

Article

Energy-Aware Ultra-Reliable Low-Latency Communication for Healthcare IoT in Beyond 5G and 6G Networks

Adeel Iqbal ^{1,†} , Tahir Khurshaid ^{2,†} , Ali Nauman ^{1,*}  and Sang-Bong Rhee ²
¹ School of Computer Science and Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea; adeeliqbal@yu.ac.kr

² Department of Electrical Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea; tahir@ynu.ac.kr (T.K.); rrsd@yu.ac.kr (S.-B.R.)

* Correspondence: anauman@ynu.ac.kr

† These authors contributed equally to this work.

Abstract: Ultra-reliable low-latency communication (URLLC) is a cornerstone of beyond 5G and future 6G networks, particularly for mission-critical applications such as the healthcare Internet of Things. In applications such as remote surgery, emergency services, and real-time health monitoring, it is imperative to ensure stringent latency and reliability requirements. However, the energy constraints of wearable and implantable medical devices pose stringent challenges to conventional URLLC methods. This paper proposes an energy-aware URLLC framework that dynamically prioritizes healthcare traffic to optimize transmission energy and reliability. The framework integrates a priority-aware packet scheduler, adaptive transmission control, and edge-enabled reliability management. Extensive Monte Carlo simulations are carried out on various network loads and varying edge computing delays to evaluate performance metrics, like latency, throughput, reliability score, energy consumption, delay violation rate, and Jain's fairness index. Results illustrate that the suggested technique achieves lower latency, energy consumption, and delay violation rates and higher throughput and reliability scores, sacrificing Jain's fairness index gracefully at peak network overload. This study is a potential research lead for green URLLC in healthcare IoT systems to come.

Keywords: ultra-reliable low-latency communication (URLLC); healthcare IoT (H-IoT); beyond 5G (B5G); 6G networks; energy efficiency; priority-aware scheduling; edge computing; Monte Carlo simulation; delay violation rate; reliability score



Academic Editor: Peter Han Joo Chong

Received: 10 April 2025

Revised: 8 May 2025

Accepted: 30 May 2025

Published: 31 May 2025

Citation: Iqbal, A.; Khurshaid, T.; Nauman, A.; Rhee, S.-B. Energy-Aware Ultra-Reliable Low-Latency Communication for Healthcare IoT in Beyond 5G and 6G Networks. *Sensors* **2025**, *25*, 3474. <https://doi.org/10.3390/s25113474>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The advent of beyond 5G (B5G) and upcoming 6G wireless communication networks marks a paradigm shift in enabling mission-critical applications that demand ultra-reliable and low-latency communication (URLLC) [1,2]. Among the most transformative of these domains is the healthcare Internet of Things (H-IoT) [3], which includes applications such as real-time patient monitoring, remote surgery, emergency response coordination, and continuous health diagnostics. These use cases impose stringent latency requirements as low as 1 ms and near-perfect reliability up to 99.999% [4,5].

While URLLC has received considerable attention in industrial automation [6], UAV [7], and vehicular communication [8], its deployment in healthcare presents unique challenges. Most H-IoT devices, such as wearable sensors, implantable monitors, and body area network nodes, operate under severe energy constraints [9]. This makes it infeasible to adopt traditional URLLC strategies that rely on high transmission power or redundant

retransmissions. Moreover, the diversity of healthcare traffic, ranging from life-critical ECG signals to non-urgent temperature readings, calls for a context-aware and differentiated approach to resource allocation [10]. Figure 1 is the visual representation of the H-IoT network where multiple biomedical sensors transmit the patient's information to a nearby edge computing node for real-time processing. The edge node is linked to the hospital network, protected by a firewall, and communicates with the Internet via a 5G network. This setup offers low-latency, secure, and efficient health monitoring.

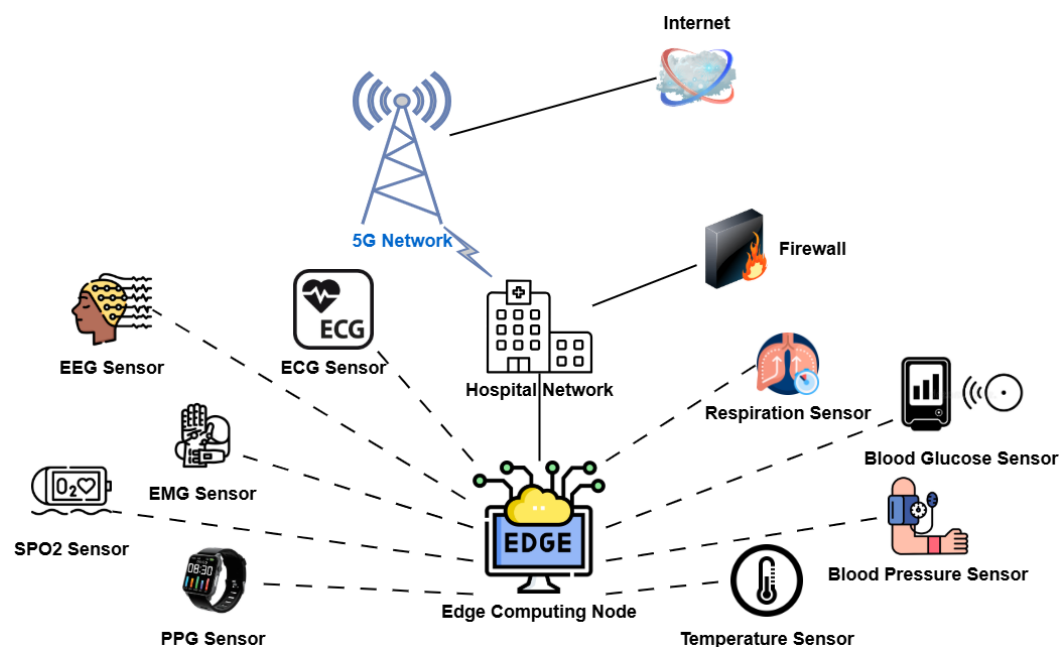


Figure 1. An edge-enabled smart healthcare architecture utilizing 5G connectivity.

Recent works have proposed adaptive scheduling [11,12] and reliability enhancement schemes [13,14]; however, few have explicitly addressed the trade-off between reliability, latency, and energy efficiency in medical scenarios. To this end, we propose an energy-aware URLLC framework tailored to H-IoT systems in B5G/6G environments. Our approach integrates (1) a priority-aware packet scheduler that allocates transmission opportunities based on urgency and device energy profile, (2) an adaptive transmission control module that dynamically adjusts modulation and power settings, and (3) an edge-assisted reliability manager that proactively minimizes packet loss and delay violation. To validate our framework, we perform extensive Monte Carlo simulations across a wide range of network loads and varying edge computing delays. We compare the proposed model with three baseline URLLC techniques: fixed scheduling [15,16], ALOHA-based random access [16,17], and priority-only URLLC [16,18]. Performance is evaluated using key metrics, including average latency, throughput, packet loss rate, energy consumption, reliability score, delay violation rate, and Jain's fairness index.

The remainder of this paper is organized as follows: Section 2 discusses related work. Section 3 introduces the system model and problem formulation. Section 4 presents the proposed energy-aware URLLC framework. Section 5 explains the simulation setup and evaluation metrics. Section 6 presents the results and discussion. Finally, Section 7 concludes this paper and outlines future work.

2. Related Work

URLLC has emerged as a fundamental enabler for mission-critical applications in B5G and upcoming 6G networks. In the context of the H-IoT, achieving sub-millisecond

latency and ultra-high reliability is vital for applications such as remote surgery, real-time patient monitoring, and emergency healthcare services. These stringent communication demands have inspired significant research focusing on energy efficiency, traffic prioritization, intelligent scheduling, and edge-based control mechanisms.

In terms of system-wide optimization, the authors in [16] conducted a comprehensive survey on URLLC scheduling in 5G networks and laid the groundwork for emerging 6G frameworks. They emphasized the importance of integrating quality-of-service (QoS) constraints with system resource constraints, particularly in latency-sensitive and energy-constrained environments.

Recent literature has addressed various aspects of URLLC design. For instance, the authors in [19] propose a latency-energy optimization model suitable for cellular IoT systems. However, their framework does not consider traffic differentiation or residual energy constraints, which are vital in H-IoT deployments. Similarly, the authors in [20] focus on reducing latency for rural smart healthcare systems using zero-forcing equalization, but the approach lacks energy-awareness and traffic-class-based prioritization.

The authors in [21] introduce a priority-aware access control framework for carbon footprint monitoring, demonstrating improved responsiveness for urgent data. However, the solution does not account for fairness among devices or energy limitations. In a broader orchestration context, the authors in [22] develop learning-based slicing and resource allocation for IoT-5G/6G systems, but their model is more aligned with network-level optimization than device-level scheduling suitable for H-IoT.

Several works have explored scheduling techniques relevant to URLLC. The authors in [23] propose a Lyapunov-drift-based scheduling method optimized for finite blocklength transmissions in short-packet communications. Although effective in balancing delay and reliability, this method overlooks energy metrics and heterogeneous traffic priorities. The authors in [24] develop a priority-aware scheduler for co-existing wireless body area networks, but the work lacks dynamic energy controls and adaptive reliability support. The authors in [25] extend this line of work by incorporating packet size and delay sensitivity in a heterogeneous multi-server environment; however, they assume static energy availability and omit edge assistance for dynamic adaptation.

The authors in [26] propose a reinforcement learning-based offloading framework for health Internet of Things, which demonstrates potential for adaptive control. However, their work does not incorporate residual battery or fairness constraints, which are critical in multi-class H-IoT deployments. Notably, none of these studies explicitly address fairness using standard metrics like Jain's Index or implement adaptive throughput equalization. As a result, the proposed framework stands out in providing energy-, latency-, and fairness-aware scheduling within an edge-centric H-IoT context. A comparative overview of these models is presented in Table 1.

Table 1. Comparison with existing works on URLLC in H-IoT context.

Study	Energy-Aware	URLLC-Oriented	H-IoT Specific	Edge-Based Logic	Traffic Prioritization	Fairness-Aware
[16]	X	✓	X	X	X	X
[19]	✓	✓	X	X	X	X
[20]	X	✓	✓	X	X	X
[21]	X	✓	✓	X	✓	X
[22]	✓	✓	X	✓	X	X
[23]	X	✓	X	X	X	X
[24]	X	✓	✓	X	✓	X
[25]	X	✓	✓	✓	✓	X
[26]	✓	✓	✓	✓	✓	X
Proposed Model	✓	✓	✓	✓	✓	✓

Cadence PCB Solutions [27] described real-world deployment scenarios such as connected ambulances and tele-ICUs, where ultra-low latency communication is essential. These use cases underline the necessity of integrating both priority and energy awareness into communication protocols to ensure uninterrupted and safe healthcare delivery.

The authors in [28] examined challenges in coexistence management for URLLC and eMBB services, pointing out that hybrid traffic handling will be a major hurdle in future wireless networks. Their findings suggest that more granular scheduling policies are needed to meet the URLLC demands of H-IoT devices without degrading the performance of other services.

Despite these advancements, most current models either focus on reliability and latency without considering energy constraints or optimize energy efficiency at the cost of QoS. The proposed energy-aware URLLC framework in this paper aims to fill this research gap by introducing a balanced, adaptable, and priority-aware solution, specifically tailored to the constraints and demands of next-generation H-IoT systems.

3. System Model and Problem Statement

In this section, the system architecture is presented, and the problem of enabling energy-aware URLLC for H-IoT beyond 5G and 6G systems is stated. Three significant layers are present in the architecture: the sensing layer, including wearable and implantable medical sensors; the edge computing layer, where local control, scheduling, and reliability management are executed; and the core network layer, which is responsible for long-term storage, diagnostics, and analytics using the cloud. Figure 2 illustrates the system architecture in detail.

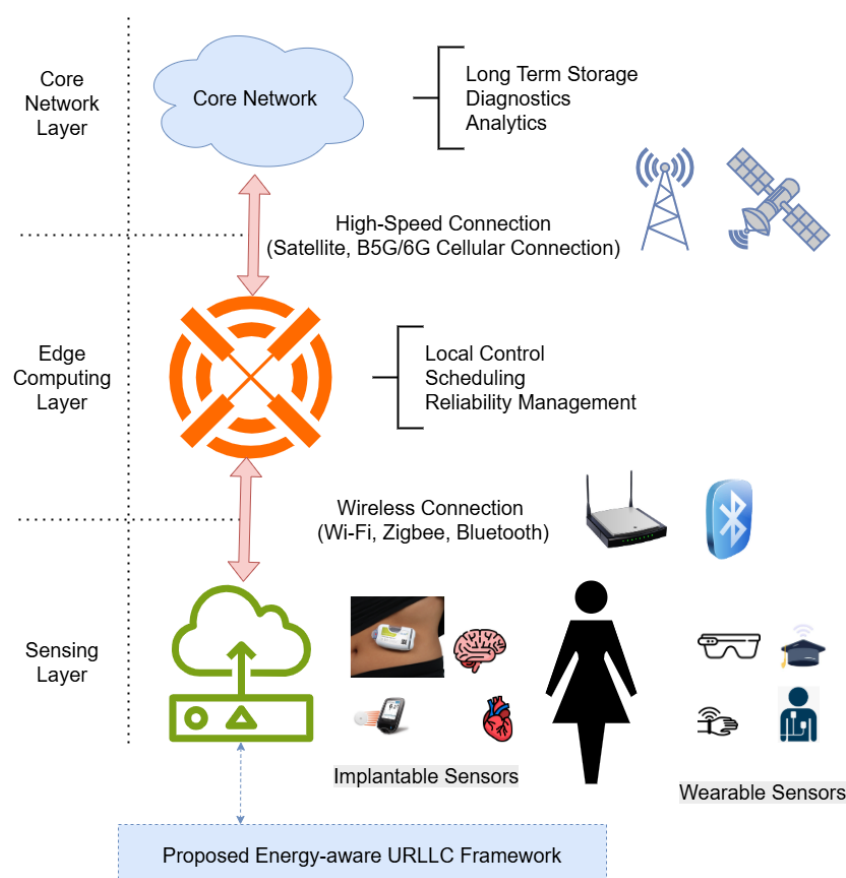


Figure 2. System architecture of the proposed energy-aware URLLC framework for H-IoT.

We consider a heterogeneous smart healthcare environment of N implantable or wearable medical sensors, denoted by $\{d_1, d_2, \dots, d_N\}$. The sensors measure a collection of physiological parameters such as ECG, heart rate, blood oxygen saturation, and body temperature. Each sensor sends data wirelessly to an access point with Mobile Edge Computing (MEC), which performs local scheduling, power control, and reliability analysis. This MEC node is also provided with a cloud backend link for logging data and diagnostic processing.

3.1. Priority-Aware Packet Scheduler (PAPS)

Each packet p_i is assigned a metadata tuple $\mathcal{M}_i = \{w_i, T_i^{\max}, B_i(t), \text{channel state}\}$, where w_i is the traffic class weight, $T_i^{\max}$ is the latency deadline, and $B_i(t)$ is the remaining device energy. An urgency score is calculated as follows:

$$U_i = \frac{w_i}{B_i(t) + \epsilon} \quad (1)$$

Packets are prioritized based on U_i and are scheduled accordingly. A fairness mechanism monitors throughput and temporarily boosts access for underperforming nodes by relaxing access thresholds.

3.2. Adaptive Transmission Control (ATC)

The energy consumed for transmission is

$$E_i = P_i^{\text{tx}} \cdot t_i^{\text{tx}} \quad (2)$$

ATC selects modulation/power profiles based on $T_i^{\max}$, $B_i(t)$, and channel SNR γ_i , ensuring low power for non-critical data and robust settings for urgent traffic. If battery levels fall below 30%, non-critical traffic is suppressed to conserve energy.

3.3. Edge-Assisted Reliability Manager (EARM)

Channel conditions are modeled via block-fading Nakagami- m distribution. The successful transmission probability is

$$P_{\text{succ}} = \mathbb{P}(\gamma_i > \gamma_{\text{th}}) \quad (3)$$

The edge node computes reliability as follows:

$$R_i = 1 - P_{\text{loss}}(i) \cdot f(D_i)$$

where $f(D_i)$ is a latency penalty function. If $R_i < R_{\min}$, corrective actions like redundancy or early retransmission are triggered.

3.4. 5G Compatibility

To ensure compatibility with real-world 5G/B5G deployments, the proposed framework is designed to operate within the 5G URLLC network slice, as shown in Figure 3. In this architecture, the edge node acts as the local slice orchestrator, interfacing with the 5G core via the Network Exposure Function (NEF) and Policy Control Function (PCF). URLLC service characteristics, such as latency budgets and reliability thresholds, are enforced using standardized 5G QoS Identifiers (5QI) and slicing APIs. The proposed modules (scheduler, control, and reliability manager) can be mapped to application and control functions running on the edge, leveraging service-level agreements (SLAs) defined in the network slice descriptor. This alignment allows seamless deployment of the proposed framework within the 5G/6G slicing infrastructure while preserving interoperability and real-time responsiveness.

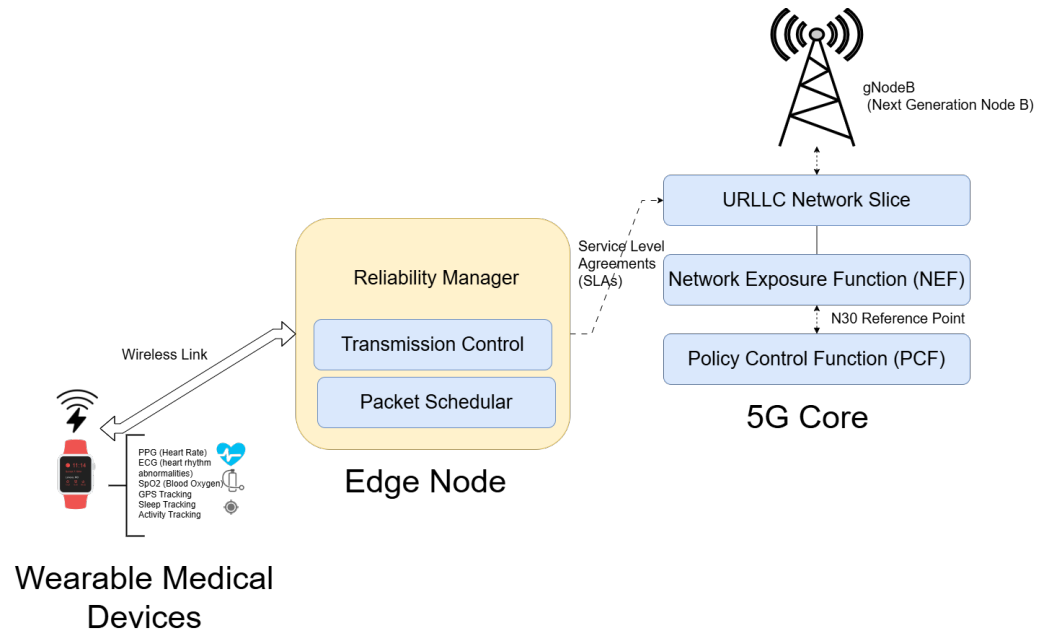


Figure 3. Integration of energy-aware URLLC framework within a 5G network architecture.

4. Proposed Energy-Aware URLLC Framework

To meet the URLLC requirements of H-IoT devices, we propose an energy-aware URLLC architecture. The architecture consists of three coordinated modules: priority-aware packet scheduler, adaptive transmission control, and edge-assisted reliability manager. These modules operate collaboratively between the device and edge layers to offer intelligent, power-efficient, and QoS-aware communication.

To formally capture the design goal, we define a mathematical optimization model that seeks to minimize the average energy consumed across devices while meeting latency, reliability, and fairness constraints. Let E_i denote the energy consumed by device i for transmitting a packet, let D_i be the observed end-to-end delay, $T_i^{\max}$ be the delay threshold, let R_i be the reliability score, and let $R_{\min}$ be the minimum required reliability. We also denote $x_i \in \{0, 1\}$ as the binary decision variable indicating whether the packet is transmitted or not. Jain's fairness index (JFI) ensures equitable scheduling.

The optimization problem is defined as

$$\min_{\{x_i\}} \frac{1}{N} \sum_{i=1}^N E_i \cdot x_i, \quad (4)$$

and it is subject

$$D_i \leq T_i^{\max}, \quad \forall i \quad (\text{Latency constraint}) \quad (5)$$

$$R_i \geq R_{\min}, \quad \forall i \quad (\text{Reliability constraint}) \quad (6)$$

$$\text{JFI}(x_1, \dots, x_N) \geq \theta, \quad (\text{Fairness constraint}) \quad (7)$$

$$x_i \in \{0, 1\}, \quad \forall i \quad (\text{Binary decision}) \quad (8)$$

Here, $E_i = P_i^{\text{tx}} \cdot t_i^{\text{tx}}$, where P_i^{tx} is the transmission power and t_i^{tx} is the duration of the transmission. The fairness constraint uses Jain's index to maintain throughput balance across users. Due to the discrete nature of x_i and non-linearity in JFI, this problem is NP-hard. Therefore, we resort to a lightweight heuristic presented in Algorithm 1, which achieves a near-optimal solution with practical complexity.

The priority-aware packet scheduler module is situated at the edge node and handles the scheduling of incoming packets based on urgency, current battery level, and latency deadlines. Each packet p_i is tagged with a weighted urgency value U_i , which is calculated

using Equation (1). Packets are prioritized by priority classes, and the scheduler always selects the most critical packet. In case of a tie when several packets have the same priority, the packet with the smallest latency deadline $T_i^{\max}$ is prioritized. This ensures the timely delivery of emergency data and saves energy by delaying non-critical transmissions from low-energy nodes.

Algorithm 1 Energy-aware URLLC packet scheduling.

```

1: for each time slot do
2:   for each packet  $p_i$  arriving at the edge do
3:     Compute urgency score:  $U_i = \frac{w_i}{B_i(t) + \epsilon}$ 
4:     Place  $p_i$  in the corresponding priority queue
5:   end for
6:   Select packet  $p_j$  with highest  $U_j$  among all queues
7:   Evaluate reliability score:  $R_j = 1 - P_{\text{loss}}(j) \cdot f(D_j)$ 
8:   if  $R_j < R_{\min}$  then
9:     Apply retransmission, multipath routing, or buffering
10:  end if
11:  Transmit  $p_j$  using profile from  $f(T_j^{\max}, B_j(t), \gamma_j)$ 
12: end for

```

The adaptive transmission control module executes on every H-IoT device and selects the most suitable transmission configuration based on the evaluation of three factors: the latency requirement, residual energy of the device, and the present channel state. On the basis of this evaluation, the device chooses from a collection of predefined profiles consisting of modulation schemes and power levels. For instance, emergency high-power QPSK traffic may be used to initiate an event, packets of a semi-critical nature may be delivered at medium power via QAM, and non-critical information may be delivered via BPSK with lower transmission power. A dynamic adaptation will render power consumption efficient with no trade-off on reliability for high-priority communications. The selected transmission profile is defined as a function in Equation (9), as follows:

$$\text{Profile}_i = f(T_i^{\max}, B_i(t), \gamma_i) \quad (9)$$

In this equation, the local SNR observed at the device is denoted as γ_i .

The edge-assisted reliability manager module executes on the edge and is tasked with imposing reliability. It calculates the reliability score for each packet using Equation (10), as follows:

$$R_i = 1 - P_{\text{loss}}(i) \cdot f(D_i) \quad (10)$$

where $P_{\text{loss}}(i)$ represents the estimated probability of packet loss and $f(D_i)$ is a delay penalty function, e.g., $e^{\alpha D_i}$. When the reliability value of a packet falls below a threshold value $R_{\min}$, the edge node takes corrective actions such as premature retransmission, multipath redundancy, or priority-based buffering. The edge-assisted reliability manager module adjusts dynamically by using history, recent trends in latency, and channel conditions while keeping URLLC conformance.

The collaborative operation of the three modules enables real-time responsiveness together with energy-aware behavior. Following the generation of the packet, the H-IoT device utilizes adaptive transmission control logic to select an optimal transmission mode. The packet is then sent to the edge, where the priority-aware packet scheduler module estimates its urgency score and schedules it according to it. When the transmission opportunity is released, the edge-assisted reliability manager module inspects the packet's reliability and takes corrective measures if necessary. This interaction ensures that each packet is

treated according to its QoS requirements without sacrificing system-wide performance. The overall logic governing these interactions is summarized in Algorithm 1.

5. Simulation Setup and Evaluation Metrics

To evaluate the performance of the proposed energy-aware URLLC framework, a discrete-event simulator was developed to model realistic network dynamics under varying H-IoT conditions. Monte Carlo simulations were employed to ensure robustness and statistical consistency. This method enables results to be averaged over different random traffic patterns and channel realizations, helping to mitigate the effects of stochastic variability and ensuring fair comparison across models. The simulation environment emulates a smart healthcare deployment, where wearable and implantable medical devices transmit physiological data to an MEC-enabled edge node. Data packets are generated according to a Poisson process and classified into emergency, semi-critical, and non-critical types, each with specific latency and reliability constraints. The wireless communication channel between devices and the edge is modeled using a block-fading Nakagami- m distribution, which is widely used to characterize small-scale fading in body-centric wireless communication. Nakagami- m offers greater flexibility than Rayleigh or Rician models, enabling accurate modeling of both line-of-sight and non-line-of-sight conditions commonly encountered in indoor hospital environments. Devices are battery-powered, and transmission policies are constrained by residual energy levels.

During operation, each device applies adaptive transmission control based on current energy availability, traffic priority, and estimated channel conditions. The edge node executes urgency-based scheduling and invokes reliability-enhancing mechanisms such as retransmissions or redundancy when necessary. The simulation tracks multiple performance indicators, including delay violation rate, energy consumption, reliability score, average throughput, and fairness index. The key parameters used in the simulation are summarized in Table 2.

Table 2. Simulation Parameters.

Parameter	Value
Number of H-IoT devices	30
Network load (λ)	0.1 to 1.0 (step size: 0.1)
Edge computing delay (d_e)	1 ms, 5 ms, 10 ms
Channel model	Nakagami- m (block fading)
Traffic arrival model	Poisson process
Traffic types	Emergency, Semi-Critical, Non-Critical
Modulation schemes	BPSK, QPSK, 16-QAM
Battery capacity	300 mAh per device

6. Results and Discussion

In this section, we present the results of our Monte Carlo simulations and compare the performance of the proposed energy-aware URLLC framework with three baseline schemes: fixed scheduling, where all packets are transmitted with fixed modulation using QPSK and medium power, ignoring traffic type; ALOHA-based random access, where packets are transmitted immediately without coordination or scheduling, leading to collisions; and priority-only URLLC, where packets are scheduled strictly by priority level without considering energy constraints or channel conditions. These models represent widely adopted benchmarks in the literature, and their operational logic is summarized in Appendix A for clarity.

6.1. Average Latency

To evaluate delay performance, we measured the average end-to-end latency, which is defined as the time from packet generation to successful delivery at the edge node, and we evaluated each scheme under varying network loads. This metric is critical in URLLC scenarios, particularly in time-sensitive H-IoT applications such as emergency response and remote surgery.

Figure 4 presents the latency trends as a function of network load. The proposed model consistently outperforms baseline schemes across all traffic intensities by leveraging latency-aware prioritization, channel-adaptive transmission gating, and congestion filtering for non-critical traffic. At a 0.1 network load, the proposed framework maintains an average latency of 12.92 ms compared to 23.14 ms for ALOHA, 24.28 ms for priority-only, and 25.11 ms for the fixed model. This translates to a reduction of 44.16% over ALOHA, 46.78% over priority-only, and 48.54% over the fixed model.

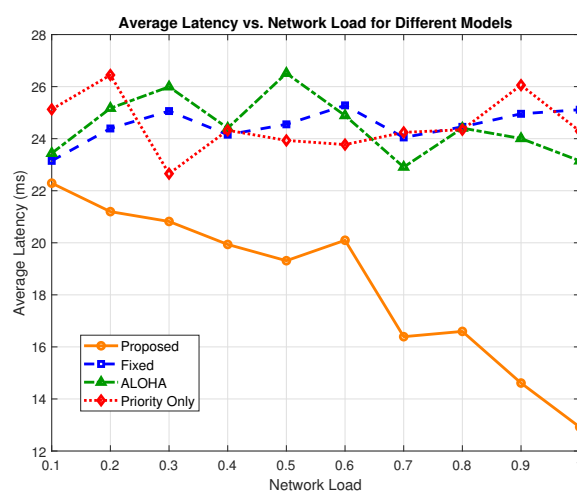


Figure 4. Average latency vs. network load.

Figure 5 shows the average latency across all load levels. The proposed model achieves an overall average of 18.41, while the ALOHA, fixed, and priority-only models yield 24.48 ms, 24.515 ms, and 24.519 ms, respectively. This corresponds to an average latency improvement of 32.97% over ALOHA, 33.16% over the fixed model, and 33.18% over the priority-only model.

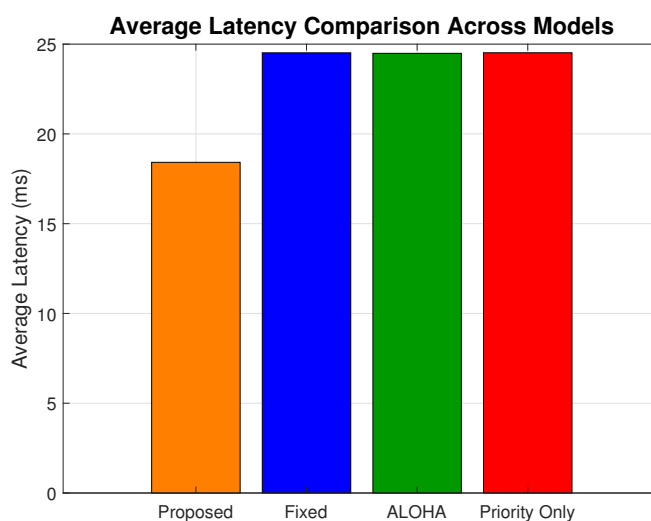


Figure 5. Average latency comparison across models.

6.2. Average Throughput

Average throughput, measured in kbps, represents the total amount of successfully received data at the edge per unit time. Figure 6 illustrates the throughput performance across varying network loads for all evaluated models. The proposed model incorporates reliability-aware retransmissions and adaptive admission control, including batched non-critical traffic and probabilistic semi-critical handling. These enhancements enable a more stable throughput profile while maintaining QoS constraints essential for H-IoT. At a 0.8 network load, the proposed scheme achieves a throughput of 481.62 kbps. In comparison, the priority-only model achieves 434.88 kbps, the fixed model achieves 441.70 kbps, and the ALOHA model trails slightly at 398.91 kbps. These values demonstrate that the proposed scheme delivers a 10.74% improvement over the priority-only model, a 9.03% improvement over the fixed model, and a 20.73% improvement over ALOHA.

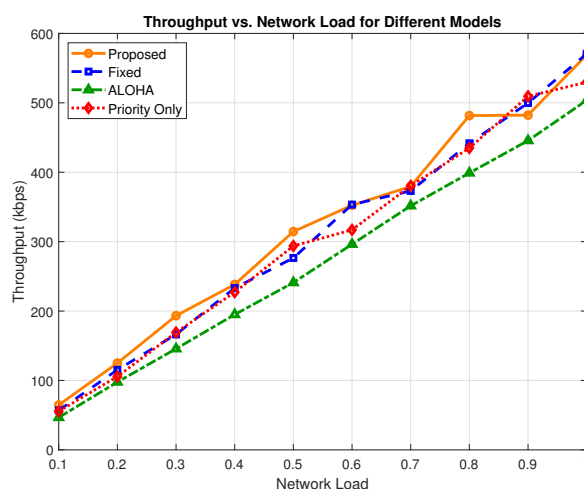


Figure 6. Throughput vs. network load.

Figure 7 presents the average throughput across all load levels. The proposed scheme achieves an overall average of 320.05 kbps, outperforming the fixed (308.64 kbps), ALOHA (272.18 kbps), and priority-only (302.26 kbps) models. This corresponds to a 17.58% gain over ALOHA, a 5.88% gain over the priority-only model, and a 3.69% gain over the fixed model.

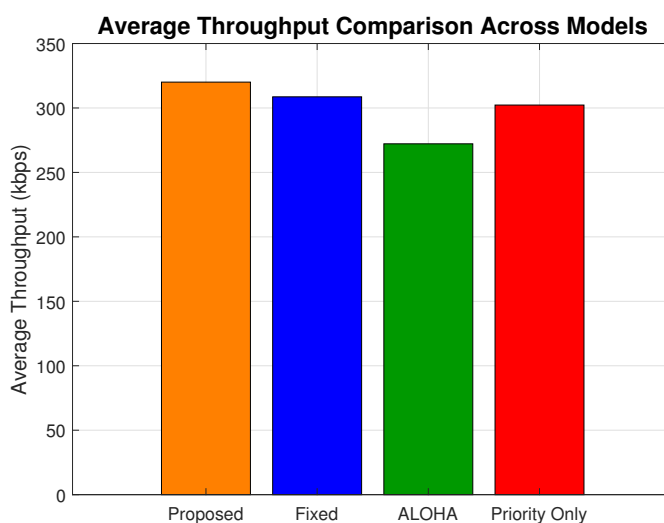


Figure 7. Average throughput comparison across models.

6.3. Reliability Score

Figure 8 illustrates the reliability score (R_s) for all evaluated models across varying network load levels. The R_s metric offers a unified evaluation by incorporating three key factors: packet delivery success, latency, and energy consumption. It is defined as follows:

$$R_s = \frac{(1 - \text{Loss Rate})}{\text{Latency} \times \text{Energy}},$$

where a higher R_s implies superior performance in delivering reliable, timely, and energy-efficient communication.

The proposed model integrates delivery-aware transmission filtering for semi-critical traffic, selectively allowing transmissions only under favorable channel and latency conditions. This enhancement significantly improves the ratio of successful packet delivery to energy-delay cost, especially in mid-to-high load scenarios.

As seen in Figure 8, the proposed scheme exhibits a better reliability profile than other models across all loads. At maximum load, the proposed model achieves an R_s value of 3.41, compared to 2.34 for the fixed model, 1.73 for the priority-only model, and 2.49 for ALOHA. The values translate to a 45.72% gain over the fixed model, a 97.10% gain over the priority-only model, and a 36.94% gain over ALOHA.

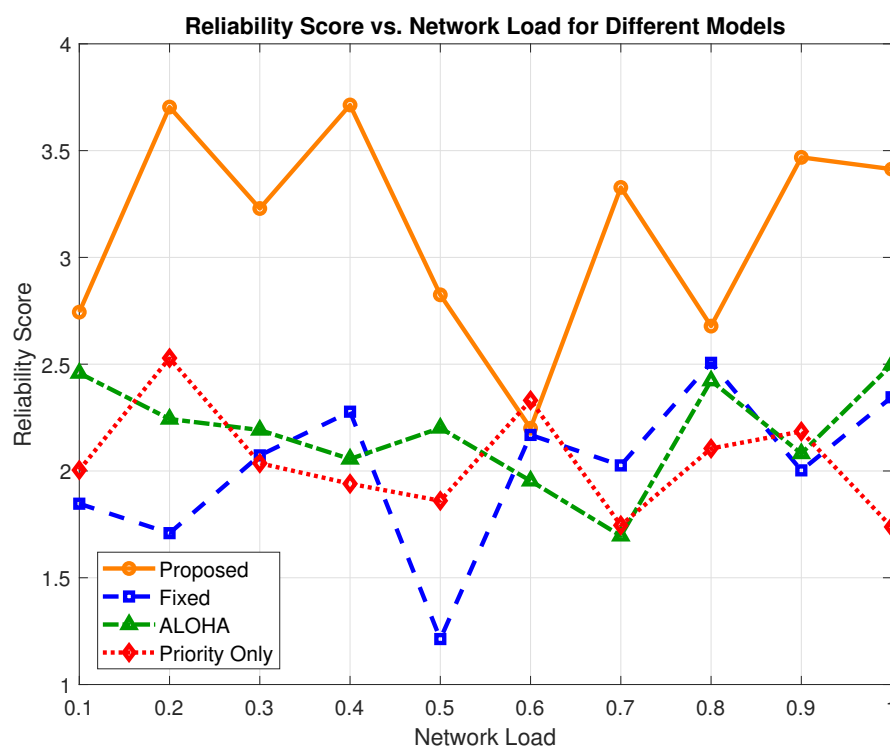


Figure 8. Reliability score (R_s) vs. network load for all evaluated models.

Figure 9 provides a bar chart comparison of the average R_s values across all load conditions. The proposed model achieves an average reliability score of 3.13, representing a 53.43% improvement over the priority-only model, which scores a value of 2.04, a 60.95% gain over the fixed model, which scores a value of 1.95, and a 46.94% gain over the ALOHA model's average of 2.13.

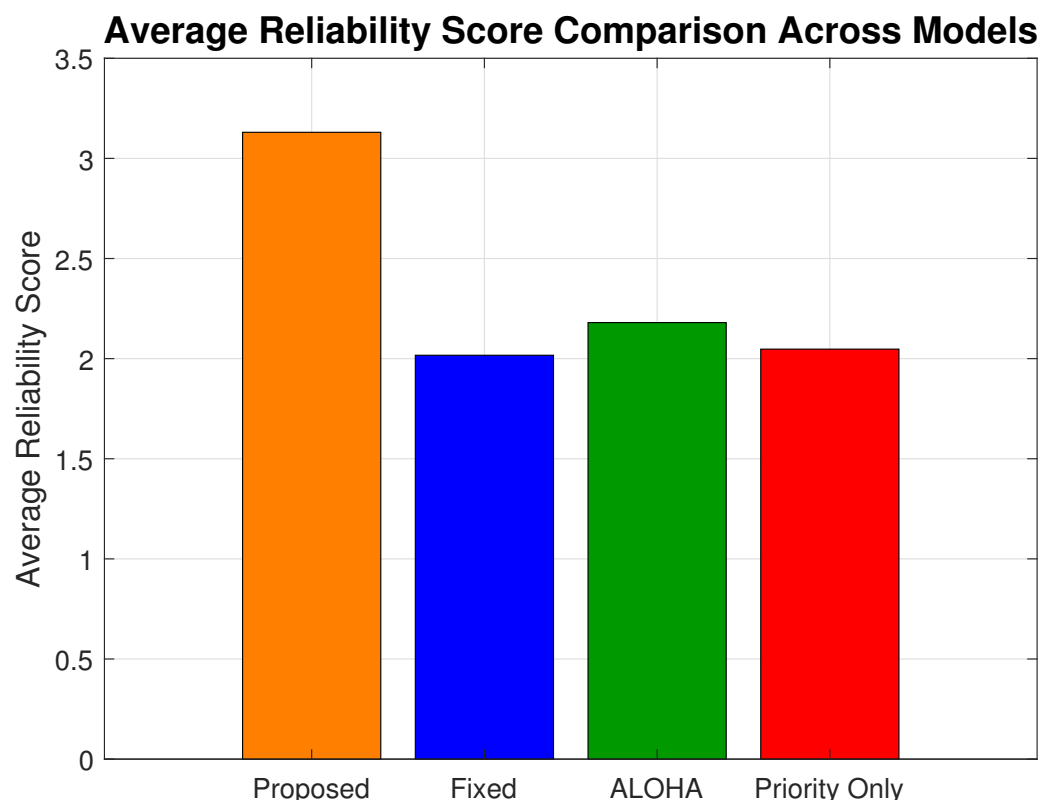


Figure 9. Average reliability score of all models.

6.4. Energy Consumption

To assess the energy efficiency of the proposed URLLC framework, we evaluated the average energy consumption per device across various network loads. The observed energy savings stem from three key design features: (i) adaptive modulation and power control that dynamically selects energy-efficient transmission profiles based on residual battery and channel conditions; (ii) battery-aware traffic suppression, which prevents non-critical transmissions when device energy falls below 30% of capacity; and (iii) congestion-aware filtering, where semi-critical traffic is selectively delayed or dropped under a high network load to avoid redundant energy expenditure. Together, these mechanisms minimize unnecessary transmissions and prioritize energy use only for delay and reliability-sensitive traffic, thereby ensuring long-term sustainability of battery-powered H-IoT devices.

As shown in Figure 10, the proposed model consistently demonstrates improved energy efficiency compared to the fixed, ALOHA, and priority-only schemes. At a maximum network load, the proposed scheme achieves an average energy consumption of 15.34 mJ, whereas the fixed model consumes 25.93 mJ, ALOHA consumes 22.68 mJ, and the priority-only model consumes 25.54 mJ. This translates to a reduction of 40.84% over the fixed model, 32.36% over ALOHA, and 39.93% over the priority-only model at peak load.

Figure 11 presents the overall average energy consumption across all load conditions. The proposed scheme consumes 9.73 mJ on average, while the fixed, priority-only, and ALOHA models consume 14.10 mJ, 14.08 mJ, and 12.60 mJ, respectively. The results translate to a reduction of 30.99% as compared to the fixed model, 30.89% as compared to the priority-only model, and 22.77% as compared to ALOHA.

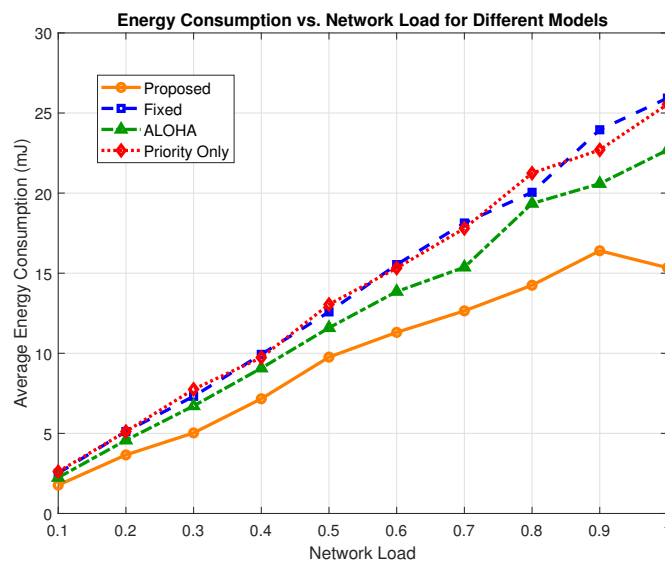


Figure 10. Average energy consumption (mJ) across network loads for all models.

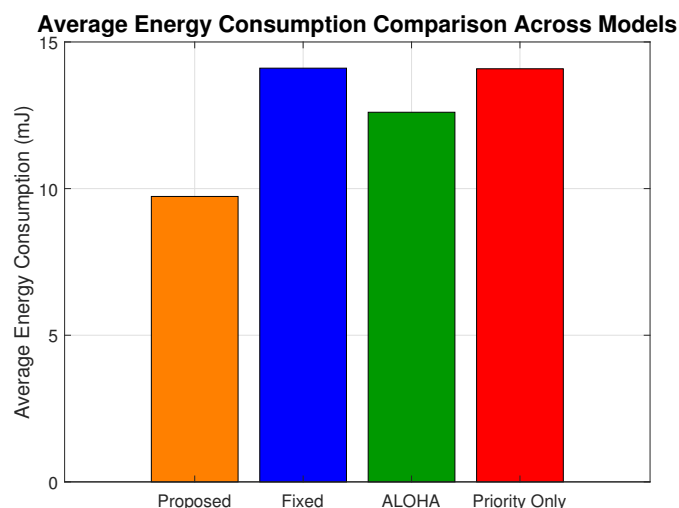


Figure 11. Mean energy consumption comparison across models.

6.5. Delay Violation Rate

To evaluate how well each scheme accommodates the stringent latency demands of heterogeneous H-IoT traffic, we measured the delay violation rate, defined as the proportion of packets whose end-to-end delay exceeds the maximum allowable threshold for their respective traffic class, i.e., emergency, semi-critical, or non-critical.

As illustrated in Figure 12, the proposed model consistently achieves lower delay violation rates across all network load conditions. This improvement is driven by its refined priority-aware scheduling strategy, reduced edge delay for urgent packets, and selective suppression of non-critical traffic under heavy congestion. At the highest network load, the proposed model registers a violation rate of 0.368, which is a significant reduction compared to the priority-only (0.551), ALOHA (0.631), and fixed (0.559) schemes. This translates to a relative improvement of up to 71.4% over the ALOHA, 51.9% over fixed, and 49.7% over priority-only models.

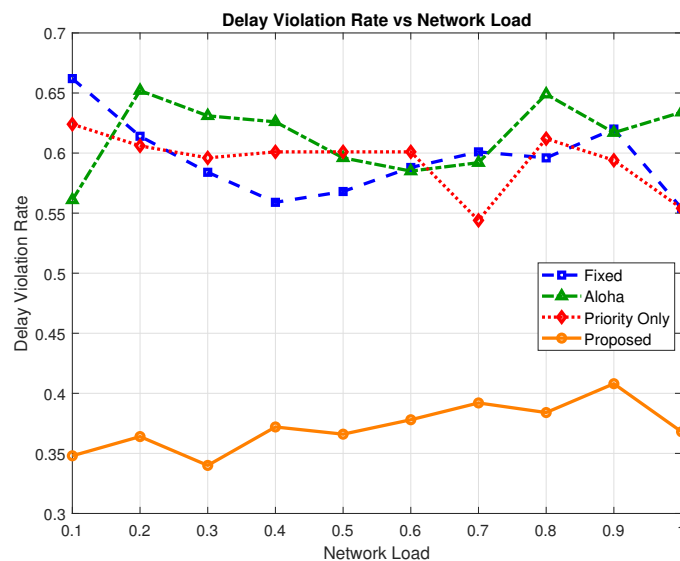


Figure 12. Delay violation rate vs. network load.

Figure 13 presents the average delay violation rate across all traffic types and load conditions. The proposed scheme achieves the lowest average value of 0.371, significantly outperforming the ALOHA (0.612), priority-only (0.590), and fixed (0.593) models. This represents an average increase of 64.9% relative to ALOHA, over 59% relative to the priority-only model, and 59.8% over the fixed model. The results validate the effectiveness of the proposed schemes as compared to the baseline schemes.

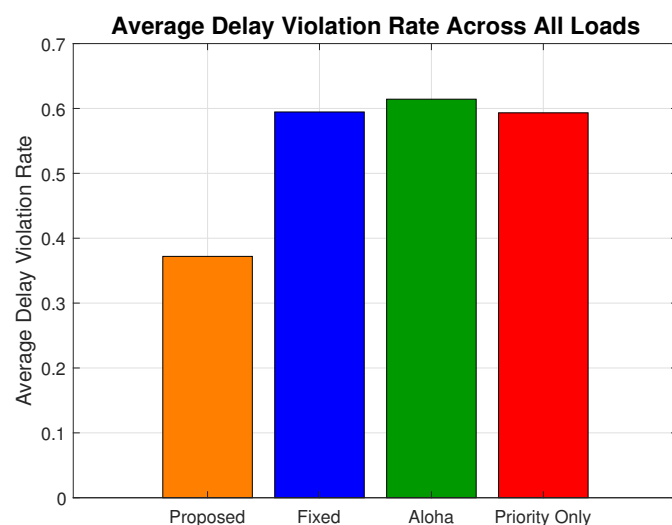


Figure 13. Average delay violation rate comparison.

6.6. Fairness Index

To assess equitable resource distribution across all sensors, we evaluated Jain's fairness index (JFI) under varying network loads. JFI measures fairness in throughput among devices as follows:

$$J(x_1, x_2, \dots, x_n) = \frac{(\sum x_i)^2}{n \cdot \sum x_i^2}$$

where x_i is the throughput of device i . JFI provides a normalized measure of throughput fairness, where a value close to 1 indicates balanced resource sharing among all devices.

Figure 14 illustrates the fairness trends across all evaluated models. The proposed scheme performs relatively similarly to all other schemes from the load value of 0 to 0.6. At a 0.6 load, the value of the proposed scheme is 0.711, the fixed scheme is 0.719, ALOHA is 0.722, and the priority-only scheme is 0.726. These values translate to a compromise of 1.11% as compared to the fixed model, 1.5% as compared to ALOHA, and 2.06% as compared to the priority-only model. At higher loads from 0.6 to 1, the value of the proposed scheme significantly drops due to the fact that at higher loads, the high-priority traffic is given precedence or lower-priority traffic, which reduces the value of JFI of the proposed scheme at higher loads.

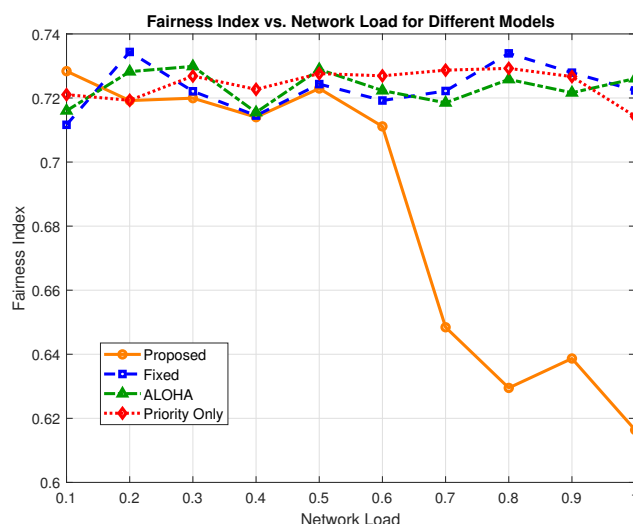


Figure 14. Fairness index vs. network load.

To address this, the proposed framework integrates a proportional fairness boost, where underperforming sensors (i.e., those with below-average throughput) receive a temporary transmission benefit by slightly relaxing their channel thresholds. This lightweight mechanism significantly improves fairness without altering the core logic or compromising reliability and delay performance.

Figure 15 provides the average fairness index value. The proposed scheme achieves a fairness index of 0.68, compared to 0.72 for ALOHA, fixed, and priority-only models, which translates to a 5.5% compromise on JFI.

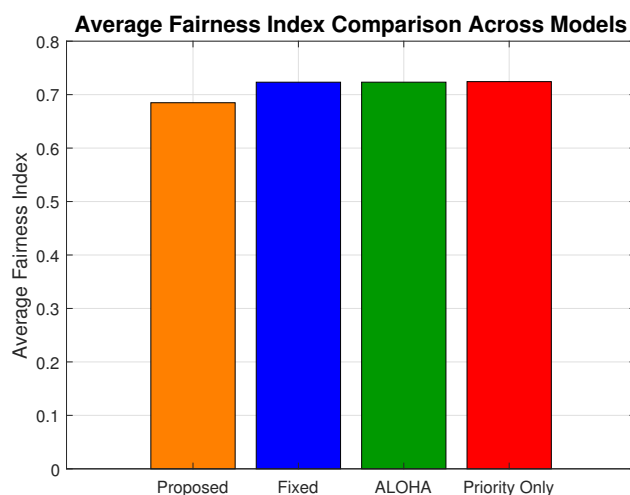


Figure 15. Average fairness index comparison.

While a marginal fairness gap persists, this is a trade-off to ensure energy efficiency and ultra-reliable low-latency performance and achieve critical design goals in H-IoT networks. The proposed scheme strikes a balanced compromise, providing consistent access while safeguarding the integrity of time-sensitive medical data.

A consolidated comparison across all performance indicators is presented in Table 3, which highlights the improvements achieved by the proposed scheme relative to baseline models. Based on the observations derived, the proposed framework proves to maintain an effective balance among energy efficiency, latency guaranteeing, and equity in URLLC applications catering to H-IoT. With a consideration of adaptive scheduling and the development of fairness-improving functionalities, the model is found to be appropriate for the extremely stringent demands posed by next-generation wireless systems. Although there are some compromises to fairness at higher loads, these are offset by the more robust delay and reliability performance that underlies its practical usefulness in H-IoT environments. As wireless communications continue to evolve, this approach opens the door to healthy, scalable, and resilient URLLC solutions in life-critical healthcare applications.

Table 3. Performance comparison of URLLC schemes in H-IoT systems.

Metric	Fixed	ALOHA	Priority-Only	Proposed Model
Average Latency (ms)	24.515	24.48	24.519	18.41
Average Throughput (kbps)	308.64	272.18	302.26	320.05
Reliability Score (R_s)	1.95	2.13	2.04	3.13
Energy Consumption (mJ)	14.10	12.60	14.08	9.73
Delay Violation Rate	0.593	0.612	0.590	0.371
Jain's Fairness Index	0.72	0.72	0.72	0.68

7. Conclusions and Future Work

This paper suggests an energy-efficient URLLC system for beyond 5G and future 6G-enabled H-IoT systems. The proposed system includes three essential modules: a priority-aware packet scheduler, an adaptive transmission control, and an edge-assisted reliability manager, to tackle latency, reliability, and energy efficiency in a comprehensive manner. Simulation results demonstrate that the new model surpasses the fixed, ALOHA, and priority-only baselines across all key metrics. It improves the latency, throughput, reliability score, energy consumption, and delay violation rate, while the fairness index shows a slight compromise at higher loads due to prioritization of critical traffic. This trade-off highlights a promising direction for future enhancement.

The framework ensures timely delivery of critical data, resource-optimized usage, and equitable access to all device classes. Future work may focus on improving fairness through adaptive learning, incorporating energy-harvesting mechanisms to extend device longevity, and validating the framework in real-world, large-scale H-IoT deployments with diverse traffic and device heterogeneity. These findings are promising for developing scalable, green, and QoS-aware communication solutions in next-generation healthcare systems.

Author Contributions: Conceptualization, A.I.; simulations, A.I., A.N. and T.K.; writing—original draft preparation, A.I. and T.K.; writing—review and editing, A.I., A.N. and S.-B.R.; supervision, S.-B.R. 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 dataset used in this study can be made available upon request from the authors.

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

Appendix A. Pseudocode for Baseline Comparison Schemes

Appendix A.1. Fixed Scheduling

Algorithm A1 Fixed Scheduling

```

1: for each device  $d_i$  do
2:   for each generated packet  $p_i$  do
3:     Set modulation = QPSK, power = medium
4:     Transmit  $p_i$  immediately
5:   end for
6: end for

```

Appendix A.2. ALOHA-Based Random Access

Algorithm A2 ALOHA Random Access

```

1: for each device  $d_i$  do
2:   for each generated packet  $p_i$  do
3:     Transmit  $p_i$  without sensing or scheduling
4:     If collision occurs, retransmit after random backoff
5:   end for
6: end for

```

Appendix A.3. Priority-Only URLLC

Algorithm A3 Priority-Based Scheduling

```

1: for each time slot do
2:   Sort incoming packets by priority level  $w_i$ 
3:   Schedule highest priority packet first
4:   Transmit using fixed modulation and power
5: end for

```

References

1. Yang, P.; Kong, L.; Chen, G. Spectrum sharing for 5G/6G URLLC: Research frontiers and standards. *IEEE Commun. Stand. Mag.* **2021**, *5*, 120–125. [\[CrossRef\]](#)
2. Rahim, M.; Basit, A.; Nguyen, T.L.; Kaddoum, G. Maximizing URLLC Reliability Through JPSSA for URLLC Services in IRS-Aided Terahertz Networks. *Authorea Prepr.* **2025**, preprint.
3. Qadri, Y.A.; Nauman, A.; Zikria, Y.B.; Vasilakos, A.V.; Kim, S.W. The future of healthcare internet of things: A survey of emerging technologies. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1121–1167. [\[CrossRef\]](#)
4. Series, M. *Guidelines for Evaluation of Radio Interface Technologies for IMT-2020*; Report ITU-R M.2412-0; International Telecommunication Union: Geneva, Switzerland, 2017; p. 2512.
5. Baek, S.; Kim, D.; Tesanovic, M.; Agiwal, A. 3GPP new radio release 16: Evolution of 5G for industrial internet of things. *IEEE Commun. Mag.* **2021**, *59*, 41–47. [\[CrossRef\]](#)
6. Hamidi-Sepehr, F.; Sajadieh, M.; Pantelev, S.; Islam, T.; Karls, I.; Chatterjee, D.; Ansari, J. 5G URLLC: Evolution of high-performance wireless networking for industrial automation. *IEEE Commun. Stand. Mag.* **2021**, *5*, 132–140. [\[CrossRef\]](#)
7. Ranjha, A.; Kaddoum, G.; Rahim, M.; Dev, K. URLLC in UAV-enabled multicasting systems: A dual time and energy minimization problem using UAV speed, altitude and beamwidth. *Comput. Commun.* **2022**, *187*, 125–133. [\[CrossRef\]](#)
8. Wu, Q.; Wang, W.; Fan, P.; Fan, Q.; Wang, J.; Letaief, K.B. URLLC-awared resource allocation for heterogeneous vehicular edge computing. *IEEE Trans. Veh. Technol.* **2024**, *73*, 11789–11805. [\[CrossRef\]](#)
9. Yadav, R.; Zhang, W.; Elgendy, I.A.; Dong, G.; Shafiq, M.; Laghari, A.A.; Prakash, S. Smart healthcare: RL-based task offloading scheme for edge-enable sensor networks. *IEEE Sens. J.* **2021**, *21*, 24910–24918. [\[CrossRef\]](#)

10. Iqbal, A.; Nauman, A.; Qadri, Y.A.; Kim, S.W. Optimizing Spectral Utilization in Healthcare Internet of Things. *Sensors* **2025**, *25*, 615. [\[CrossRef\]](#)
11. Nagarajan, S.M.; Devarajan, G.G.; Mohammed, A.S.; Ramana, T.; Ghosh, U. Intelligent task scheduling approach for IoT integrated healthcare cyber physical systems. *IEEE Trans. Netw. Sci. Eng.* **2022**, *10*, 2429–2438. [\[CrossRef\]](#)
12. Iqbal, N.; Imran, A.; Ahmad, S.; Ahmad, R.; Kim, D.H. A scheduling mechanism based on optimization using IoT-tasks orchestration for efficient patient health monitoring. *Sensors* **2021**, *21*, 5430. [\[CrossRef\]](#)
13. Taimoor, N.; Rehman, S. Reliable and resilient AI and IoT-based personalised healthcare services: A survey. *IEEE Access* **2021**, *10*, 535–563. [\[CrossRef\]](#)
14. Rahmani, A.M.; Babaei, Z.; Souri, A. Event-driven IoT architecture for data analysis of reliable healthcare application using complex event processing. *Clust. Comput.* **2021**, *24*, 1347–1360. [\[CrossRef\]](#)
15. Meng, J.; Tan, H.; Li, X.Y.; Han, Z.; Li, B. Online deadline-aware task dispatching and scheduling in edge computing. *IEEE Trans. Parallel Distrib. Syst.* **2019**, *31*, 1270–1286. [\[CrossRef\]](#)
16. Haque, M.E.; Tariq, F.; Khandaker, M.R.; Wong, K.K.; Zhang, Y. A survey of scheduling in 5G URLLC and outlook for emerging 6G systems. *IEEE Access* **2023**, *11*, 34372–34396. [\[CrossRef\]](#)
17. Munari, A. Modern random access: An age of information perspective on irregular repetition slotted ALOHA. *IEEE Trans. Commun.* **2021**, *69*, 3572–3585. [\[CrossRef\]](#)
18. Chen, Y.; Wang, G.; Yi, H.; Zhang, W. Priority-based distributed queuing random access mechanism for mMTC/uRLLC terminals coexistence. In Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Helsinki, Finland, 25–28 April 2021; pp. 1–6.
19. Elgarhy, O.; Reggiani, L.; Alam, M.M.; Zoha, A.; Ahmad, R.; Kuusik, A. Energy efficiency and latency optimization for IoT URLLC and mMTC use cases. *IEEE Access* **2024**, *12*, 23132–23148. [\[CrossRef\]](#)
20. Kumar, A.; Gaur, N.; Nanthamornphong, A. Improving the latency for 5G/B5G based smart healthcare connectivity in rural area. *Sci. Rep.* **2024**, *14*, 6976. [\[CrossRef\]](#)
21. Su, X.; Liu, P.; Wang, Z.; Liao, H.; Zhou, Z.; Xu, G.; Wang, Z.; Qin, Y. Priority-aware intelligent device access management for carbon footprint monitoring in sustainable cities and society. *IET Commun.* **2023**, *17*, 409–417. [\[CrossRef\]](#)
22. Abba Ari, A.A.; Samafou, F.; Ndam Njoya, A.; Djedouboum, A.C.; Aboubakar, M.; Mohamadou, A. IoT-5G and B5G/6G resource allocation and network slicing orchestration using learning algorithms. *IET Netw.* **2025**, *14*, e70002. [\[CrossRef\]](#)
23. Liu, Y.; Lee, J.; Han, Y.; Chen, W. Lyapunov drift-based scheduling for short-packet transmission with finite blocklength coding. In Proceedings of the GLOBECOM 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; pp. 2068–2073.
24. Roobini, S.; Jacob, L. Priority Aware and Spectrum Efficient Scheduling of Co-Existing Wireless Body Area Networks. *Wirel. Pers. Commun.* **2022**, *122*, 3371–3392. [\[CrossRef\]](#)
25. Asingwire, B.K.; Sibomana, L.; Ngenzi, A.; Kabiri, C. Delay and size-dependent priority-aware scheduling for IoT-based healthcare traffic using heterogeneous multi-server priority queueing system. *IET Commun.* **2023**, *17*, 1877–1887. [\[CrossRef\]](#)
26. Zhou, Z.; Wang, Z.; Yu, H.; Liao, H.; Mumtaz, S.; Oliveira, L.; Frascolla, V. Learning-based URLLC-aware task offloading for Internet of Health Things. *IEEE J. Sel. Areas Commun.* **2020**, *39*, 396–410. [\[CrossRef\]](#)
27. Cadence PCB Solutions. URLLC Use Cases in Healthcare and Beyond. 2023. Available online: <https://resources.pcb.cadence.com/blog/2023-urllc-use-cases-in-healthcare-and-beyond> (accessed on 7 April 2025).
28. Humayun, M.; Hamid, B.; Jhanjhi, N.; Suseendran, G.; Talib, M. 5G network security issues, challenges, opportunities and future directions: A survey. *J. Physics: Conf. Ser.* **2021**, *1979*, 012037. [\[CrossRef\]](#)

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.