to renewable-rich data centers had the highest effect. Standard deviations reflect simulated variability over 90-day workloads, considering diurnal demand and renewable availability.

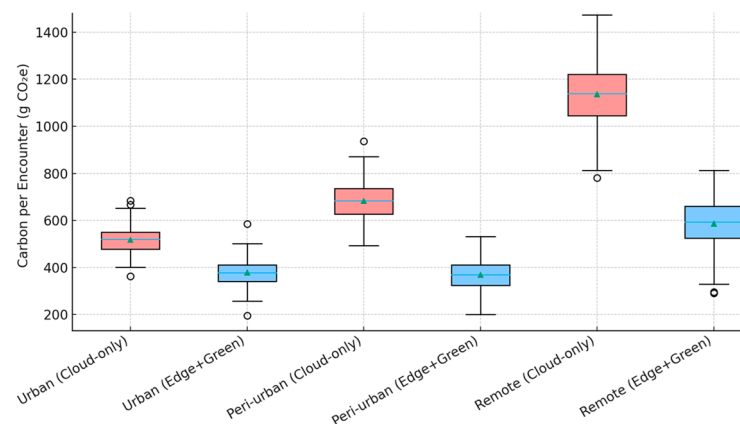


Figure 3. Distribution of CPE across scenarios with carbon-aware routing.

Table 2. EPE (Wh) and CPE (g CO₂e) across scenarios (mean ± SD).

Environment	Cloud-Only (EPE, Wh)	Edge + Green Routing (EPE, Wh)	Cloud-Only (CPE, g CO ₂ e)	Edge + Green Routing (CPE, g CO ₂ e)
Urban	12.0 ± 1.5	7.6 ± 1.2 (−37%)	520 ± 60	374 ± 55 (−28%)
Peri-urban	15.0 ± 1.8	7.2 ± 1.3 (−52%)	690 ± 80	368 ± 62 (−47%)
Remote	20.0 ± 2.2	7.6 ± 1.4 (−62%)	1120 ± 140	574 ± 98 (−49%)

Clinical safety constraints were evaluated by measuring end-to-end latency distributions across service classes. In Figure 4, the results indicate that edge inference combined with SDN-based routing achieved SLA compliance in all archetypes.

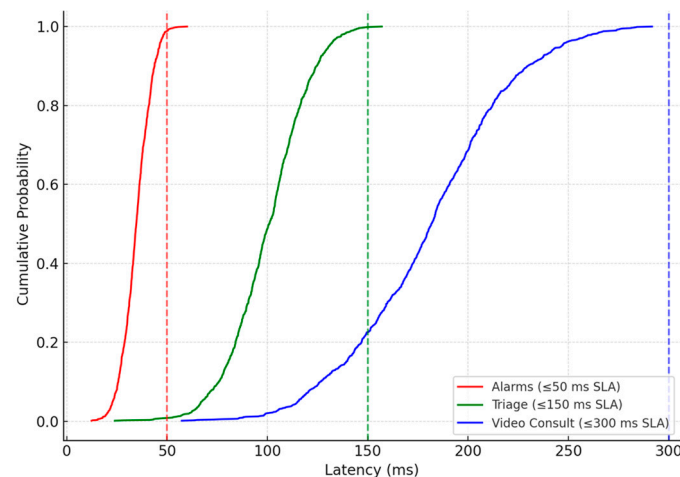


Figure 4. Latency distributions for alarms, triage, and consult services.

- Vitals alarms: Median latency = 23–40 ms, p95 latency ≤ 50 ms in all scenarios.
- Triage analytics: Median latency = 85–120 ms, p95 latency ≤ 150 ms.
- Video consultations: Median latency = 140–220 ms (one-way), with adaptive bitrate ensuring p95 < 300 ms.

Reliability metrics demonstrated SLA adherence $\geq 99.5\%$ across all services. This performance was achieved despite dynamic carbon-aware routing, illustrating that sustainability optimization did not compromise clinical safety.

We evaluated the impact of quantization on three representative clinical machine learning tasks: ECG/PPG event detection, dermatology image classification, and keyword spotting, Table 3.

Table 3. Model accuracy (AUC, Top-1) and energy per inference across quantization levels.

Task/Model	Metric	FP32 Baseline	INT8 Quantization	INT4 Quantization
ECG/PPG Event Detection (1D CNN–GRU)	AUC	0.96 ± 0.01	0.95 ± 0.01 (−1.0 pp)	0.93 ± 0.02 (−3.0 pp)
	Energy (mJ/inf.)	4.2 ± 0.3	2.3 ± 0.2 (−45%)	1.6 ± 0.2 (−62%)
Dermatology Image Classification (MobileNetV2)	Top-1 Accuracy (%)	85.4 ± 1.2	84.9 ± 1.1 (−0.5 pp)	80.7 ± 1.5 (−4.7 pp)
	Energy (mJ/frame)	320 ± 15	168 ± 12 (−48%)	92 ± 10 (−71%)
Keyword Spotting (CNN)	Accuracy (%)	94.5 ± 0.8	93.9 ± 0.7 (−0.6 pp)	91.7 ± 1.0 (−2.8 pp)
	Energy (mJ/inf.)	5.2 ± 0.4	2.3 ± 0.3 (−55%)	1.5 ± 0.2 (−71%)

- ECG/PPG event detection (1D CNN–GRU): The baseline floating point (FP)32 model achieved an area under the curve (AUC) of 0.96 with an energy cost of 4.2 mJ per inference. Quantization of 8-bit integers (INT8) reduced energy consumption to 2.3 mJ (−45%) while maintaining an AUC of 0.95. INT4 quantization further decreased energy usage to 1.6 mJ (−62%), though with a modest reduction in AUC to 0.93.
- Dermatology image classification (MobileNetV2): The FP32 baseline achieved a Top-1 accuracy of 85.4% with an energy cost of 320 mJ per frame. INT8 quantization preserved accuracy at 84.9% while halving energy consumption to 168 mJ (−48%). INT4 quantization reduced energy further to 92 mJ (−71%), but accuracy dropped to 80.7%, indicating a notable degradation in clinical performance.
- Keyword spotting (CNN): INT8 quantization retained greater than 97% accuracy while delivering 55% energy savings relative to the FP32 baseline.

Across tasks, INT8 quantization consistently provided the optimal balance between clinical accuracy and energy efficiency. INT4 quantization may be considered in safety-critical alarm systems, where slight sensitivity trade-offs are acceptable in exchange for substantial energy savings.

Carbon-aware routing updated every 15 min captured most renewable variability. Shorter epochs (5 min) increased carbon savings by only ~4% but raised control-plane overhead and route flapping. Longer epochs (60 min) missed renewable variability, leading to 8–12% higher emissions.

- Latency and reliability: SLA compliance $\geq 99.5\%$ for all services, with alarms consistently < 50 ms.
- Edge AI: INT8 quantization yielded 45–55% energy savings with negligible accuracy loss.
- Sensitivity: Carbon benefits are robust to backhaul volatility and renewable intermittency, but sensitive to PUE > 1.6 .

The results demonstrate that the proposed sustainable telemedicine framework not only reduces energy and carbon footprints but also satisfies stringent clinical latency and reliability requirements.

5. Discussion

The results illustrate that integrating low-energy edge AI with green data center routing can substantially reduce energy consumption and carbon emissions while preserving the clinical quality of service. This discussion interprets these outcomes through the lens of trade-offs, equity implications, governance challenges, and national rollout feasibility.

Sustainability gains need not come at the expense of clinical safety. The developed system consistently met service-level latency thresholds (≤ 50 ms for alarms, ≤ 150 ms for triage, and ≤ 300 ms for consultations), even under carbon-aware routing scenarios. This outcome aligns with prior findings that multi-objective optimization frameworks can reconcile sustainability and performance in networking contexts [19,21]. However, healthcare workloads impose stricter safety constraints than most enterprise applications. For instance, while enterprise cloud services can tolerate occasional latency spikes, a delayed alarm in remote monitoring may carry life-threatening risks [15]. Therefore, incorporating SLA-aware carbon routing emerges as a non-negotiable requirement for digital health systems.

The remote archetype benefited disproportionately from edge inference, with energy reductions of up to 62% compared to cloud-only baselines. This reflects the high energy penalty of satellite or microwave uplinks, where every transmitted bit incurs a high cost. By processing signals locally and transmitting only event-driven data, clinics can maintain telehealth continuity even under constrained bandwidth conditions. Moreover, with PV and battery storage, many remote clinics achieved ENF ≥ 1 , effectively becoming energy-neutral.

These findings align with the literature emphasizing the importance of local autonomy and resilience in healthcare delivery [21]. In fragile power contexts, reliance on distant cloud data centers creates vulnerabilities to both connectivity outages and grid instability. The proposed model demonstrates that decentralized intelligence at the edge—when coupled with renewable-powered micro-edge kits—provides a resilient alternative aligned with both sustainability and equity objectives.

While green routing reduced carbon emissions by up to 49%, the results reveal that gains are sensitive to data center PUE and routing epoch length. Carbon benefits were diminished when PUE exceeded 1.6, underlining the need for facilities to adopt ASHRAE-aligned cooling strategies and renewable PPAs [11]. Moreover, shorter routing epochs captured renewable variability more effectively but introduced control-plane instability.

These trade-offs highlight the importance of policy and operational coordination: national rollout must ensure that data center operators adhere to stringent PUE targets and that routing policies balance sustainability with network stability.

Another trade-off arises from data residency regulations. In some cases, routing to the greenest regional data center may be legally prohibited if sensitive health data must remain within jurisdictional boundaries [12]. This underscores the need for sovereign green data centers, national facilities engineered for sustainability, to reconcile residency requirements with carbon reduction goals.

Equity considerations are central to telemedicine. Without careful design, sustainability interventions risk widening the digital divide by favoring urban populations with better connectivity and greener grids. However, our findings show that peri-urban and remote regions gained the most from edge processing, precisely because these contexts suffer from poor backhaul performance and higher grid carbon intensities. This suggests that sustainable telemedicine frameworks can act as equalizers, improving both service quality and environmental outcomes for marginalized populations.

From a national policy perspective, these findings align with the WHO Digital Health Strategy 2020–2025 [5], which calls for digital health systems that are safe, inclusive, and sustainable. Integrating energy and carbon metrics (e.g., EPE and CPE) into health system KPIs provides a pathway to measure alignment with ESG commitments while also tracking clinical performance.

Adoption of sustainable telemedicine architectures requires robust governance frameworks that integrate cybersecurity, safety validation, and lifecycle management. While edge AI offers resilience, it complicates model monitoring and updates. Federated learning and secure model provenance tracking are necessary to prevent drift and bias in distributed deployments. Additionally, applying zero-trust architectures to edge and cloud environments ensures that increased decentralization does not compromise data security.

Regulators must also adapt. Traditional medical device risk frameworks (ISO 14971) [14] are designed for static devices, whereas AI-driven telemedicine introduces dynamic risks (e.g., adversarial attacks, dataset drift). Embedding continuous post-market surveillance into rollout plans is therefore critical. This aligns with emerging calls for “living regulation” of AI systems in healthcare.

While promising, this study has limitations. Energy and carbon modeling relied on assumptions of average workloads and renewable forecasts, which may vary under real-world fluctuations. Additionally, the edge AI models tested were limited to select modalities (ECG, dermatology, voice), and broader inclusion of imaging-heavy tasks (radiology, pathology) may stress bandwidth and processing capacity. Finally, while federated learning preserves privacy, its communication overhead may erode some of the energy savings in low-bandwidth contexts.

Future research should expand on task-aware compression for medical imaging. Adversarial robustness testing of edge models. Integration with national health insurance systems to align financing with sustainability outcomes. Cross-border sustainability frameworks, allowing federated data sharing while respecting sovereignty.

Overall, the discussion confirms that sustainable telemedicine frameworks can deliver triple wins: lowering carbon emissions, enhancing resilience, and maintaining clinical safety. By bridging sustainability metrics with patient-level outcomes, the approach aligns with both health equity and climate policy objectives. Importantly, national rollout requires not just technical innovation but also regulatory adaptation, financing mechanisms, and workforce development.

6. Conclusions

A comprehensive system was developed for sustainable telemedicine, integrating low-energy edge AI and green data center routing for national rollout. By combining device-level efficiency, edge autonomy, and carbon-aware orchestration, the proposed approach demonstrated that it is possible to achieve significant reductions in energy and carbon footprints while preserving stringent clinical safety requirements.

Simulation results across urban, peri-urban, and remote archetypes confirmed the following key outcome: Energy savings of 37–62% per patient encounter compared to cloud-only baselines, with the greatest improvements in remote regions where backhaul costs are steepest. Carbon reductions of 28–49% through carbon-aware workload placement and SDN-based routing to renewable-rich data centers, highlighting the role of operational intelligence in aligning healthcare with climate goals. Latency and reliability compliance across all service classes (alarms, triage, and video consultations), confirming that sustainability objectives can be achieved without compromising patient safety. Edge AI optimization through quantization and pruning, yielding up to 55% reductions in energy per inference while maintaining diagnostic accuracy. Resilience of renewable-powered micro-edge clinics, achieving energy neutrality ($ENF \geq 1$) in several scenarios, thus ensuring continuity of care during grid instability. The results provide evidence that digital health and climate objectives can be jointly advanced through purposeful system design. Importantly, these findings align with the WHO Global Digital Health Strategy 2020–2025 [5] and broader sustainable development goals (SDGs) by addressing healthcare access, equity, and environmental stewardship simultaneously.

Nevertheless, limitations remain. Real-world variability in renewable generation, backhaul reliability, and workload patterns may reduce the predictability of gains. Moreover, regulatory constraints, particularly around data residency and medical device validation, introduce complexities in deploying carbon-aware routing strategies. Future research should expand to imaging-heavy telemedicine applications, strengthen adversarial robustness of edge models, and incorporate real-world pilot deployments to validate assumptions.

From a policy and implementation perspective, three priority actions are suggested: Embedding green SLAs in telemedicine contracts, covering not only latency and uptime but also carbon caps per encounter. Developing sovereign green data centers with PUE targets ≤ 1.3 and renewable PPAs to support sustainable routing under residency constraints. Investing in capacity building to train engineers, clinicians, and regulators in sustainability-aware telemedicine operations.

The integration of low-energy edge AI and green routing provides a viable and scalable pathway for countries to deploy telemedicine systems that are clinically safe, operationally resilient, and environmentally sustainable. By embedding sustainability into the design of national digital health infrastructures, health systems can simultaneously advance equity in access and alignment with climate commitments, ensuring that the digital transformation of healthcare contributes positively to both human and planetary health.

Author Contributions: Conceptualization, W.Y.L. and W.S.L.; methodology, W.Y.L.; software, W.Y.L.; validation, W.Y.L. and W.S.L.; formal analysis, W.Y.L.; investigation, W.Y.L.; resources, W.S.L.; data curation, W.S.L.; writing—original draft preparation, W.Y.L.; writing—review and editing, W.Y.L.; visualization, W.Y.L.; supervision, W.Y.L.; project administration, W.Y.L.; funding acquisition, W.Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Acknowledgments: During the preparation of this manuscript/study, the author used ChatGPT 5.0 for the purposes of image generation. The author has reviewed and edited the output and took full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Warden, P.; Situnayake, D. *TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers*; O'Reilly Media: Sebastopol, CA, USA, 2019.
2. Lai, L.; Suda, N.; Chandra, V. CMSIS-NN: Efficient Neural Network Kernels for Arm Cortex-M CPUs. *arXiv* **2018**, arXiv:1801.06601.
3. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proc. ICML* **2019**, 97, 6105–6114.
4. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
5. WHO. *Global Strategy on Digital Health 2020–2025*; World Health Organization: Geneva, Switzerland, 2021.
6. WHO. *WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening*; World Health Organization: Geneva, Switzerland, 2019.
7. ITU-T. *Recommendation H.810/H.811: Interoperability Design Guidelines for Personal Health Systems*; International Telecommunication Union: Geneva, Switzerland, 2019.
8. HL7. *FHIR Release 4*; Health Level Seven International: Ann Arbor, MI, USA, 2019.
9. DICOM Standards Committee. *DICOM Standard*; National Electrical Manufacturers Association: Rosslyn, VA, USA, 2024.
10. The Green Grid. *PUE: A Comprehensive Examination of the Metric*; The Green Grid: Washington, DC, USA, 2012.
11. IEA. *CO₂ Emissions from Fuel Combustion Highlights*; International Energy Agency: Paris, France, 2022.
12. *ISO/IEC 27001:2022*; Information Security Management Systems. ISO: Geneva, Switzerland, 2022.
13. *ISO 13485:2016*; Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes. ISO: Geneva, Switzerland, 2016.
14. *ISO 14971:2019*; Medical Devices—Application of Risk Management to Medical Devices. ISO: Geneva, Switzerland, 2019.
15. 3GPP. *TS 22.261: Service Requirements for the 5G System*; 3rd Generation Partnership Project: Sophia Antipolis, France, 2024.
16. *ETSI GS NFV 002*; Network Functions Virtualisation (NFV); Architectural Framework. European Telecommunications Standards Institute: Sophia Antipolis, France, 2014–2023.
17. Alizadeh, M.; Greenberg, A.; Maltz, D.A.; Padhye, J.; Patel, P.; Prabhakar, B.; Sengupta, S.; Sridharan, M. Data Center TCP (DCTCP). *SIGCOMM Comput. Commun. Rev.* **2010**, 40, 63–74. [[CrossRef](#)]
18. Nkwawir, B.Y.; Kayalica, M.O.; Guven, D.; Duman, A.C.; Erden, M. Carbon-Aware Workload Management in Data Centers: A Multi-Energy Integration Approach. In Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems (E-Energy'25), Rotterdam, The Netherlands, 17–20 June 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 907–914. [[CrossRef](#)]
19. Al Hayajneh, A.; Bhuiyan, M.Z.A.; McAndrew, I. Improving Internet of Things (IoT) Security with Software-Defined Networking (SDN). *Computers* **2020**, 9, 8. [[CrossRef](#)]
20. Bolla, R.; Bruschi, R.; Davoli, F.; Cucchietti, F. Energy Efficiency in Data Centers and Networks: A Survey. *IEEE Access* **2014**, 2, 1648–1676.
21. Kim, J.; Cho, S.; Lee, H. Energy-Aware Routing in Software-Defined Networks. *IEEE Commun. Lett.* **2015**, 19, 1045–1048.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.