

Article

Large AI Model-Enhanced Digital Twin-Driven 6G Healthcare IoE

Haoyuan Hu, Ziyi Song and Wenzao Shi *

College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China; 136152023067@student.fjnu.edu.cn (H.H.); 136132024017@student.fjnu.edu.cn (Z.S.)

* Correspondence: swz@fjnu.edu.cn

Abstract

The convergence of the Internet of Everything (IoE) and healthcare requires ultra-reliable, low-latency, and intelligent communication systems. Sixth-generation (6G) wireless networks, coupled with digital twin (DT) models and large AI models (LAMs), are envisioned to promise substantial and practically meaningful improvements in smart healthcare by enabling real-time monitoring, diagnosis, and personalized treatment. In this article, we propose an LAM-enhanced DT-driven network slicing framework for healthcare applications. The framework leverages large models to provide predictive insights and adaptive orchestration by creating virtual replicas of patients and medical devices that guide dynamic slice allocation. Reinforcement learning (RL) techniques are employed to optimize slice orchestration under uncertain traffic conditions, with LAMs augmenting decision-making through cognitive-level reasoning. Numerical results show that the proposed LAM-DT-RL framework reduces service-level agreement (SLA) violations by approximately 42–43% compared to a reinforcement-learning-only slicing strategy, while improving spectral efficiency and fairness among heterogeneous healthcare services. Finally, we outline open challenges and future research opportunities in integrating LAMs, DTs, and 6G for resilient healthcare IoE systems.

Keywords: 6G; large AI model; digital twin; IoE

1. Introduction

The digitalization of healthcare has accelerated in recent years, driven by the proliferation of wearable sensors, tele-medicine platforms, robot-assisted surgery, and AI-supported clinical diagnostics. This transformation is increasingly tied to the paradigm of the Internet of Everything (IoE), which integrates not only devices and data streams but also people, processes, and intelligent services into a unified system [1–4]. In this context, healthcare applications place highly stringent demands on communication infrastructures [5]. They require not only ultra-reliable and low-latency data delivery but also intelligent adaptation to patient-specific needs, seamless integration of multi-modal data, and the capacity to scale across heterogeneous hospital environments. Existing fifth-generation (5G) networks have enabled preliminary advances in tele-consultation, continuous patient monitoring, and mobile health applications, yet they are insufficient for emerging use cases such as remote robotic surgery, real-time digital twin-enabled diagnostics, and AI-assisted personalized treatment. These new demands point directly to the transformative potential of sixth-generation (6G) communication networks, which promise terabit-level throughput, sub-millisecond latency, and pervasive AI-native intelligence.



Academic Editor: Yung-Fa Huang

Received: 13 December 2025

Revised: 17 January 2026

Accepted: 21 January 2026

Published: 31 January 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.

One of the enabling technologies in this vision is network slicing [6–9], which allows multiple logical networks to coexist on shared physical infrastructure, each optimized for a specific quality of service (QoS) requirement. For instance, robotic surgery requires an ultra-reliable low-latency communication (URLLC) slice, while fitness tracking or wellness monitoring can be supported by slices optimized for energy efficiency and scalability. Digital twin (DT) technology adds another dimension by creating virtual replicas of patients, medical devices, and hospital environments that continuously synchronize with their physical counterparts through real-time data streams [10,11]. DTs allow physicians to simulate interventions, detect anomalies, and anticipate patient needs, thereby introducing predictive intelligence into the healthcare workflow. Together, 6G-enabled slicing and DTs can transform healthcare into a proactive, personalized, and adaptive system.

However, as powerful as these technologies are, they remain incomplete without the cognitive reasoning and multi-modal understanding provided by Large AI Models (LAMs). LAMs includes large language models (LLMs) such as GPT-4 and LLaMA, vision models, such as SAM and DINO, and multi-modal models such as Gemini, have demonstrated unprecedented capabilities in processing unstructured data, extracting semantic knowledge, and supporting context-aware decision-making [12–14]. In the healthcare IoE, these models can analyze complex multi-modal streams that include text from electronic health records, sensor readings from wearable devices, high-resolution medical images, and even clinical notes. Unlike traditional AI algorithms, which are often domain-specific and limited in scalability, LAMs can generalize across tasks, adapt to dynamic environments, and provide a semantic layer that empowers DTs and network slicing managers to reason at the cognitive level.

Despite their promise, integrating LAMs into healthcare IoE systems raises new challenges [15,16]. Healthcare traffic is inherently heterogeneous: critical haptic feedback in robotic surgery requires sub-millisecond latency, high-bandwidth imaging demands sustained throughput, and large populations of wearable sensors generate constant but delay-tolerant data streams. Static allocation policies or rule-based slicing are unable to adapt to these rapid fluctuations, resulting in resource under-utilization or service-level agreement (SLA) violations. Reinforcement learning (RL) offers adaptive decision-making by dynamically reallocating resources in response to traffic variations. Yet RL alone is reactive, learning from past interactions without sufficient foresight. DTs complement this reactivity with predictive insights, while LAMs further enrich the process by contextualizing predictions, reasoning over multi-modal data, and enabling proactive orchestration of healthcare IoE services.

In this article, we propose a Large AI Model-enhanced DT-driven network slicing framework tailored for healthcare applications in 6G-enabled IoE environments. The framework introduces a three-layer architecture: (1) the physical layer composed of patients, devices, and medical infrastructure; (2) the digital twin layer that mirrors these entities with realtime synchronization and predictive modeling; and (3) the network slicing layer, where reinforcement learning agents allocate resources dynamically, guided by LAM-augmented DT insights. By integrating these elements, the framework not only addresses latency, reliability, and efficiency requirements but also aligns communication resources with patient-specific health profiles, thereby moving toward precision healthcare. Simulation results confirm that the proposed approach significantly improves SLA compliance, spectral efficiency, and latency compared with baseline methods.

2. State of the Art

The integration of network slicing, DT technology, RL, and LAMs in healthcare represents an emerging frontier of research at the intersection of communications, artificial

intelligence, and medical informatics. This section reviews the state of the art in each of these areas, highlighting their potential and limitations when applied to the Internet of Everything (IoE) in the 6G era. By synthesizing these perspectives, we identify critical research gaps that motivate the proposed framework.

2.1. Network Slicing in Healthcare IoE

Network slicing has been widely recognized as a key enabler to support heterogeneous quality-of-service (QoS) requirements in 5G/6G-enabled IoE systems, including healthcare. Early works tailored to e-health mainly focus on mapping medical applications to different slice templates and demonstrating feasibility. Jain et al. propose a 5G network slice architecture for digital real-time healthcare, where biometric data from wearable devices are transported over dedicated slices and processed by a data-analytics framework to support remote diagnosis and tele-consultation [17]. Bërdufi et al. investigate how end-to-end network slicing can be leveraged to enable and enhance IoT-based e-health services, highlighting issues such as QoS isolation, mobility support, and interoperability with legacy systems [18].

More recent studies move towards 6G and large-scale IoT deployments. Alwakeel et al. introduce a 6G network slicing framework for smart city IoT, explicitly considering verticals such as e-health and discussing slice admission, scaling and monitoring functions at the management plane [19]. Abba Ari et al. review machine learning-based resource allocation and slicing orchestration schemes for IoT-5G and B5G/6G systems, emphasizing that AI-native control will be essential to meet diverse service-level agreements (SLAs) [20]. At the same time, several works point out that network slicing raises new security and privacy risks for healthcare applications, since patient data may traverse multiple virtualized functions with slice-specific attack surfaces [21].

Overall, existing network-slicing solutions for healthcare IoE primarily treat slices as static service templates with pre-defined QoS classes (e.g., eMBB-like tele-medicine vs. URLLC-like remote surgery). They rarely exploit rich cross-layer context—such as patient status predictions or semantic understanding of clinical workflows—to drive proactive slice orchestration.

2.2. Digital Twin for Smart Healthcare

Digital twins (DTs) are emerging as a central paradigm for smart healthcare, where virtual replicas of patients, organs, medical devices or hospital environments are synchronized with real-time data streams to enable monitoring, prediction and personalized treatment. Recent systematic surveys show a rapid growth of DT-based healthcare research since 2020, covering digital patient models, virtual ICU rooms, surgical planning, and hospital-wide operational twins [22]. Kabir et al. review DTs in healthcare IoT, identifying key use cases such as digital patient avatars for chronic-disease management, DT-enabled operating theatres, and device-level twins for predictive maintenance; they also emphasize the need for scalable data integration from heterogeneous IoT sensors [22]. Ringeval et al. conduct a meta-review of DT applications in healthcare and highlight three primary application clusters: patient-centric prediction and decision support, workflow optimization in hospitals, and population-level public-health modeling [23]. Security and privacy are another major concern. Hemdan et al. propose a “smart and secure healthcare with digital twins” framework combining blockchain and federated learning to protect DT data integrity and support decentralized model training [24].

These investigations demonstrate the promise of DTs for smart healthcare, but they typically operate at the application layer and focus on clinical decision-making or asset management, rather than closing the loop with communication network resource orchestration.

2.3. Role of Large AI Models in IoE Healthcare

Large AI models—most prominently large language models (LLMs) and large multimodal models—are rapidly transforming digital health. Recent surveys highlight that LLMs can summarize medical records, support clinical documentation, generate differential diagnoses, and provide patient-specific treatment recommendations by integrating unstructured notes, lab results and imaging reports [25]. Maity et al. systematically review LLMs in healthcare and categorize applications into decision support, documentation, patient engagement and biomedical research, while also emphasizing risks such as hallucinations and bias [26]. Zhou et al. discuss how foundation models can act as general-purpose engines for biomedical and healthcare tasks when combined with domain-specific fine-tuning and retrieval [25].

Beyond “stand-alone” clinical AI, several works begin to explore LLMs as interfaces or orchestrators for complex digital-health ecosystems. Imrie et al. propose using LLMs as natural language interfaces that mediate interactions between clinicians and AI tools, including digital twins and decision-support systems [27]. Zonayed et al. review the integration of machine learning and IoT in healthcare, pointing to the need for AI methods that can reason over continuous streams of sensor data from wearables and medical IoT devices [28]. Very recent position papers further argue that LLM-based health applications will have to satisfy medical device regulations and be embedded into safe clinical workflows, rather than used as isolated chatbots [29].

Despite these efforts, the role of large AI models in network-level control remains largely unexplored. Existing works mostly deploy LLMs at the application or clinical decision-support layer; there is little work on using LAMs as semantic engines that feed predictive and contextual features into network slicing or resource allocation mechanisms for IoE healthcare.

2.4. Research Gap and Motivation

Examining the above strands of literature reveals several gaps.

1. Simple treatment of network slicing and clinical intelligence. Current network-slicing solutions for healthcare IoE primarily focus on defining slice templates and guaranteeing QoS through traffic-centric monitoring and control [17,19]. At the same time, digital-twin and clinical AI research concentrates on patient modeling, disease prediction, and workflow optimization, largely decoupled from the underlying communication network management [22,23]. There is a lack of architectures that jointly consider DT-based predictions and network-slicing decisions in a closed loop.
2. Under-utilization of large AI models at the network layer. LLMs and other large AI models have shown impressive capabilities in understanding medical text, summarizing multi-modal data and interacting with clinicians [25,26,28,29]. Yet, existing work almost exclusively positions them at the application layer; they are not systematically integrated into network control loops to provide semantic, patient-aware context for slice orchestration. The potential of LAMs to serve as semantic front ends for DTs and RL agents in IoE healthcare networks remains largely untapped.

Motivated by these gaps, this paper proposes a unified architecture in which (i) digital twins maintain predictive models of patients and medical devices, (ii) a large AI model (LAM) extracts high-level semantic features and risk indicators from heterogeneous clinical and operational data, and (iii) an RL-based network-slicing orchestrator uses this enriched state to proactively allocate resources across critical, monitoring, and routine slices. By tightly coupling DTs, LAMs and RL within a 6G-ready healthcare IoE framework, we aim to move from reactive, traffic-only control towards semantics- and patient-aware

resource orchestration that better aligns communication resources with clinical priorities and SLA requirements.

3. Proposed Framework

In this section, we present the proposed LAM-enabled, DT-driven network slicing architecture for healthcare IoE systems. As illustrated in Figure 1, the overall framework consists of three logical layers and a side LAM engine. The physical layer aggregates patients, wearable sensors, imaging devices, surgical robots and hospital infrastructure, which continuously generate heterogeneous multi-modal data such as ECG waveforms, medical images, logs and haptic signals. On top of this, the DT layer maintains virtual replicas of patients and medical devices, fuses the incoming measurements, and produces short-term predictions on emergencies, workload spikes and device failures. The network slicing layer hosts multiple logical slices that are tailored to heterogeneous healthcare services. In this work, we focus on three canonical slice categories, i.e., critical, monitoring, and routine slice, which are already depicted in Figure 1 as running on top of the DT layer. Their detailed definitions, typical applications, and QoS requirements will be elaborated in Section 3.3 and summarized in Table 1.

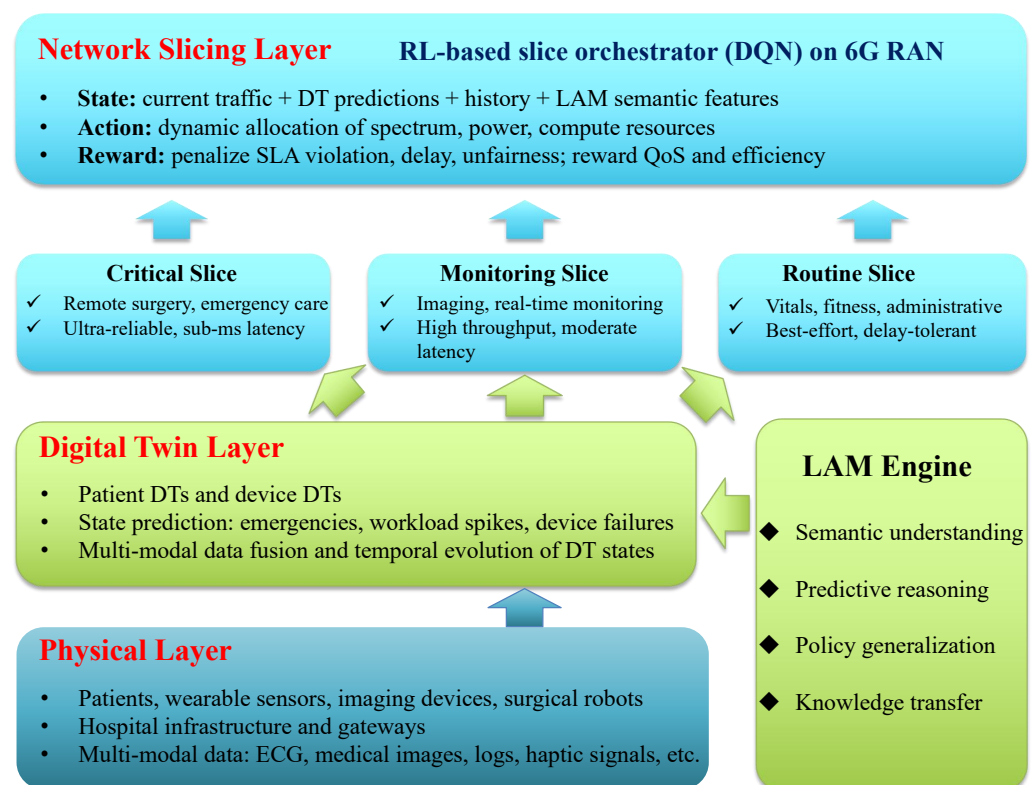


Figure 1. Illustration of the proposed framework.

More specifically, a dedicated LAM engine is employed to provide semantic understanding of clinical notes and operational logs, as well as predictive reasoning, policy generalization, and knowledge transfer. Its outputs are injected into the state representation used by the RL agent. In this work, the LAM engine is conceptualized as a multi-modal module, similar in spirit to GPT-4V or Gemini-like architectures, which can jointly process clinical text, structured physiological signals, and imaging-derived descriptors. In practice, such a model can be instantiated as a transformer-based multi-modal LLM that maps these heterogeneous inputs into a compact set of quantitative semantic factors (for example,

risk scores and urgency indicators). These semantic factors are then made available to the slicing controller to augment the network state.

Table 1. Representative healthcare IoE applications and slice-level QoS requirements.

Slice Type	Latency (ms)	Reliability (%)	Data Rate (Mb/s)	Traffic Pattern
Critical	$\leq 1\text{--}2$	≥ 99.999	10–100	Short packets, periodic with bursty peaks
Monitoring	10–50	≥ 99.9	10–500	Bursty, session-based streams
Routine	$\leq 100\text{--}500$	≥ 99	≤ 10	Low-rate, periodic or delay-tolerant

3.1. Role of Large AI Models

The novelty of the framework lies in its integration of LAMs into the orchestration process. LAMs provide cognitive intelligence beyond the capabilities of DTs and RL alone. Their contributions can be summarized as follows:

- **Semantic Understanding:** LAMs process heterogeneous healthcare IoE data, including clinical notes, medical images, sensor streams, and contextual hospital information. This allows the slicing manager to reason at a semantic level rather than only numerical indicators.
- **Predictive Reasoning:** LAMs enhance DT forecasts by recognizing cross-modal patterns, such as correlating abnormal ECG signals with textual physician reports to anticipate emergencies more accurately.
- **Policy Generalization:** When coupled with RL, LAMs provide foresight into state transitions, enabling policies that generalize across patients, devices, and varying traffic conditions. This reduces the time required for RL convergence and improves responsiveness.
- **Adaptive Knowledge Transfer:** LAMs can transfer insights learned from global datasets into local hospital environments, ensuring that orchestration policies benefit from broader medical and communication knowledge.

3.2. Reinforcement Learning for Slice Orchestration

To enable closed-loop and proactive slice orchestration, we cast the problem into a Markov decision process (MDP), as illustrated in Figure 2. At each decision epoch (t), the state s_t aggregates four types of information: (i) instantaneous traffic measurements across different slices and base stations, (ii) short-term workload and risk predictions provided by the patient/device DTs, (iii) historical key performance indicators (KPIs) such as latency and SLA violation statistics, and (iv) high-level semantic features extracted by the LAM from clinical notes, patient profiles and operational logs. Given s_t , a DQN-based RL agent computes the action a_t , which corresponds to a joint resource allocation decision over all active slices, including spectrum, power, edge computing capacity and admission control. The environment (i.e., the underlying 6G RAN and core network with dynamic traffic and queueing) responds to a_t , yielding the next state s_{t+1} and a scalar reward r_t that encodes SLA satisfaction, spectral efficiency and fairness among slices. The reward function r_t is given by

$$r_t = w_{\text{SLA}} r_{\text{SLA}}(t) + w_{\text{SE}} r_{\text{SE}}(t) + w_{\text{fair}} r_{\text{fair}}(t), \quad (1)$$

where $r_{\text{SLA}}(t)$ captures slice-level SLA satisfaction, r_{SE} captures the normalized spectral efficiency, and r_{fair} rewards fairness among slices. In particular, $r_{\text{SLA}}(t)$ is set to +1 if, within the current time window, all SLA constraints of the critical care slice are satisfied and to −1 otherwise, reflecting the higher priority of safety-critical services. The term r_{SE} is obtained by normalizing the aggregate spectral efficiency across slices to $[0, 1]$, while r_{fair} is derived from the Jain’s fairness index on the achieved throughput. By iteratively interacting with this MDP, the agent learns a slicing policy that leverages both DT predictions and LAM-enhanced semantic information to anticipate future demand and minimize long-term SLA violations.

Concretely, at each decision epoch t , the LAM engine outputs a low-dimensional semantic feature vector $\mathbf{f}_{\text{LAM}}(t) \in \mathbb{R}^{d_{\text{LAM}}}$. This vector contains interpretable scores such as: (i) a clinical risk score in $[0, 1]$ that summarizes the likelihood of acute deterioration, (ii) an urgency level for the current tasks, (iii) a resource criticality index that reflects the sensitivity to bandwidth and latency, and (iv) a compressed representation of associated free-text notes. The overall state fed to the RL agent is then formed by concatenating traffic measurements, DT predictions, historical KPIs, and $\mathbf{f}_{\text{LAM}}(t)$. In other words, all unstructured or multimodal inputs (e.g., free-text clinical notes and imaging-derived descriptors) are first processed by the LAM into the fixed-size semantic feature vector $\mathbf{f}_{\text{LAM}}(t)$, so that the DQN always operates on a well-defined, low-dimensional state representation.

Additionally, from a timing perspective, we adopt a multi-timescale design: LAM inference is executed at a coarse timescale (e.g., every few seconds or minutes) to update patient risk profiles and semantic scores, whereas the RL-based slicing decisions are made at the granularity of the 6G transmission time interval (TTI), typically in the sub-millisecond to millisecond range. The RL agent therefore operates on the most recently cached LAM outputs, and LAM inference is not required on a per-TTI basis. This separation decouples the potentially higher latency of LAM inference from the tight real-time constraints of TTI-level scheduling.

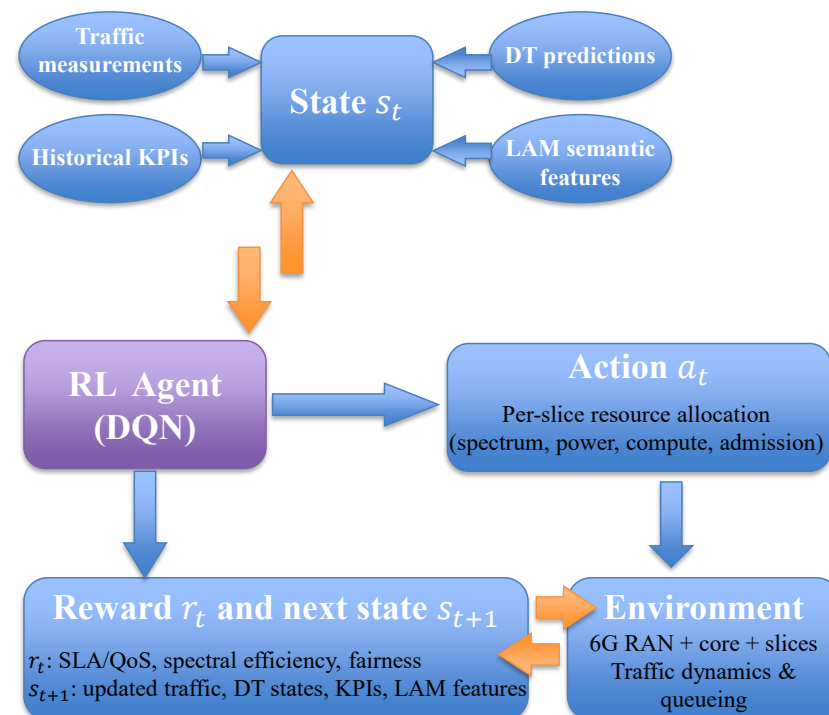


Figure 2. Illustration of the Markov decision process.

3.3. Slice Categories

We define three representative slice categories tailored to healthcare IoE scenarios. The goal is not to enumerate all possible services, but to capture the dominant classes of QoS requirements that drive the design of our DT+RL+LAM-based orchestration framework. These slices serve as logical abstractions for grouping traffic with similar latency, reliability, and throughput demands, and they are used consistently in the traffic models and simulations in Section 3.3.

- **Critical slice:** This slice is designed for life-critical, closed-loop services such as tele-surgery, emergency response, and haptic feedback control of surgical robots. The traffic is typically composed of short packets and control messages that must be delivered with extremely low latency and jitter to avoid destabilizing the control loop. We target sub-millisecond end-to-end latency and ultra-high reliability (on the order of $>99.999\%$), with moderate data rates sufficient to carry command, haptic, and status information. In practice, the critical slice receives the highest scheduling priority and is tightly coupled with DT predictions of imminent emergencies as well as LAM-derived semantic cues from clinical notes and triage records, so that additional resources can be proactively reserved when a high-risk situation is anticipated.
- **Monitoring slice:** The monitoring slice supports continuous observation and diagnostic services, including real-time medical imaging streams, ICU video monitoring, remote consultations, and high-resolution vital sign telemetry. Compared to the critical slice, these applications tolerate slightly higher latency (on the order of 10–50 ms) but often require much higher throughput to sustain video and image data streams. Reliability is still stringent (e.g., $\geq 99.9\%$) to avoid interruptions in continuous monitoring. The associated traffic is usually session-based and bursty, reflecting the start and end of imaging procedures or consultations. DT models provide forecasts of workload spikes (e.g., scheduled imaging sessions or ICU admissions), while the RL agent and LAM engine jointly adjust the resource allocation of the monitoring slice to prevent congestion and maintain smooth service quality.
- **Routine slice:** The routine slice aggregates non-critical and delay-tolerant services such as wearable fitness tracking, periodic wellness monitoring, electronic health record synchronization, and administrative or logistics traffic. These applications generate low- to moderate-rate data flows and can tolerate much larger end-to-end latency (on the order of 100–500 ms) and slightly relaxed reliability targets (e.g., around 99%), as long as the information is eventually delivered. From an orchestration perspective, the routine slice is treated as best-effort traffic that fills the residual capacity left after serving critical and monitoring slices; nevertheless, fairness constraints in the reward design ensure that it is not completely starved under heavy load. DT and LAM modules still contribute by identifying long-term usage patterns and background tasks that can be scheduled to off-peak periods.

These three slice categories are aligned with the heterogeneous QoS profiles summarized in Table 1 and form the basis on which the subsequent MDP formulation, traffic models, and performance evaluation are built.

3.4. Inter-Layer Coordination

The strength of the proposed framework lies in the interlayer synergy. Physical data streams feed DTs, which predict patient and device states. LAMs interpret and enrich these predictions, offering semantic insights and cross-modal reasoning. RL agents then leverage this contextual knowledge to orchestrate slices adaptively. The slicing decisions influence subsequent physical operations, creating a continuous feedback loop. This closed-loop

integration ensures that healthcare systems are not only adaptive but also anticipatory, aligning resources with patient-specific and system-wide requirements in real time.

3.5. Illustrative LAM–DT–RL Interaction Example

To illustrate how the LAM concretely alters the RL agent’s input state, consider a patient whose digital twin predicts abnormal ECG dynamics indicating a high risk of a cardiac episode. The DT summarizes these numerical signals into a set of predictive features that suggest an elevated probability of arrhythmia. In parallel, the physician’s clinical note contains phrases such as ‘sudden chest pain’ and ‘history of myocardial infarction’. The LAM jointly processes the DT-derived features and the clinical note. It outputs a high emergency risk score (e.g., close to 1), a ‘critical care’ urgency label, and a resource-criticality index indicating strong sensitivity to latency and reliability. These scores form the semantic feature vector $\mathbf{f}_{\text{LAM}}(t)$. The RL state at decision epoch t is then augmented as

$$s_t = [\text{traffic measurements, DT predictions, historical KPIs, } \mathbf{f}_{\text{LAM}}(t)]. \quad (2)$$

As a result, the learned policy is biased towards assigning this patient’s flows to the ultra-reliable, low-latency slice, even under congestion, whereas a DT-only or RL-only policy that does not see the clinical context may under-prioritize this session. This example demonstrates how the LAM provides complementary cognitive intelligence beyond the numerical metrics captured by the DT and the network alone.

3.6. Summary

In summary, the proposed LAM-enhanced DT-driven slicing framework introduces a holistic and future-ready design for healthcare IoE. By combining real-time mirroring through DTs, adaptive optimization via RL, and semantic reasoning through LAMs, the framework overcomes the limitations of static slicing and purely reactive systems. This integration ensures resilience against workload fluctuations, personalization of communication resources, and efficiency in large-scale healthcare IoE deployments, paving the way for precision medicine and intelligent healthcare services in the 6G era.

4. Simulation

In this section, we present a numerical analysis of the proposed framework based on Monte Carlo simulations. The evaluation focuses on how the joint use of DT predictions and LAM-derived semantic features affects slicing decisions under heterogeneous healthcare traffic. It is important to note that we do not emulate a fully deployed system with an online large AI model in the loop. The evaluation focuses on system reliability, latency performance, spectral efficiency, and fairness across heterogeneous slices. Comparative experiments against baseline strategies are also provided to demonstrate the advantages of DTs, RL, and LAMs.

4.1. Simulation Setup

The simulation environment emulates a hospital-scale deployment operating within a 6G IoE network. Specifically, the network supports 200 patients equipped with wearable devices, continuous monitoring sensors, and medical imaging equipment, alongside mission-critical applications such as telesurgery and emergency response systems. Each patient is associated with a DT, which continuously synchronizes with real-world data streams including ECG signals, imaging data, and hospital workflow logs. These ECG and wearable data are represented by synthetic summary indicators rather than full raw signals. For each patient, we generate a small set of synthetic high-level indicators that summarize

the condition of different data sources, such as an ECG abnormality score, a wearable activity index, and a binary flag indicating whether an imaging-intensive examination is involved. These indicators are statistically correlated with the traffic class (critical care, monitoring, routine) and the underlying patient state: emergency sessions tend to exhibit higher ECG abnormality scores and critical imaging flags, whereas routine sessions exhibit low-risk values. The DT module consumes these indicators, together with recent network statistics, to produce short-term risk and load predictions, while the LAM uses them to derive semantic scores that augment the RL agent's state. Regarding imaging data, we adopt an abstract representation instead of simulating concrete medical images or vision models. Each session is assigned a binary 'imaging-intensive' flag indicating whether it involves resource-demanding imaging procedures (e.g., CT or MRI). This flag is treated as an additional input feature to the DT and LAM modules, influencing the predicted risk and resource-criticality scores. The slicing manager is implemented as a centralized RL agent, trained via a DQN with experience replay and target networks to ensure stable convergence. The total network bandwidth is set to 2 GHz and divided into dynamic resource blocks assigned to different slices. Simulation time is divided into epochs, each representing one second of real-world operation, during which traffic demands fluctuate based on stochastic models of patient conditions and hospital workloads.

In our simulation study, the LAM engine is instantiated at an abstract level. Instead of deploying a full multi-modal large AI model, we emulate its effect by generating a set of synthetic semantic features that are correlated with the simulated traffic events and DT predictions. Specifically, for each session we assign a clinical risk category and task urgency label based on the traffic class (critical care, monitoring, routine) and the DT-predicted condition, and then map these labels to numerical scores in $[0, 1]$ with small random perturbations. The resulting semantic feature vector $\mathbf{f}_{\text{LAM}}(t)$ is concatenated with the conventional network statistics to form the RL state. In practice, we implement a simplified mock LAM pipeline directly at the feature level. For each session, we derive semantic scores from the traffic class and the DT-predicted condition using rule-based mappings. For example, emergency critical-care flows whose DT state indicates elevated risk are assigned a high clinical risk score (e.g., $f_{\text{LAM},\text{risk}} \in [0.8, 1.0]$) and a high urgency score, whereas routine or stable monitoring flows are assigned lower values (e.g., $f_{\text{LAM},\text{risk}} \in [0.0, 0.3]$). As for the weights in the reward function r_t , we use $(w_{\text{SLA}}, w_{\text{SE}}, w_{\text{fair}}) = (0.5, 0.3, 0.2)$, which balances the emphasis on SLA satisfaction with spectral efficiency and fairness after preliminary tuning.

The proposed DT+RL+LAM slicing framework was implemented in Python using the PyTorch deep learning library. The DQN-based slice orchestrator adopts a fully connected neural network with three hidden layers of 128, 64, and 32 neurons, respectively, and ReLU activation. The discount factor is set to $\gamma = 0.99$, the learning rate to 10^{-3} , and an ϵ -greedy exploration strategy is used, where ϵ is linearly annealed from 1.0 to 0.05 over the first 10^4 decision steps and then kept constant. The replay buffer stores 5×10^4 transitions, and the target network is updated every 100 training steps to stabilize learning. These hyperparameters are chosen to balance modeling capacity and training stability. In particular, the three fully connected hidden layers with 128, 64, and 32 neurons provide sufficient expressiveness to capture the nonlinear relationship between the augmented state and the optimal slicing action, while larger networks yielded only marginal performance gains at the cost of increased training time in preliminary experiments [30]. Additionally, each simulation run consists of 10,000 epochs (seconds), and results are averaged over 20 Monte Carlo runs with different random seeds to smooth out stochastic variations. All experiments are executed on a workstation equipped with a 16-core Intel Xeon CPU at 3.0 GHz, 64 GB of RAM, and a single NVIDIA RTX-3090 GPU. Although a GPU is used

to accelerate training, the proposed algorithm can also be executed on a CPU-only server with unchanged algorithmic behavior.

4.2. Traffic Model

The traffic model is designed to capture the heterogeneity of healthcare IoE environments. Three major categories are defined in alignment with the slice design:

- **Critical Traffic:** This includes tele-surgery haptic control signals and emergency alerts. These streams have strict requirements of sub-millisecond latency and packet loss probability below 10^{-5} .
- **Monitoring Traffic:** This covers high-bandwidth applications such as MRI scans, real-time ultrasound video, and continuous ECG streams. These flows require stable throughput with moderate latency tolerance, typically below 100 ms.
- **Routine Traffic:** This consists of fitness tracking data, hospital administrative logs, and wellness monitoring from large populations of wearable devices. These flows are delay-tolerant but highly scalable, reflecting the mMTC dimension of healthcare IoE.

Traffic arrivals are modeled using Poisson processes with variable intensities to mimic realistic fluctuations in patient workloads. Bursty events, such as emergencies or scheduled imaging procedures, are introduced to test the robustness of the proposed orchestration framework. It is worth noting that, in this simulation study, we do not explicitly generate multi-modal clinical notes or medical images; instead, the effect of such inputs is abstracted through synthetic LAM-derived semantic features that are correlated with the traffic classes and DT predictions.

4.3. Performance Metrics

System performance is evaluated using four key metrics, which are defined more formally as follows.

- (1) **SLA Violation Ratio:** Each task i is associated with an end-to-end latency L_i and a packet loss probability p_i , and belongs to one of the three slices $s \in \text{critical, monitoring, routine}$. For each slice we specify latency and reliability thresholds (L_{th}^s, p_{th}^s) reflecting the QoS requirements in Section 4.2, e.g., $L_{th}^{\text{critical}} = 1 \text{ ms}$, $p_{th}^{\text{critical}} = 10^{-5}$, $L_{th}^{\text{monitoring}} = 100 \text{ ms}$, $p_{th}^{\text{monitoring}} = 10^{-3}$, and $L_{th}^{\text{routine}} = 500 \text{ ms}$, $p_{th}^{\text{routine}} = 10^{-2}$. A task is counted as violating the SLA if either the latency or reliability constraint is not met:

$$v_i = \begin{cases} 1, & L_i > L_{th}^{s(i)} \text{ or } p_i > p_{th}^{s(i)}, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where $s(i)$ denotes the slice of task i . If N is the total number of tasks in the simulation, the SLA violation ratio is

$$\text{SLA-Violation} = \frac{1}{N} \sum_{i=1}^N v_i \times 100\%. \quad (4)$$

- (2) **Spectral Efficiency:** Let B_{tot} denote the total system bandwidth and T_{sim} the simulation duration. Let R_i (in bits) be the successfully delivered payload of task i that meets its SLA. The achieved spectral efficiency is computed as

$$\text{SE} = \frac{\sum_{i=1}^N R_i}{B_{\text{tot}}, T_{\text{sim}}} \quad [\text{bits/s/Hz}], \quad (5)$$

which reflects how efficiently the available spectrum is utilized under QoS constraints.

- (3) **Latency Distribution:** For each slice we record the end-to-end latency L_i of all tasks and then compute the empirical cumulative distribution function (CDF)

$F_L^s(\ell) = \Pr(L_i \leq \ell)$. These CDFs provide a detailed view of how often ultra-low latency requirements are met, beyond the average values reported in the tables and heatmaps.

- (4) **Fairness Index:** Fairness across slices is quantified using Jain's index. Let x_s denote the average satisfied throughput (or equivalently, the number of SLA-satisfied tasks) for slice s over the simulation horizon. For $S = 3$ slices, Jain's fairness index is given by

$$\mathcal{F} = \frac{\left(\sum_{s=1}^S x_s\right)^2}{S \sum_{s=1}^S x_s^2}, \quad 0 < \mathcal{F} \leq 1, \quad (6)$$

where higher values indicate a more balanced allocation across slices while still allowing critical traffic to be prioritized.

In addition, we report the Accuracy of the slicing policy, which we define as the task-level QoS satisfaction ratio. Specifically, Accuracy is given by the percentage of user sessions for which all end-to-end QoS constraints (latency, reliability, and bandwidth) are simultaneously satisfied, over the total number of sessions generated during the simulation. This provides a complementary, session-centric view of QoS performance to the time-window-based SLA-violation metric.

4.4. Baseline

To highlight the benefits of the proposed framework, we compared it against three baselines:

- **Static Slicing:** Resources are pre-allocated to each slice, independent of traffic fluctuations.
- **Rule-Based Dynamic Slicing:** Threshold-based adjustments are made when traffic exceeds predefined limits.
- **RL-Based Slicing without DT/LAMs:** Reinforcement learning is applied for resource allocation but without predictive insights from DTs or semantic reasoning from LAMs.

4.5. Results and Analysis

Simulation results are provided to demonstrate that the effectiveness and efficiency of our proposed framework. Specifically, Figure 3 compares the SLA violation probability of four slicing strategies. Static slicing exhibits the highest SLA violation at about 18.6%, while introducing simple rule-based adaptation reduces the violation rate to 12.3%. A pure RL-driven orchestrator further decreases the violation to 9.8% by reacting to traffic dynamics. In contrast, the proposed DT+RL+LAM framework achieves the lowest SLA violation of only 5.6%, corresponding to a reduction of nearly 70% compared with static slicing and about 43% compared with the RL-only baseline. This confirms that coupling DT predictions and LAM-extracted semantic features with RL enables more proactive and fine-grained resource orchestration.

Table 2 further compares the four schemes in terms of task accuracy, latency, spectral efficiency, and fairness. Static slicing yields the worst performance (accuracy 81.2%, latency 22.5 ms, spectral efficiency 3.1 bits/s/Hz, fairness index 0.76), and rule-based control provides moderate improvements. RL-only orchestration boosts accuracy to 89.3%, lowers latency to 15.7 ms, and improves spectral efficiency and fairness (4.2 bits/s/Hz, 0.84). The proposed DT+RL+LAM solution consistently outperforms all baselines, achieving the highest accuracy (94.1%), the lowest latency (11.3 ms), the best spectral efficiency (5.1 bits/s/Hz), and the most balanced resource allocation (fairness index 0.91). Overall, these results demonstrate that integrating DT-driven state prediction and LAM-enabled semantic understanding into the RL orchestrator not only suppresses SLA violations

but also jointly enhances reliability, timeliness, efficiency, and fairness of 6G healthcare network slicing.

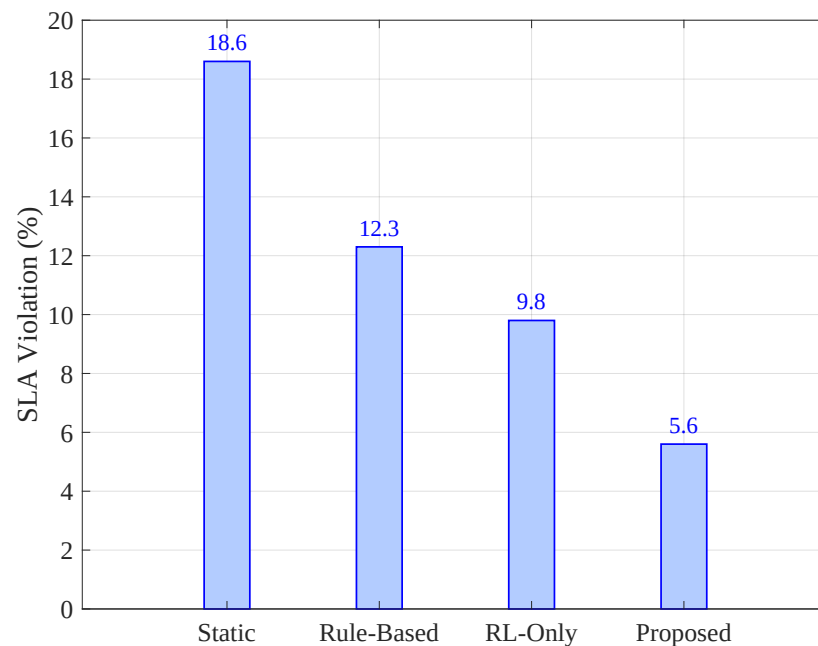


Figure 3. Comparison of SLA violation ratio across different slicing methods, in which the proposed DT+RL+LAM framework achieves the lowest violation rate.

Table 2. Performance Comparison of Different Slicing Strategies.

Method	QoS Satisfaction (%)	Latency (ms)	Spectral Efficiency (bits/s/Hz)	Fairness Index
Static Slicing	81.2%	22.5	3.1	0.76
Rule-Based	85.6%	18.9	3.8	0.81
RL-Only	89.3%	15.7	4.2	0.84
Proposed (DT+RL+LAM)	94.1%	11.3	5.1	0.91

Figure 4 presents a heatmap of the average end-to-end latency for three slice types (critical, monitoring and routine) under the four slicing strategies. For the static configuration, the critical slice already experiences around 10 ms latency, while monitoring and routine services suffer from delays above 20 ms and 30 ms, respectively. Rule-based control reduces latency moderately for all slices but still cannot guarantee ultra-low delay for critical tasks. The RL-only orchestrator yields further improvements, particularly for the critical slice, whose latency drops to about 4 ms. The proposed DT+RL+LAM framework achieves the best overall performance, bringing the critical-slice latency close to 1–2 ms and significantly reducing the delay of monitoring and routine slices at the same time. This per-slice view confirms that the proposed architecture not only prioritizes life-critical services, but also avoids starving non-critical traffic, leading to a more balanced and clinically meaningful QoS provisioning. Notably, the heatmap visually confirms that our framework can simultaneously maintain near-theoretical latency for the critical slice (1–2 ms) while still delivering clear latency improvements for the monitoring and routine slices compared with all baselines, highlighting a balanced multi-slice performance.

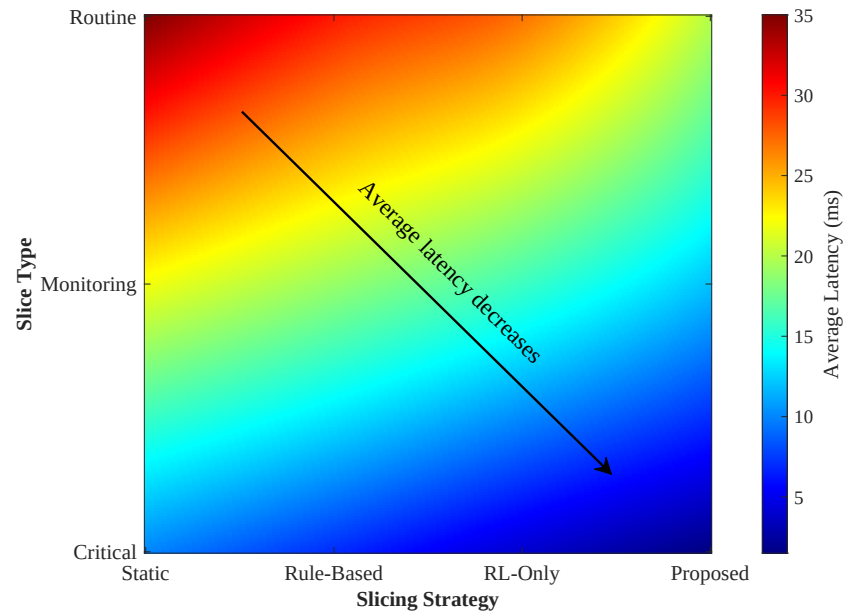


Figure 4. Heatmap of the average end-to-end latency for three slice types.

4.6. Discussion

These findings confirm the synergistic role of DTs, RL, and LAMs in healthcare IoE. DTs provide predictive foresight, RL ensures adaptive policy learning, and LAMs add semantic reasoning that bridges multi-modal data and contextual knowledge. Together, these components create a proactive–reactive orchestration loop that cannot be achieved by static or reactive methods alone. While integrating LAMs introduces computational overhead, the simulation results indicate that the added processing cost is negligible compared with the performance improvements. This is because in our simulation the LAM is instantiated only as a lightweight feature-generation block that maps traffic labels and DT predictions to synthetic semantic scores, rather than as a full large language model. Overall, the proposed framework demonstrates technical feasibility and significant advantages for large-scale deployment in smart healthcare IoE systems.

5. Discussion and Future Challenges

The proposed LAM-enhanced DT-driven slicing framework demonstrates clear advantages in healthcare IoE systems by combining predictive foresight, adaptive optimization, and semantic reasoning. Compared with static or rule-based approaches, it achieves superior SLA compliance, latency reduction, and fairness, proving its potential to support mission-critical applications such as telesurgery and real-time diagnostics. The layered design also enables personalization, where patient-specific DTs directly influence slice orchestration, aligning communication resources with individualized health conditions.

Despite these benefits, several challenges remain before large-scale deployment is feasible. First, scalability is a critical issue: maintaining thousands of DTs synchronized in real time generates significant computation and communication overhead, requiring distributed or federated RL mechanisms to avoid bottlenecks. Hierarchical orchestration strategies or hybrid cloud–edge architectures may also be necessary to balance computation between centralized servers and hospital edge nodes.

Secondly, data privacy and security remain major concerns. DTs and LAMs inevitably process highly sensitive medical information, including patient records and imaging data. Ensuring compliance with healthcare regulations such as HIPAA or GDPR requires robust privacy-preserving mechanisms, including federated DT synchronization, homomorphic

encryption, and differential privacy. Furthermore, LAMs themselves may be vulnerable to adversarial attacks or membership inference, necessitating robust defense techniques and blockchain-based accountability frameworks.

Third, interoperability across heterogeneous devices and protocols poses a barrier to adoption. Current hospital environments are composed of legacy equipment, proprietary communication systems, and emerging IoT devices, often lacking unified standards. Achieving seamless integration of DT-driven orchestration will require standardized interfaces, lightweight communication protocols, and ontology-based knowledge representations to harmonize heterogeneous healthcare data.

Finally, the explainability of RL and LAMs is an open challenge. Both reinforcement learning and large AI models are often perceived as “black boxes,” which raises concerns for clinical practitioners and regulators who require transparent decision-making in safety-critical environments. Integrating explainable AI techniques into DTs and LAMs will be necessary to build trust, improve accountability, and facilitate clinical acceptance.

Future research should therefore explore interdisciplinary innovations, including privacy-preserving federated DTs, explainable RL for medical decision-making, and energy-aware orchestration protocols. Emerging technologies such as blockchain for decentralized accountability, semantic communication in 6G networks, and quantum-safe cryptography may further enhance the robustness of the proposed framework. Ultimately, collaboration across the fields of wireless communications, AI, and healthcare informatics will be essential to translate this conceptual framework into resilient and trustworthy 6G-enabled healthcare systems.

6. Conclusions

In this article, we proposed a LAM-enhanced DT-driven network slicing framework for smart healthcare in 6G-enabled Internet of Everything (IoE). By tightly integrating DTs for predictive insights, RL for adaptive orchestration, and LAMs for semantic reasoning, the system achieves proactive, patient-centric, and resilient resource management. Simulation results confirmed significant performance gains, including an approximately 42–43% reduction in SLA violations, a 31% improvement in spectral efficiency, and a 29% reduction in latency compared with baseline methods, thereby validating the feasibility of the proposed framework.

Beyond quantitative improvements, the framework highlights the importance of semantic intelligence in healthcare IoE. LAMs provide cross-modal reasoning that bridges textual, visual, and sensor data, while DTs enable real-time synchronization of patient and device states. Together, they empower RL agents to make proactive allocation decisions rather than purely reactive adjustments. This multi-layered synergy ensures reliability and personalization for diverse applications ranging from tele-surgery to routine health monitoring.

Looking forward, several challenges must be addressed to realize large-scale deployment, including scalability of DTs, privacy and security of sensitive data, interoperability across heterogeneous systems, and explainability of AI-driven orchestration. Future work should investigate lightweight LAMs tailored for edge deployment, federated DT synchronization mechanisms, and hybrid blockchain solutions for accountability. With these advancements, the proposed framework has the potential to become a cornerstone of intelligent, secure, and patient-centric healthcare services in the 6G era, paving the way toward resilient medical IoE infrastructures.

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

Funding: This work was supported in part by the Natural Science Foundation of Fujian Province, China (Grant No. 2025J01633).

Data Availability Statement: Data are contained within the article.

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

References

1. Wang, C.X.; You, X.; Gao, X.; Zhu, X.; Li, Z.; Zhang, C.; Wang, H.; Huang, Y.; Chen, Y.; Haas, H.; et al. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 905–974. [\[CrossRef\]](#)
2. Akan, O.B.; Dinc, E.; Kuscü, M.; Cetinkaya, O.; Bilgin, B.A. Internet of Everything (IoE) - From Molecules to the Universe. *IEEE Commun. Mag.* **2023**, *61*, 122–128. [\[CrossRef\]](#)
3. Kong, X.; Wu, Y.; Wang, H.; Xia, F. Edge Computing for Internet of Everything: A Survey. *IEEE Internet Things J.* **2022**, *9*, 23472–23485. [\[CrossRef\]](#)
4. Chen, Y.; Guo, X.; Zhou, G.; Jin, S.; Ng, D.W.K.; Wang, Z. Unified Far-Field and Near-Field in Holographic MIMO: A Wavenumber-Domain Perspective. *IEEE Commun. Mag.* **2025**, *63*, 30–36. [\[CrossRef\]](#)
5. Aqeel, I.; Khormi, I.M.; Khan, S.B.; Shuaib, M.; Almusharraf, A.; Alam, S.; Alkhalidi, N.A. Load Balancing Using Artificial Intelligence for Cloud-Enabled Internet of Everything in Healthcare Domain. *Sensors* **2023**, *23*, 5349. [\[CrossRef\]](#)
6. Foukas, X.; Patounas, G.; Elmokashfi, A.; Marina, M.K. Network Slicing in 5G: Survey and Challenges. *IEEE Commun. Mag.* **2017**, *55*, 94–100. [\[CrossRef\]](#)
7. Zhang, S. An Overview of Network Slicing for 5G. *IEEE Wirel. Commun.* **2019**, *26*, 111–117. [\[CrossRef\]](#)
8. Vergutz, A.; Noubir, G.; Nogueira, M. Reliability for Smart Healthcare: A Network Slicing Perspective. *IEEE Netw.* **2020**, *34*, 91–97. [\[CrossRef\]](#)
9. Chen, Y.; Wang, Y.; Liu, M.; Zhang, J.; Jiao, L. Network Slicing Enabled Resource Management for Service-Oriented Ultra-Reliable and Low-Latency Vehicular Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 7847–7862. [\[CrossRef\]](#)
10. Lu, Q.; Li, M. Fault Prediction Method Towards Rolling Element Bearing Based on Digital Twin and Deep Transfer Learning. *Appl. Sci.* **2025**, *15*, 12509. [\[CrossRef\]](#)
11. Addula, S.R.; Ramaswamy, Y.; Dawadi, D.; Khan, Z.; Veeramachaneni, P.; Pamidi Venkata, A.K. Blockchain-Enabled Healthcare Optimization: Enhancing Security and Decision-Making Using the Mother Optimization Algorithm. In Proceedings of the 2025 International Conference on Intelligent and Cloud Computing (ICoICC), Bhubaneswar, India, 2–3 May 2025; pp. 1–8.
12. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 39. [\[CrossRef\]](#)
13. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *Natl. Sci. Rev.* **2024**, *11*, nwae403. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 106. [\[CrossRef\]](#)
15. Qiu, J.; Li, L.; Sun, J.; Peng, J.; Shi, P.; Zhang, R.; Dong, Y.; Lam, K.; Lo, F.P.W.; Xiao, B.; et al. Large AI Models in Health Informatics: Applications, Challenges, and the Future. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 6074–6087. [\[CrossRef\]](#)
16. Dreyer, M.; Berend, J.; Labarta, T.; Vielhaben, J.; Wiegand, T.; Lapuschkin, S.; Samek, W. Mechanistic understanding and validation of large AI models with semanticons. *Nat. Mach. Intell.* **2025**, *7*, 1572–1585. [\[CrossRef\]](#)
17. Jain, H.; Chamola, V.; Jain, Y.; Naren. 5G network slice for digital real-time healthcare system powered by network data analytics. *Internet Things Cyber-Phys. Syst.* **2021**, *1*, 14–21. [\[CrossRef\]](#)
18. Berdufi, E.; Slamnik-Kriještorac, N.; Marquez-Barja, J. Leveraging on Network Slicing to Enable and Enhance IoT-based e-Health services. In Proceedings of the GoodIT '22: Proceedings of the 2022 ACM Conference on Information Technology for Social Good, Limassol, Cyprus, 7–9 September 2022; pp. 431–436. [\[CrossRef\]](#)
19. Alwakeel, A.M.; Alnaim, A.K. Network Slicing in 6G: A Strategic Framework for IoT in Smart Cities. *Sensors* **2024**, *24*, 4254. [\[CrossRef\]](#)
20. 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. [\[CrossRef\]](#)

21. Dias, J.; Pinto, P.; Santos, R.; Malta, S. 5G Network Slicing: Security Challenges, Attack Vectors, and Mitigation Approaches. *Sensors* **2025**, *25*, 3940. [[CrossRef](#)]
22. Kabir, M.R.; Shishir, F.S.; Shomaji, S.; Ray, S. Digital twins in healthcare IoT: A systematic review. *High-Confid. Comput.* **2025**, *5*, 100340. [[CrossRef](#)]
23. Ringeval, M.; Etindele Sosso, F.A.; Cousineau, M.; Paré, G. Advancing Health Care With Digital Twins: Meta-Review of Applications and Implementation Challenges. *J. Med. Internet Res.* **2025**, *27*, e69544. [[CrossRef](#)]
24. Hemdan, E.E.D.; Sayed, A. Smart and Secure Healthcare with Digital Twins: A Deep Dive into Blockchain, Federated Learning, and Future Innovations. *Algorithms* **2025**, *18*, 401. [[CrossRef](#)]
25. Zhou, J.; Li, H.; Chen, S.; Chen, Z.; Han, Z.; Gao, X. Large language models in biomedicine and healthcare. *npj Artif. Intell.* **2025**, *1*, 44. [[CrossRef](#)]
26. Maity, S.; Saikia, M.J. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **2025**, *12*, 631. [[CrossRef](#)]
27. Imrie, F.; Rauba, P.; van der Schaar, M. Redefining digital health interfaces with large language models. *arXiv* **2023**, arXiv:2310.03560. [[CrossRef](#)] [[PubMed](#)]
28. Zonayed, M.; Tasnim, R.; Jhara, S.S.; Mimona, M.A.; Hussein, M.R.; Mobarak, M.H.; Salma, U. Machine learning and IoT in healthcare: Recent advancements, challenges & future direction. *Adv. Biomark. Sci. Technol.* **2025**, *7*, 335–364. [[CrossRef](#)]
29. Freyer, O.; Wiest, I.C.; Kather, J.N.; Gilbert, S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit. Health* **2024**, *6*, e662–e672. [[CrossRef](#)]
30. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)] [[PubMed](#)]

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.