Dual-LLM Integration with Reconfigurable Intelligent Surface for Healthcare Networks

Sravani Kurma, *Student Member, IEEE*, Keshav Singh, *Member, IEEE*, Anal Paul *Member, IEEE*, Shahid Mumtaz, *Senior Member, IEEE*, and Chih-Peng Li, *Fellow, IEEE*

***Abstract*—The increasing complexity of real-time healthcare necessitates intelligent systems for dynamic data management and personalized assistance. This paper proposes a novel dual-LLM framework that integrates large language models (LLMs) into wireless healthcare networks. The first LLM powers an interactive artificial intelligence module (IAIM) embedded within a mobile edge computing (MEC) environment, which dynamically optimizes user-specific data routing and reconfigurable intelligent surface (RIS) configurations via a modified proximal policy optimization (PPO) algorithm. A novel Greedy Look-Ahead Algorithm (GLAA) is introduced for real-time path selection based on signal strength, emergency factors, and user-specific parameters. The second LLM, utilizing a retrieval-augmented generation (RAG) approach, serves as a personalized healthcare chat assistant that delivers context-aware patient support using real-time and historical data. Simulation results demonstrate that the proposed IAIM achieves a 9.6% reduction in network overhead compared to manual modeling and reduces latency by up to 52.5% over baseline PPO approaches, thus enabling enhanced user experience and responsiveness in healthcare systems.**

***Index Terms*—Deep reinforcement learning, Interactive artificial intelligence, Large language models, Mobile edge computing, Reconfigurable intelligent surfaces, Retrieval-augmented generation, Health management.**

I. Introduction

LARGE language models (LLMs) have emerged as transformative tools in wireless communication, particularly within 5G, the forthcoming 6G, and broader wireless technologies. These sophisticated models, such as GPT-4 and its successors, possess unparalleled capabilities to process and comprehend vast volumes of textual data [1]–[4]. The transition from 4G to 5G marked a vital moment, introducing unprecedented data speeds and connectivity. As we advance to 6G networks, intelligent and adaptive solutions are crucial to meet escalating demands while ensuring network resilience. LLMs, with their real-time decision-making capabilities, are well-positioned to address these challenges by ensuring optimal network performance under evolving conditions [5]–[7]. The advantages of integrating LLMs into wireless communications are manifold. They can anticipate and mitigate environmental impacts on network performance, predict network congestion, identify straggler nodes, and adapt to adverse weather conditions, enabling networks to operate at peak efficiency [8]. Moreover, LLMs excel in the dynamic allocation of network resources, optimizing connectivity in both high-demand and low-connectivity areas [9].

As LLMs continue to redefine artificial intelligence (AI) and natural language understanding, their potential extends far beyond current applications. The shift from rule-based inferences to advanced learning models necessitates more sophisticated AI solutions to manage the growing complexity of data. Enhancing LLMs with interactive AI (IAI) enables more dynamic and user-responsive networking solutions [10]. Combining IAI with LLMs allows systems to react to changes and proactively manage resources through direct interactions with users and real-time data feeds. Developing LLM-based chatbots for healthcare utilizes patient data and advanced techniques like retrieval-augmented generation (RAG) to provide personalized, context-aware interactions, revolutionizing patient support [11]. Applications of LLMs and IAI in wireless communications include network optimization, predictive maintenance, improved security, and personalized user experiences. Integrating IAI with technologies like RAG and LangChain enriches AI responses by extracting information from vast databases, offering tailored solutions that align with user needs [11]. This combination enhances flexibility, minimizes human bias, and optimizes network resource utilization, paving the way for significant advancements in AI-driven networking.

In wireless communications, routing mechanisms determine optimal paths for data transmission between source and destination nodes, ensuring timely information delivery. Traditional methods, such as proactive protocols like optimized link state routing (OLSR) and reactive protocols like ad-hoc on-demand distance vector (AODV), have been widely used in dynamic networks [12].Proactive protocols maintain up-to-date routing information but generate significant overhead due to constant control message exchanges. Reactive protocols reduce overhead by initiating route discovery only when needed, though this often results in delayed route establishment due

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC 112-2221-E-110-038-MY3, NSTC 113-2218-E-110-009 and NSTC 113-2222-E-110-008-MY3, and also in part by the Sixth Generation Communication and Sensing Research Center funded by the Higher Education SPROUT Project, the Ministry of Education of Taiwan. The work of S. Mumtaz was part of the 6G-SENSES project from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101139282. (*Corresponding author:Keshav Singh*).

S. Kurma, K. Singh, and C-P. Li are with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (Email: sravani.phd.nsysu.21@gmail.com, keshav.singh@mail.nsysu.edu.tw, cpli@faculty.nsysu.edu.tw).

A. Paul is with the Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320315, Taiwan (Email: apaul@saturn.yzu.edu.tw)

S. Mumtaz is with the Department of Applied Informatics, Silesian University of Technology Akademicka Gliwice, Poland (Email: dr.shahid.mumtaz@ieee.org).

to on-demand setup. Hybrid routing mechanisms combine the strengths of both approaches, offering the persistent connectivity of proactive methods while mitigating overhead through reactive features. However, their complexity poses implementation and maintenance challenges. This adaptability is crucial in modern wireless systems, including internet of things (IoT) devices and reconfigurable intelligent surfaces (RIS), where network dynamics and performance requirements vary significantly.

A. Literature review

In this paper, we explore both LLM and wireless communication-based approaches.

Under LLM based approach, we comment on existing literature focusing on the broader aspects of the current topic. In the area of chatbots, the AI-powered medical chatbot [13] aims to improve healthcare access and reduce costs by offering preliminary disease diagnosis and information using techniques like n-gram, term frequency-inverse document frequency (TFIDF), and cosine similarity. Complex queries are handled by a third-party expert system, enhancing initial medical consultations. The "CataractBot" study [14] developed with an eye hospital in India, uses AI and expert verification to provide accurate, multilingual information about cataract surgery, enhancing the trustworthiness and accessibility of health information. A trial with 49 participants demonstrated its effectiveness in making health information accessible and trustworthy, thereby easing the burden on healthcare professionals. Comparatively, both CataractBot and the general AI medical chatbot [13] enhance access to healthcare information, with CataractBot offering expert-verified cataract surgery information, focusing on building trust. The study [15] improves interactions with RAG-based agents by developing a suggestion question generator using dynamic contexts, such as few-shot examples and retrieved information. Experiments show this approach generates better questions, helping users communicate more effectively with the system. The REALM framework [16] enhances clinical predictions by integrating multimodal electronic health records (EHR) data with external knowledge graphs, using an LLM to process clinical notes and a gated recurrent unit (GRU) model for time-series data, ensuring consistency and reducing errors. Tested on MIMIC-III mortality and readmission tasks, REALM significantly outperforms traditional models, demonstrating its effectiveness in refining clinical insights. Each approach tackles different aspects of healthcare digitization with AI. Studies like [17] and [14] focus on user acceptability and specific informational needs, while technologies in [13] and [15] aim to enhance the efficiency of medical consultations and user interactions with AI. The REALM framework [16] illustrates the advanced application of AI in processing and integrating complex healthcare data for better clinical outcomes. Furthermore, Park *et al.* [18] provided a timely analysis of generative AI's potential in automating knowledge-intensive tasks, particularly within the healthcare domain. Their study positions healthcare as a key sector poised for LLM deployment, reinforcing the rationale behind our dual-LLM integration in smart healthcare systems. In a more targeted clinical application, Hu *et al.* [19] developed a GPT-4-powered RAG-enhanced system tailored for dementia care. Achieving a diagnostic accuracy of 90%, their model not only offers high clinical readability but also generates personalized care plans reviewed by medical professionals. This closely aligns with our proposed dual-LLM architecture, which emphasizes context-awareness and personalized medical support through a collaborative model involving both cloud-based and edge-side LLMs. Collectively, these works represent a multifaceted advancement in healthcare digitization. While studies such as [17] and [14] prioritize patient acceptability and specialized health education, others like [13], [15], and [16] focus on enhancing clinical effectiveness through AI-powered interaction, offloading, and decision support.

To address the increasing complexity of real-time healthcare and the growing demand for advanced IoT applications, the integration of cutting-edge technologies such as mobile edge computing (MEC) within 5G networks is essential [20]. MEC, a key enabler of IoT, offers cloud computing capabilities at the network's edge, thereby enabling faster response times and enhancing computational efficiency for end users [21], [22]. Under the wireless communication domain, we focus on mobile edge computing (MEC) systems, and the relevant literature survey is as follows: Computation offloading and resource allocation are of paramount importance in MEC networks, garnering significant attention in recent years [23], [24]. Performance evaluation often considers energy consumption [25], [26], [27] and latency [28], [29] as key criteria. Munoz *et al.* [27] minimized energy consumption by optimizing transmission time and offloaded data to a femto access point (AP). A low-complexity Lyapunov optimization-based dynamic computation offloading algorithm was proposed in [28] to reduce execution time. Yang *et al.* [29] designed a heuristic method to partition users' computation tasks to minimize average completion time. Ni *et al.* [30] proposed a resource allocation strategy using priced timed Petri nets, considering cost and credibility evaluations of users and fog resources. Wang *et al.* [31] addressed energy consumption and execution latency minimization by optimizing computation speed and transmission power. Several studies have explored the intricate trade-off between energy consumption and execution latency in mobile edge computing (MEC) systems. For instance, Hong *et al.* [32] modeled data offloading scheduling as a dynamic programming problem, introducing a weighting factor to balance the combined impact of energy usage and latency in their optimization framework. Recent advances in artificial intelligence (AI) and reconfigurable intelligent surfaces (RISs) have opened up transformative possibilities for MEC networks, enabling smarter, more adaptive resource allocation and performance enhancement. Notably, Ni *et al.* [33] conducted a comprehensive analytical modeling and simulation-based outage performance analysis of RIS-assisted device-to-device (D2D) communications tailored for healthcare IoT environments, highlighting the potential of RIS in improving link reliability and spectral efficiency. Similarly, Mercuri *et al.* [34] investigated the integration of RIS into ambient assisted living (AAL) and smart hospital infrastructures, emphasizing its effectiveness in enhancing radar-based indoor human monitor-

ing. Furthermore, AI-driven techniques are increasingly being leveraged to optimize computation offloading strategies and resource management in MEC systems, particularly in latency-sensitive and energy-constrained scenarios, thereby advancing the operational intelligence and responsiveness of next-generation healthcare networks. For example, AI algorithms can predict network congestion, dynamically adjust resource allocation, and optimize energy consumption in real-time [35]. RIS technology leverages AI to intelligently manipulate electromagnetic waves, improving signal strength and coverage. This enhances the efficiency of MEC systems by dynamically adjusting phase shifts to optimize wireless communication [36]. The integration of RIS with MEC systems allows for better resource utilization and improved energy efficiency. Distinct from previous studies, this paper proposes an energy-aware offloading scheme that balances energy consumption and execution latency by jointly optimizing central processing unit (CPU)-cycle frequency, transmission power, and channel resource allocation. Additionally, this work incorporates AI and RIS technologies to further enhance the performance and adaptability of MEC networks. The weighting factor is specifically defined based on the residual energy of the IoT sensor battery, ensuring efficient and sustainable network operations. We briefly summarize the state-of-the-art in Table I.

B. Motivation and Contributions

1) Need for Dual-LLM implementation in healthcare: The integration of LLMs within healthcare systems is driven by the critical need to enhance data management and provide personalized assistance. Traditional healthcare systems face significant challenges such as limited personalized patient interaction, the inability to provide real-time responses, and a lack of contextual awareness. These limitations necessitate the deployment of advanced AI systems capable of addressing these gaps effectively.

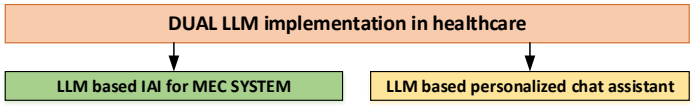


Fig. 1: An illustration of proposed Dual-LLM implementation in healthcare.

As depicted in Fig. 1, the Dual-LLM system integrates two LLM applications within the healthcare framework: one for optimizing MEC network efficiency and data routing through IAI, and the other for providing personalized patient interactions through an advanced chat assistant. The first application of LLMs within our proposed framework is LLM-based IAI framework which adjusts parameters for dynamic user-specific data routing (DUDR) and RIS in a MEC environment. This integration enhances network efficiency and optimizes data flow in dynamic healthcare settings. The IAI uses predictive analytics and advanced data processing to adjust DUDR and RIS configurations, ensuring optimal data flow and signal integrity, especially in real-time health monitoring scenarios. During medical emergencies, the system prioritizes and routes data from critical monitoring devices to healthcare providers, ensuring timely intervention and reducing adverse outcomes. Integrating IAI, the optimization framework enhances network efficiency and user experience under variable conditions and high-stakes healthcare demands. The IAI continuously adapts to changes in the network environment, ensuring appropriate responses to real-time data. During peak usage, such as a pandemic or mass casualty event, the system dynamically allocates bandwidth and resources to maintain efficient communication between medical staff and patients.

A second application involves an LLM-based chat assistant that offers real-time personalized recommendations and guidance to patients. By comparing real-time data with historical records, the assistant provides tailored advice, enhancing patient outcomes and engagement. Using state-of-the-art LLM technologies, this assistant delivers highly personalized, context-aware interactions essential for immediate and accurate patient responses. For instance, a diabetic patient can receive real-time dietary advice and blood sugar updates based on their current and historical data. Employing advanced LLMs, the chat assistant improves patient interaction by offering customized guidance, crucial for managing chronic conditions and adhering to treatment plans. An example includes a post-surgery patient receiving daily check-ins and reminders for medication, physical therapy, and follow-up appointments, thus improving recovery outcomes.

This dual-model architecture is irreplaceable because it decouples responsibilities across two critical layers: (i) the network control plane, managed by LLM-1, and (ii) the patient interaction layer, governed by LLM-2. LLM-1 handles real-time MEC-based optimization and RIS configuration through Interactive AI (IAI), ensuring that time-sensitive physiological data from patients is transmitted with prompt responsiveness and robust signal integrity. Simultaneously, LLM-2 leverages retrieval-augmented generation (RAG) to provide semantic-level, personalized responses based on both current sensor readings and electronic health record histories. A single-model solution cannot simultaneously optimize physical-layer routing and deliver accurate, patient-facing semantic inference. Thus, the proposed Dual-LLM system offers a specialized, mutually reinforcing intelligence architecture essential for modern, context-aware smart healthcare systems.

2) Need for DUDR: In wireless communication, various approaches have addressed challenges in heterogeneous and dynamic network environments, focusing on network optimization, resource allocation, and signal enhancement. However, a comprehensive solution remains elusive. We propose DUDR, a novel approach that optimizes data routing by considering user profiles, device capabilities, and real-time network conditions, thereby redefining wireless communication efficiency. This is particularly critical in healthcare, where reliable and timely data transmission is essential due to patient mobility and dynamic medical device usage. Traditional wireless communication protocols, including proactive (e.g., OLSR), reactive (e.g., AODV), and hybrid (e.g., ZRP) routing schemes, are not inherently designed to address the stringent requirements of healthcare-centric networks. These methods exhibit critical limitations in scalability, context awareness, and real-time adaptability. Moreover, traditional schemes lack contextual intelligence and cannot prioritize data based on

TABLE I: Comparison with significant existing works in AI- and RIS-enabled healthcare communication.

Reference	Key Contributions	Limitations
J. Shao et al., *J. Commun. Inf. Netw., 2024* [8]	Introduces WirelessLLM, a framework for adapting large language models to wireless communication networks, addressing challenges in wireless intelligence.	Conceptual framework; lacks practical implementation and specific focus on healthcare applications.
P. Ramjee et al., *arXiv preprint arXiv:2402.04620, 2025* [14]	Develops CataractBot, an LLM-powered expert-in-the-loop chatbot for cataract patients, improving multilingual access and patient engagement.	Limited to static patient interaction. No integration with RIS, MEC, or real-time data routing frameworks.
Y. Zhu et al., *arXiv preprint arXiv:2402.07016, 2024* [16]	Proposes a RAG-enhanced LLM framework to analyze multimodal electronic health records (EHRs) for clinical outcome prediction.	Operates in batch-mode without real-time adaptation, wireless context-awareness, or integration with RIS-enabled infrastructure.
C. Huang et al., *IEEE Trans. Wireless Commun., 2019* [37]	Investigates the potential of reconfigurable intelligent surfaces (RIS) to enhance energy efficiency in wireless communication systems. Provides theoretical analysis and practical insights into RIS implementations.	Focuses on energy efficiency in wireless communications; does not address AI integration, MEC, or patient-centric healthcare applications.
Our Work (This Paper)	Integrates a Dual-LLM system for RIS-assisted smart healthcare. One LLM controls MEC-level dynamic user-specific data routing and RIS configuration using interactive AI, while the second LLM acts as a patient-specific chat assistant for personalized recommendations. Utilizes a modified PPO algorithm for joint latency-energy optimization, combining communication-level optimization and AI-driven patient interfacing.	

patient vitals or emergency level, undermining the responsiveness of healthcare systems. These protocols also fail to leverage advancements such as RIS, which are crucial for enhancing signal coverage and reliability in complex indoor hospital settings. Finally, conventional routing algorithms are often energy-agnostic, overlooking the critical need for energy-efficient communication among battery-powered medical sensors. These limitations highlight the inadequacy of traditional methods and motivate the need for intelligent, context-aware, and RIS-integrated solutions, such as the proposed Dual-LLM-enabled DUDR framework.

In modern healthcare settings, network congestion is increasingly becoming a critical concern due to the proliferation of medical devices and data-intensive medical applications. Facilities often experience traffic overload, especially during emergencies or pandemics, impacting the timely delivery of patient data. Moreover, critical applications such as remote diagnostics, robotic surgery, and smart intensive care units demand URLLC, where even minor delays can lead to adverse patient outcomes. These challenges necessitate intelligent, edge-driven, and adaptive communication solutions. Recent advancements in LLMs have demonstrated significant potential in healthcare. Our research introduces a multi-source, multi-modal LLM-based chat assistant designed for personalized patient support through advanced retrieval mechanisms. Unlike prior works such as [37] that address RIS optimization or task offloading without AI-driven adaptability, recent LLM-based efforts like CataractBot [14] and REALM for multimodal EHRs [16] focus on static patient interaction and offline clinical analysis, respectively. In contrast, our framework uniquely combines dual-LLMs with real-time RIS control and MEC-based IAI to support both dynamic network optimization and personalized healthcare assistance.

Our integration of RIS and Dual-LLMs addresses healthcare network bottlenecks by enhancing signal propagation, reducing transmission latency, and minimizing network overhead. The RIS dynamically strengthens weak communication links, while the LLM-IAI engine optimizes user-specific routing and resource allocation via deep reinforcement learning. Additionally, the chat-based assistant module provides real-time, privacy-preserving patient interactions and decision support within edge environments, which are essential for responsive and secure healthcare service delivery. By leveraging state-of-the-art LLM technologies, our healthcare assistant provides context-aware, personalized interactions, enhancing patient care and support.

The key contributions of this paper are as follows:

1) Introduction of DUDR System: Developed a novel DUDR system within a MEC environment that adapts routing based on real-time user-specific data. We employ a Greedy Look-Ahead Algorithm (GLAA) for path selection, leveraging a dynamic scoring mechanism tailored to the DUDR framework. This system significantly enhances network performance and ensures consistent data delivery by intelligently managing traffic flow in accordance with the specific demands of healthcare applications.
2) Optimization Framework Using LLM-IAI: Created an LLM-based IAI framework to dynamically adjust DUDR parameters and RIS configurations. This optimization framework aims to enhance network efficiency, optimize data flow, and minimize overhead in a highly dynamic and variable healthcare environment.
3) Real-Time Personalized Healthcare Assistance: Implemented an LLM-based chat assistant that provides real-time personalized recommendations and guidance to patients. This assistant compares real-time data with historical records to offer tailored advice, significantly improving patient outcomes and engagement.
4) Enhanced Network Efficiency and User Experience: Employed advanced deep reinforcement learning (DRL) driven modified PPO algorithm to demonstrate substantial improvements in data transmission efficiency, latency reduction, and overhead minimization compared to conventional network management methods, thereby enhancing overall user experience.

C. Organization

The remainder of this paper is structured as follows: Section II and Section III describe the considered LLM-IAI-based

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2025.3587065

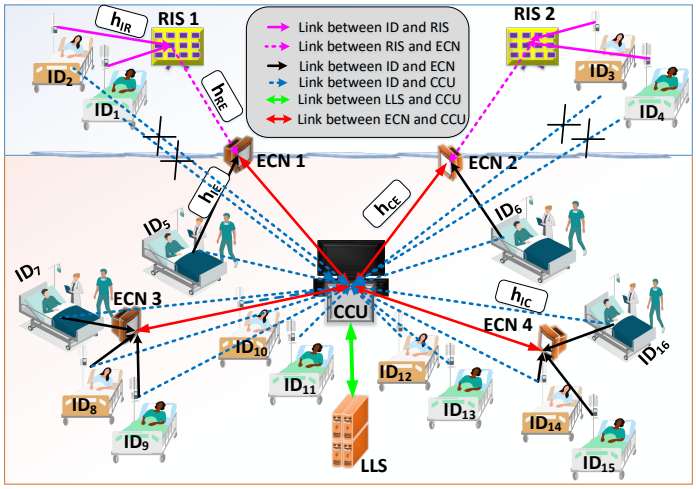


Fig. 2: An illustration of proposed LLMs with IAI network architecture.

MEC system and its signal model, respectively. The optimization problem formulation is presented in Section IV. Section V outlines the proposed solution using the standard PPO and modified PPO algorithms. In Section VI, the proposed solution to implement DUDR is presented. Numerical simulations are presented in Section VII to verify the theoretical results. Finally, Section VIII provides the conclusions and future work of the paper.

II. System Model

In our advanced system model, as depicted in Fig. 2, we integrate LLMs[1] with Interactive AI (IAI) to strengthen modern healthcare communication systems. Edge computing nodes (ECNs) facilitate localized data processing and real-time decision-making, thereby enhancing system responsiveness and ensuring dependable information delivery. The integration of DUDR and RIS optimizes network performance and user experience, managed by a central control unit (CCU) that coordinates signals and manipulates electromagnetic waves to improve signal coverage. LLMs within ECNs analyze healthcare data in real-time, supporting dynamic DUDR management by adjusting data pathways based on user inputs and network conditions for efficient and adaptive delivery. RIS units, governed in real-time by insights from LLM analysis, optimize signal propagation, improve quality, and reduce interference. Our proposed smart healthcare system employs beyond 5G (B5G) networking, the IoT, and AI, shifting from traditional cloud dependencies to edge-centric processing and storage. This system comprises interconnected components for seamless data acquisition, processing, and analysis, with IAI enhancing user interaction and adaptability, thus improving patient care and operational efficiency. Key components of the framework include perception units within ECNs to interpret real-time data from sensors and user inputs, refining data handling protocols and routing decisions adaptively. Action units execute strategies based on real-time analyses, adjusting network configurations and resource allocations instantly. Brain units, utilizing RAG and LLMs, form the decision-making core, synthesizing information from Perception Units and orchestrating the overall network strategy to meet the dynamic needs of the healthcare environment.

[1]LLMs embedded in MEC nodes enhance RIS adaptability by predicting channel dynamics and prioritizing emergency health data using contextual cues derived via RAG and LangChain pipelines. They generate real-time phase adjustment vectors and score-based routing decisions, enabling low-latency, energy-efficient, and emergency-aware communication, which are critical for applications like continuous monitoring and remote triage.

A. Connectivity and Data Flow

As shown in the Fig. 2, IoT devices are widely deployed in the healthcare IoT ecosystem for patient monitoring and data collection, capturing vital signs like heart rate, blood pressure, and temperature. These devices are strategically placed to ensure optimal coverage and patient comfort, and they utilize smart algorithms for adaptive data collection based on patient needs and environmental conditions. ECNs, positioned within healthcare facilities, process this data locally to reduce latency and improve response times, leveraging IAI to enable real-time health monitoring and adaptive responses to patient condition changes. The CCU orchestrates operations, ensuring seamless coordination among components by processing aggregated data with sophisticated AI algorithms, including an IAI framework for enhanced decision-making and dynamic resource allocation. RIS units, controlled by the CCU's IAI-enhanced algorithms, dynamically improve wireless signal quality across the facility, ensuring uninterrupted and optimized data transmission. Interactive AI units, distributed throughout the network, facilitate direct interactions with healthcare professionals and automated systems using RAG and LLMs to provide contextually relevant information and predictive analytics. The LLM server (LLS) has been upgraded with the latest AI models, enabling complex data interactions and real-time learning, supporting predictive diagnostics and personalized treatment plans, thereby extending its capabilities to proactive health management.

The functionality of our advanced healthcare system critically depends on the efficient flow and processing of data, facilitated by state-of-the-art networking technique namely DUDR. Data collected by IoT devices, including critical patient vitals, undergoes initial processing at nearby ECNs. This proximal data handling minimizes latency, enabling rapid responses to essential health metrics. RIS significantly enhance the quality of data transmission within the healthcare facility. By dynamically optimizing signal paths, RIS units ensure robust and efficient communication links between IoT devices and ECNs, thereby strengthening the overall stability and continuity of the network.

Once processed, data is forwarded from the ECNs to the CCU for more comprehensive system-wide analysis. This stage involves sophisticated routing mechanisms that ensure fast and secure data handling across the network. After analysis, the CCU conveys data to the LLS for advanced analytics. The LLS leverages sophisticated AI models to delve deeper into patient health trends, facilitating predictive diagnostics and more informed medical decision-making. DUDR is instrumental in optimizing network traffic flow within the

system. It intelligently prioritizes and routes data based on specific user needs and the urgency of medical situations. For example, in emergencies, DUDR ensures that critical patient information receives precedence, rapidly reaching the required medical personnel while deprioritizing less urgent data. The interaction between the CCU and the LLS embodies the core of our system's data analytics capability. Equipped with advanced AI technologies, the LLS provides deep insights into patient conditions, supporting complex decision-making processes that enhance overall patient care and operational efficiency.

III. Signal Model

The signal model of the smart healthcare system encapsulates the data transmission and processing mechanisms, starting from the collection of data by IoT devices to its final analysis in the LLS. We consider a scenario where multiple single-antenna IoT devices communicate their data through a wireless medium facilitated by an RIS[2] and processed by ECNs, which are equipped with multiple antennas, before reaching the CCU and LLS.

A. Overall Signal Reception at ECNs

Each IoT device is equipped with a single transmitting antenna, whereas each ECN possesses Q number of multiple receiving antennas. The RIS involved in the communication path is composed of N adjustable elements, enhancing signal transmission through dynamic phase adjustments.

The transmitted signal from the k^{th} IoT device, denoted as $x_k(t)$, can be expressed as:

$$x_k(t) = \sqrt{P_k(t)} s_k(t), \tag{1}$$

where $P_k(t)$ is the transmission power of the k^{th} IoT device at time instant t, and $s_k(t)$ is the signal symbol at time t.

Now, the total signal received at the i^{th} ECN, factoring in the path selection, is given by

$$\begin{aligned}\mathbf{y}_{i,k}(t) = {} & a_{i,k}^{\text{dir}} \mathbf{h}_{i,k}^{\text{dir}}(t) x_k(t) \\ & + a_{i,k}^{\text{RIS}} \mathbf{H}_{i,k}^{\text{RIS}}(t) \mathbf{\Phi}(t) \mathbf{g}_k(t) x_k(t) + \mathbf{n}_{i,k}(t),\end{aligned} \tag{2}$$

where $a_{i,k}^{\text{dir}}$ and $a_{i,k}^{\text{RIS}}$ are binary variables indicating the use of the direct and RIS-enhanced paths, respectively. Here, $\mathbf{h}_{i,k}^{\text{dir}}(t) \in \mathbb{C}^{Q\times 1}$ is a vector of channel gains from the k^{th} IoT device to the i^{th} ECN at time instant t, $\mathbf{H}_{i,k}^{\text{RIS}}(t) \in \mathbb{C}^{Q\times N}$ is a matrix representing the channel from the RIS to the i^{th} ECN at time instant t, $\mathbf{\Phi}(t) \in \mathbb{C}^{N\times N}$ is a diagonal matrix representing the phase shifts introduced by the RIS at time instant t, $\mathbf{g}_k(t) \in \mathbb{C}^{N\times 1}$ is the channel gain from the k^{th} IoT device to the RIS at time t, and $\mathbf{n}_{i,k}(t) \in \mathbb{C}^{Q\times 1}$ is a vector of additive noise at time t.

[2]To address practical deployment challenges, we note that RIS panels can be non-invasively integrated within hospital ceilings or walls due to their thin, passive, and low-power nature. This facilitates enhanced indoor wireless coverage without disrupting existing infrastructure.

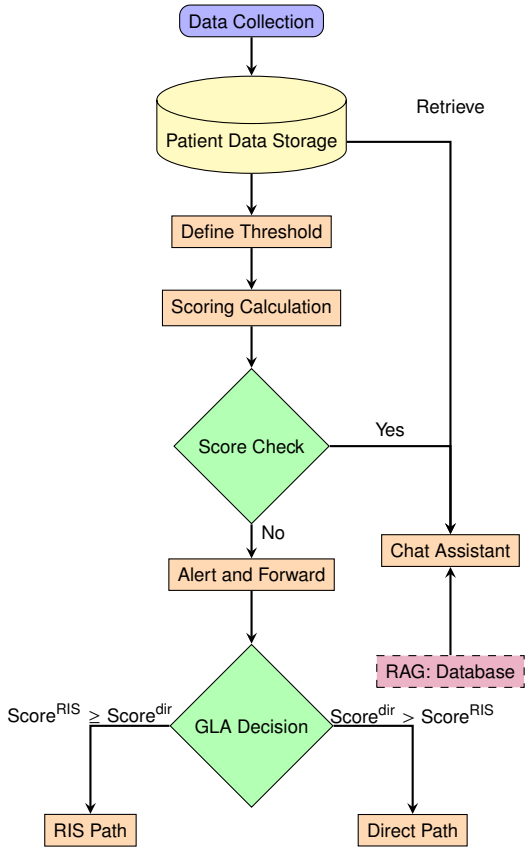


Fig. 3: DUDR process flowchart.

B. RIS Phase Shift Matrix

The effectiveness of the RIS-enhanced signal is significantly influenced by the phase shifts of the RIS elements. These enhancements can be mathematically represented by a diagonal matrix $\mathbf{\Phi}(t)$, which combines both phase adjustments and gain factors:

$$\mathbf{\Phi}(t) = \text{diag}(e^{j\phi_1(t)}, e^{j\phi_2(t)}, \ldots, e^{j\phi_N(t)}), \tag{3}$$

where $\phi_k(t)$ is the phase shift introduced by the n^{th} element of the RIS at time instant t, with N being the total number of elements in the RIS. The term $e^{j\phi_n(t)}$ signifies that each RIS element adjusts the phase of the incoming signal, aligning it optimally for improved propagation and reception. RIS phase shifts are dynamically optimized by a DRL agent guided by LLM-derived policies, which incorporate real-time healthcare context. Factors such as residual device energy influence passive reflection prioritization to conserve power, while sudden changes in patient vitals trigger emergency-aware RIS configurations for low-latency, high-reliability transmission.

C. Dynamic User-Specific Data Routing

The flowchart in Fig. 3 illustrates the DUDR process in a smart healthcare system. It begins with data collection from IoT devices, followed by real-time storage of the collected data. For this purpose, a time-series database, InfluxDB, is employed due to its capability to handle high-frequency write operations and complex queries. These databases are strategically positioned between the IoT devices and the edge server

to ensure efficient data flow. InfluxDB is particularly well-suited for managing time-sensitive health metrics, enabling timely ingestion and consistent availability of patient data for downstream processing.

We now incorporate a real-time scoring function guided by LLM-based IAI at MEC nodes, which interprets patient context, signal metrics, and emergency factors to optimize routing and dynamically reconfigure RIS via a DRL agent. Subsequently, scoring calculations are conducted using a predefined path selection scoring mechanism, as detailed in Subsection III-C-1.

1) Path Selection Scoring Mechanism: The adoption of a scoring mechanism for path selection in smart healthcare systems is driven by the need to optimize data transmission under varying network conditions and clinical urgencies. The scoring approach ensures that every data path is evaluated against a set of critical performance metrics such as signal strength, proximity, user-specific needs, and emergency medical factors. This comprehensive evaluation is crucial because it allows the system to adapt dynamically to changes in the environment or user conditions, ensuring that the most efficient and effective transmission routes are selected for every situation.

The flexibility to use direct links, RIS-enhanced links, or a combination of both based on their scores further enhances the system's robustness, providing redundancy and ensuring continuous service even under suboptimal conditions. This dual-path utilization is particularly advantageous in e-health applications where network stability and data fidelity are paramount. For instance, in remote patient monitoring systems, selecting the most dependable communication path ensures uninterrupted data flow and preservation of critical health information, thereby enabling healthcare providers to make timely and well-informed clinical decisions.

Overall, this scoring method aligns with the goals of modern e-health systems, which strive to offer timely, and patient-centered care. By intelligently routing data based on real-time assessments of network and user conditions, the system supports a wide range of e-health applications, from telemedicine and remote diagnostics to emergency medical response, thereby playing a pivotal role in the digital transformation of healthcare.

Hence, we introduce a comprehensive scoring mechanism for path selection at time instant t which depends on the following parameters:

- Distance (d): The physical distance between the IoT device and the ECN at time instant t, represented by $d_{i,k}(t)$. The distance between two points (x_1, y_1) and (x_2, y_2) at time instant t is calculated as:

$$d_{i,k}(t) = \sqrt{(x_i(t) - x_k(t))^2 + (y_i(t) - y_k(t))^2}. \quad (4)$$

- Signal Strength (SS): Measured for both the direct and RIS-enhanced paths at time instant t, denoted as $SS^{\text{dir}}_{i,k}(t)$ and $SS^{\text{RIS}}_{i,k}(t)$, respectively. The signal strength at time instant t is calculated based on the Log-normal path loss model:

$$SS_{i,k}(t) = P_t(t) - PL_{d0} - 10\gamma \log_{10}\left(\frac{d_{i,k}(t)}{d_0}\right) + \mathcal{N}(0,\sigma). \quad (5)$$

  Here, PL_{d0} represents the path loss at the reference distance d_0, and $\mathcal{N}(0,\sigma)$ denotes the Gaussian random variable with zero mean and standard deviation σ accounting for shadow fading.
- User-Specific Parameters ($U_{i,k}(t)$): These parameters reflect unique user requirements and device characteristics at time instant t influencing data routing decisions.
- Emergency Factor ($E_{i,k}(t)$): The emergency factor at time instant t prioritizes data routing based on the criticality of the monitored condition. It is calculated as a weighted sum of various health indicators, including heart rate, blood pressure, respiratory rate, oxygen saturation, glucose level, ECG readings, and consciousness level. These indicators range from non-critical (0) to life-threatening (1).

The proposed scoring evaluates each potential path i.e., Direct path and RIS-enhanced path, and selects the optimal path or paths based on a composite score as provided below

$$\text{Score}^{\text{dir}}_{i,k}(t) = w_1 SS^{\text{dir}}_{i,k}(t) - w_2 d_{i,k}(t) + w_3 U_{i,k}(t) + w_4 E_{i,k}(t), \quad (6)$$

$$\text{Score}^{\text{RIS}}_{i,k}(t) = w_1 SS^{\text{RIS}}_{i,k}(t) - w_2 d_{i,k}(t) + w_3 U_{i,k}(t) + w_4 E_{i,k}(t), \quad (7)$$

$$\text{Total\_score}(t) = \text{Score}_{\text{user\_to\_RIS}}(t) + \text{Score}_{\text{RIS\_to\_ECN}}(t). \quad (8)$$

where w_1 to w_4 are weights assigned to signal strength, distance, user-specific parameters, and the emergency factor, respectively. These scores help the path selection algorithm dynamically prioritize routes based on an integrated assessment of signal quality, proximity, user-specific needs, and the urgency of the medical situation. In this scenario, the priority of weights would be $w_4 > w_1 > w_2 > w_3$ reflecting the criticality of emergency response and the importance of accurate heart rate monitoring in the e-health system.

The calculated scores are checked against the threshold via a microcontroller:

- Both Scores Below Threshold: If both $\text{Score}^{\text{dir}}$ and $\text{Score}^{\text{RIS}}$ are below the threshold, the data is sent to the LLM-based chat assistant. This process involves retrieving patient details stored in the real-time database and utilizing the RAG mechanism to enrich the LLM's responses.
- Either Score Above or Equal to Threshold: If either $\text{Score}^{\text{dir}}$ or $\text{Score}^{\text{RIS}}$ scores are above or equal to the threshold, the microcontroller further evaluates the emergency level. In cases of high emergency, both a short message service (SMS) and an email are sent to the doctor's mobile number and email address respectively. The SMS is sent via a Global System for Mobile Communication (GSM) module and includes the measured values and the Global Positioning System (GPS) position of the patient [38]. The email, sent through SMTP, provides a more

Algorithm 1 GA for Optimal User-ECN and User-RIS-ECN Selection at Time Instant t.

Initialize: Set of users U, set of ECNs E, set of RISs R, transmit power $P_t(t)$, distance threshold D^{th}
Initialize an empty list Results
for each user $u \in U$ **do**
  best_combination $\leftarrow None$
  best_score $\leftarrow -\infty$
  for each ECN $e \in E$ **do**
    Calculate the $D(u,e,t)$ using (4)
    if $D(u,e,t) > D^{\text{th}}$ **then**
      continue
    end if
    Calculate the $SS^{\text{dir}}(t)$ using (5).
    Calculate the $\text{Score}^{\text{dir}}_{i,k}(t)$ using (6).
    if $\text{Score}_{\text{dir}}(t) \geq$ best_score **then**
      best_score $\leftarrow \text{Score}^{\text{dir}}(t)$
      best_combination $\leftarrow (u, None, e)$
    end if
    for each RIS $r \in R$ **do**
      Calculate the $D(u,r,t)$ using (4).
      Calculate the $D(r,e,t)$ using (4).
      Calculate $D(ur,re,t) = D(u,r,t) + D(r,e,t)$.
      if $D(ur,re,t) > D^{\text{th}}$ **then**
        continue
      end if
      Calculate the $SS^{\text{RIS}}(t)$ using (5).
      Calculate the $\text{Score}^{\text{RIS}}_{i,k}(t)$ using (7)
      if $\text{Score}^{\text{RIS}}(t) >$ best_score **then**
        best_score $\leftarrow \text{Score}^{\text{RIS}}(t)$
        best_combination $\leftarrow (u,r,e)$
      end if
    end for
  end for
  Add best_combination and best_score to Results
end for
return Results

detailed report. For less severe emergencies, only an email notification is sent, providing comprehensive details and allowing the doctor to assess and prioritize the situation based on the detailed information provided.

Finally, after passing the score check, the path selection process determines the routing path using either the Greedy algorithm (GA) or the GLAA, depending on the scenario. The details of these algorithms are described in the following subsections.

2) GA for Optimal User-ECN and User-RIS-ECN Selection: The GA is designed to make immediate, local optimizations by selecting the best combination of users, ECNs, and RISs based on the highest signal strength at the current moment. The steps of the algorithm are provided in **Algorithm 1**.

The problem with the GA is that it only makes decisions based on the immediate scores without considering the potential future states of the network. This can lead to suboptimal path selections since it does not account for the possibility that a slightly lower current score might lead to a much better overall score when future conditions are taken into account. As a result, the GA might commit to paths that seem best in the short term but are less optimal in the long run. Additionally, the execution time of the GA can be high due to the need to evaluate all potential paths without any intermediate filtering.

Algorithm 2 GLAA for Optimal User-ECN and User-RIS-ECN Selection at Time Instant t.

Initialize: Set of users U, set of ECNs E, set of RISs R, transmit power $P_t(t)$, distance threshold D^{th}
Initialize an empty list Results
for each user $u \in U$ **do**
  best_combination $\leftarrow None$
  best_score $\leftarrow -\infty$
  for each ECN $e \in E$ **do**
    Calculate $D(u,e,t)$ using (4)
    if $D(u,e,t) \leq D^{\text{th}}$ **then**
      Calculate $SS^{\text{dir}}(t)$ using (5)
      Calculate $\text{Score}_{\text{dir}}(t)$ using (6)
      if $\text{Score}_{\text{dir}}(t) >$ best_score **then**
        best_score $\leftarrow \text{Score}_{\text{dir}}(t)$
        best_combination $\leftarrow (u, None, e)$
      end if
    end if
    for each RIS $r \in R$ **do**
      Calculate $D(u,r,t)$ using (4)
      if $D(u,r,t) \leq D^{\text{th}}$ **then**
        Calculate $SS_{\text{user\_to\_RIS}}(t)$ using (5)
        Calculate $\text{Score}_{\text{user\_to\_RIS}}(t)$ using (7)
        if $\text{Score}_{\text{user\_to\_RIS}}(t) >$ best_score **then**
          for each ECN $e' \in E$ **do**
            Calculate $D(r,e',t)$ using (4)
            if $D(r,e',t) \leq D^{\text{th}}$ **then**
              Calculate $SS_{\text{RIS\_to\_ECN}}(t)$ using (5)
              Calculate $\text{Score}_{\text{RIS\_to\_ECN}}(t)$ using (7)
              Calculate Total_score(t) using (8)
              if Total_score$(t) >$ best_score **then**
                best_score$\leftarrow$Total_score(t)
                best_combination$\leftarrow(u,r,e')$
              end if
            end if
          end for
        end if
      end if
    end for
  end for
  Add best_combination and best_score to Results
end for
return Results

3) GLAA Algorithm for Optimal User-ECN and User-RIS-ECN Selection: The GLAA improves upon the basic GA by considering potential future network states. This algorithm evaluates immediate and potential future scores to make more strategic routing decisions. The steps of the algorithm are provided in **Algorithm 2**.

The GLAA addresses the shortcomings of the basic GA by incorporating a look-ahead mechanism that evaluates potential future states of the network. It first checks if the user-to-RIS score is greater than the current best score, which acts as a filter to identify promising intermediate links. If the user-to-RIS score is high enough, the algorithm then evaluates the combined user-RIS-ECN score. This approach ensures that decisions are not just based on immediate scores but also consider the potential future benefits, leading to more strategic and optimal path selections. By focusing on promising intermediate links, the GLAA effectively balances immediate and future network performance, resulting in more efficient and effective routing decisions. Furthermore, the intermediate filtering step reduces unnecessary computations, which can

lead to a shorter execution time compared to the basic GA, despite the additional evaluations.

IV. Optimization Framework

Each computational task in the system is characterized by the following parameters: $\tau_{i,k}(t) = \{D_{i,k}(t), C_{i,k}(t), L_{i,k}^{max}(t)\}$, where $D_{i,k}(t)$ represents the data size, $C_{i,k}(t)$ denotes the required CPU cycles per byte to process the data task corresponding to user k and edge computing node (ECN) i at time instant t, and $L_{i,k}^{max}(t)$ specifies the maximum latency tolerance in ms.

A. Decision Making for Task Offloading

The decision to offload a task to the ECN is based on optimizing the trade-off between execution time and energy consumption, under the constraint of maximum tolerance latency. The decision variable is denoted as

$$\beta_{i,k}(t) = \begin{cases} 1, & \text{if the task is offloaded to the ECN,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

In the context of smart healthcare systems, efficient task execution is critical for both local and edge computing environments. This section delineates the computation models for tasks executed on IoT devices and through ECNs.

B. Local Computing on IoT Devices

Local computing is defined based on the computational capability of the IoT device, denoted as $\Omega_{i,k}^{\mathrm{IoT}}(t)$, which represents the CPU cycles per second available for task execution. For a given task processed locally, the execution time $T_{i,k}^{\mathrm{IoT}}(t)$ is determined by

$$T_{i,k}^{\mathrm{IoT}}(t) = \frac{C_{i,k}(t)}{\Omega_{i,k}^{\mathrm{IoT}}(t)}. \tag{10}$$

The energy consumption $E_{i,k}^{\mathrm{IoT}}(t)$ for executing the task locally at the IoT device is calculated as

$$E_{i,k}^{\mathrm{IoT}}(t) = \omega \left(\Omega_{i,k}^{\mathrm{IoT}}(t)\right)^2 C_{i,k}(t), \tag{11}$$

with $\omega = 10^{-26}$ watts/cycle^3 being a coefficient dependent on the chip architecture, reflecting the energy efficiency of the IoT device's processor. This model enables dynamic adjustment of the CPU-cycle frequency via technologies like dynamic voltage and frequency scaling (DVFS) to optimize for energy consumption while meeting execution time requirements.

C. Edge Computing via ECN

In the edge computing model, tasks are offloaded to an ECN for processing. The transmission rate $R_{i,k}(t)$ for sending input data from the IoT device to the ECN over a wireless channel after applying maximum-ratio combining is given by (12), which is presented on the top of the next page. Here, w denotes the bandwidth allocated for transmission ($w = B/N$) and $\sigma_{i,k}^2(t)$ represents the noise power at the i^{th} ECN. $I_{i,k}(t)$ is the total interference power received at the ith ECN from other transmitting devices, which is given as follows

$$I_{i,k}(t) = \sum_i^I \sum_{v\neq k}^K \left| a_{i,v}^{\mathrm{d}}(t)\mathbf{h}_{i,v}^{\mathrm{d}}(t) + a_{i,v}^{\mathrm{RIS}}(t)\mathbf{H}_{v,r}^{\mathrm{RIS}}(t)\mathbf{\Phi}(t)\mathbf{g}_{r,v}(t)\right|^2 P_v(t). \tag{13}$$

The total execution time for edge computing $T_{i,k}^{C}(t)$ includes both the transmission time to the ECN and the computation time at the ECN, and thus computed as follows:

$$T_{i,k}^{\mathrm{ECN}}(t) = \frac{D_{i,k}(t)}{R_{i,k}(t)} + \frac{C_{i,k}(t)}{\Omega^{\mathrm{ECN}}(t)}, \tag{14}$$

where $D_{i,k}(t)$ denotes the data size of the task, and $\Omega^{\mathrm{ECN}}(t)$ the fixed CPU-cycle frequency of the ECN.

The energy consumption for offloading and executing the task at the ECN $E_{i,k}^{\mathrm{ECN}}(t)$ is given as follows

$$E_{i,k}^{\mathrm{ECN}}(t) = \frac{P_{i,k}(t)D_{i,k}(t)}{R_{i,k}(t)}. \tag{15}$$

This model assumes the outcome data size from the ECN back to the IoT device is significantly smaller than the input data size, hence the energy and time costs for returning results are negligible.

In the realm of IoT-based healthcare systems, the dual objectives of minimizing energy consumption and execution latency are paramount, directly impacting both the patient experience and the operational longevity of IoT devices. To address these concerns, our model introduces a novel approach to balancing these objectives, employing a dynamic weighting factor $\lambda_{i,k}(t)$ ($\lambda_{i,k}(t) \in [0, 1]$), which allows for a flexible trade-off between energy savings and latency reduction, tailored to meet patient-specific requirements.

Furthermore, recognizing the critical role of battery sustainability in IoT applications, our framework integrates the battery's residual energy rate $\rho_{i,k}(t)$ into the weighting factor. This adjustment ensures that decision-making aligns with the device's current energy state, promoting energy-efficient operations without compromising service quality. The modified weighting factor is given by

$$\lambda_{i,k}^*(t) = \lambda_{i,k}(t)\rho_{i,k}(t), \tag{16}$$

where $\rho_{i,k}(t) = \frac{E_{i,k}^{\mathrm{res}}(t)}{E^{\mathrm{total}}(t)}$, with $E_{i,k}^{\mathrm{res}}(t)$ representing the maximum available residual energy of the IoT device connected to ECN j and $E^{\mathrm{total}}(t)$ denoting the total battery capacity.

Given this foundation, the overhead for executing a task directly on the IoT device, encapsulating both energy consumption and latency, is formulated as follows

$$O_{i,k}^{\mathrm{IoT}}(t) = \lambda_{i,k}^*(t)T_{i,k}^{\mathrm{IoT}}(t) + (1 - \lambda_{i,k}^*(t))\beta E_{i,k}^{\mathrm{IoT}}(t), \tag{17}$$

where β serves as a normalization factor, equalizing the units between energy consumption and latency to facilitate their direct comparison. This factor is derived from the ratio of the average latency to the average energy consumption across all tasks and devices. Here, $\lambda_{i,k}^t = \lambda_{i,k}^*(t)$ and $\lambda_{i,k}^{\mathrm{ECN}} = (1 - \lambda_{i,k}^*(t))\beta$ respectively represent the adjusted weights for execution latency and energy consumption.

$$R_{i,k}(t) = w \log_2 \left(1 + \frac{\left\| a^{\mathrm{d}}_{i,k}(t)\mathbf{h}^{\mathrm{d}}_{i,k}(t) + a^{\mathrm{RIS}}_{i,k}(t)\mathbf{H}^{\mathrm{RIS}}_{i,r}(t)\mathbf{\Phi}(t)\mathbf{g}_{r,k}(t) \right\|^2 P_k(t)}{\sigma^2_{i,k}(t) + I_{i,k}(t)} \right). \tag{12}$$

In parallel, for tasks offloaded to the ECN, the overhead is calculated as follows

$$O^{\mathrm{ECN}}_{i,k}(t) = \lambda^t_{i,k} T^{\mathrm{ECN}}_{i,k}(t) + \lambda^{\mathrm{ECN}}_{i,k} E^{\mathrm{ECN}}_{i,k}(t). \tag{18}$$

Consequently, the cumulative overhead for the IoT device in executing or offloading the task is determined by

$$O_{i,k}(t) = \beta_{i,k}(t) O^{\mathrm{ECN}}_{i,k}(t) + (1 - \beta_{i,k}(t)) O^{\mathrm{IoT}}_{i,k}(t). \tag{19}$$

This comprehensive approach ensures an optimal balance between energy efficiency and latency, enhancing device autonomy and patient satisfaction in IoT-driven healthcare services.

D. Optimization Problem Formulation using MM

Our objective is to minimize the total overhead in the system by optimizing the allocation and routing of computational tasks and signal paths. Thus, the optimization problem can be formulated as follows:

$$\min_{a_t} \sum\nolimits_{i,k} \left(\beta_{i,k}(t) O^{\mathrm{ECN}}_{i,k}(t) + (1 - \beta_{i,k}(t)) O^{\mathrm{IoT}}_{i,k}(t) \right) \tag{20a}$$

$$\text{s.t.} \quad T^{\mathrm{IoT}}_{i,k}(t) \le L^{max}_{i,k}(t) \quad \text{if } x_i(t) = 0, \tag{20b}$$

$$T^{\mathrm{ECN}}_{i,k}(t) \le L^{max}_{i,k}(t) \quad \text{if } x_i(t) = 1, \tag{20c}$$

$$E^{\mathrm{IoT}}_{i,k}(t) \le E^{\mathrm{res}}_{i,k}(t) \quad \text{if } x_i(t) = 0, \tag{20d}$$

$$E^{\mathrm{ECN}}_{i,k}(t) \le E^{\mathrm{res}}_{i,k}(t) \quad \text{if } x_i(t) = 1, \tag{20e}$$

$$\mathrm{Var}(T_{i,k}(t)) \le V_{\max} \quad \forall i,k, \tag{20f}$$

$$|\phi_n(t)| \in 1, \quad \forall n \in \mathcal{N}, \tag{20g}$$

where the constraints are described as follows. Constraints (20b) and (20c) ensure that the latency for executing tasks, whether locally on the IoT devices or at the ECNs, does not exceed the maximum tolerance threshold $L^{max}_{i,k}(t)$ at time instant t. These constraints are pivotal for real-time healthcare applications where delays can be critical. Constraints (20d) and (20e) limit the energy consumption for executing tasks to the residual energy available on the IoT devices ($E^{\mathrm{res}}_{i,k}(t)$) at time instant t. This is crucial for managing the energy efficiency and operational longevity of battery-powered IoT devices in a healthcare monitoring context. Finally, constraint (20g) ensures that the magnitude of each phase shift $\phi_n(t)$ introduced by the RIS elements is equal to 1 at time instant t.

E. Optimization Problem Formulation using IAIM

Our objective is to minimize the total overhead in the system by optimizing the allocation and routing of computational tasks and signal paths. Thus, the optimization problem can be formulated as follows:

$$\min_{a_t} \sum\nolimits_{i,k} \left(\beta_{i,k}(t) O^{\mathrm{ECN}}_{i,k}(t) + (1 - \beta_{i,k}(t)) O^{\mathrm{IoT}}_{i,k}(t) \right) \tag{21a}$$

$$\text{s.t.} \quad 0 \le a^{\mathrm{dir}}_{i,k}(t) + a^{RIS}_{i,k}(t) \le 1 \quad \forall i,k, \tag{21b}$$

$$\sum_k x_{i,k}(t) = 1 \quad \forall i, \tag{21c}$$

$$R_{i,k}(t) \ge R_{\min} \quad \forall i,k, x_i(t) = 1, \tag{21d}$$

$$(20c), (20e), (20d), (20b), (20f), \tag{21e}$$

where the constraints are described as follows. Constraint (21b) ensures that either a direct or an RIS-enhanced path is selected for each signal transmission between IoT devices and ECNs at time instant t, enforcing exclusivity in path usage. Constraint (21c) ensures that each IoT device is assigned to a unique path at time instant t. Constraint (21d) ensures that the data rate for each IoT device-ECN pair meets the minimum required rate $R_{\min}$ at time instant t. Constraint (20f) ensures that the variance in latency does not exceed the maximum allowed variance $V_{\max}$ at time instant t.

The problem involves significant complexities, including nonlinearity and time-variant dynamics. Hence, we employ DRL techniques due to their adaptability, ability to balance exploration and exploitation, and robustness in handling such challenging environments.

V. Proposed Solution for Optimization problem

Standard proximal policy optimization (PPO) is implemented due to its balance between simplicity and stability, offering a robust training process through its clipped surrogate objective function, which limits policy updates and ensures sample efficiency. This makes PPO a widely adopted choice for reinforcement learning tasks. However, to further enhance performance, modified PPO is introduced, incorporating an adaptive clipping parameter that dynamically adjusts based on the behaviour of the policy. Particularly in non-stationary contexts, this enhancement improves stability and convergence by enabling more flexible and controlled policy updates. Modified PPO's adaptive nature and enhanced stability make it particularly suitable for complex and evolving scenarios, providing more accurate learning outcomes compared to standard PPO.

A. Standard Proximal Policy Optimization (PPO)

The aim of the standard PPO policy optimization technique is to identify an improved policy through recurrent policy updates based on observed trajectories. Moreover, standard PPO is effective because it balances the need for robust updates with the simplicity of implementation. The algorithm achieves this through a clipped surrogate objective function,

which limits the size of policy updates to prevent significant deviations from the old policy, thus ensuring stability during training. The policy is defined by π with the parameter θ_π. In the process of training, stochastic gradient descent (SGD) is used on a mini-batch of L_t transitions (S_t, A_t, R_t, S_{t+1}) in order to identify an optimal policy π^*. The framework for the proposed algorithm is defined as follows:

- State Space:The agent uses the variables in the state space to make decisions and is dependent on the environment. The state space at time instant t is represented as $\boldsymbol{S}_t = \{\boldsymbol{S}_1, \boldsymbol{S}_2\}$, where $\boldsymbol{S}_1$ encapsulates the channel conditions: $\left\{\mathbf{h}_{i,k}^{\text{dir}}(t), \mathbf{H}_{i,k}^{\text{RIS}}(t), \mathbf{g}_k(t)\right\}, \forall i, k$, and $\boldsymbol{S}_2$ covers synchronization parameter: $\{T_{\text{RIS}}\}$, where the system model synchronizes with the RIS operational time-scale $\{T_{\text{RIS}}\}$ to avoid synchronization conflicts.
- Action Space: All the optimizing variables are kept in the action space and are represented by $\boldsymbol{a}_t^n = \left\{\mathbf{x}(t), \lambda(t), \boldsymbol{\Omega}^{\text{IoT}}(t), \beta(t), P(t), \boldsymbol{\Phi}(t)\right\}$, where $a_t^n \in a_t$. The set of all possible actions at time instant t is denoted by $\boldsymbol{a}_t = \left\{\boldsymbol{a}_t^1, \boldsymbol{a}_t^2, \boldsymbol{a}_t^3, \ldots, \boldsymbol{a}_t^n, \ldots, \infty\right\}$. Each element of this set, $\boldsymbol{a}_t^i$ for $i = 1, 2, 3, \ldots$, corresponds to a specific combination of the optimization variables. The action space $\mathscr{A} = \{a_1, a_2, \ldots, a_t,\ a_{t+1}, \ldots, \infty\}$ consists of all possible actions at different time instants.
- Reward function (r_t) : A reward function calculates the obtained reward at time instant t is designed as

$$r_t\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) = \sum\nolimits_{i,k} \Big(\beta_{i,k}(t) O_{i,k}^{\text{ECN}}(t) + (1-\beta_{i,k}(t)) O_{i,k}^{\text{IoT}}(t)\Big) \tag{22}$$

The objective function to update the policy parameters is $\theta_{t+1}^{\pi} = \underset{\theta^{\pi}}{\text{argmax}} \frac{1}{L_t} \sum_{i=1}^{L_t} \nabla_{a_i} \mathcal{L}\left(S_i, a_i; \theta^{\pi}\right)$. In standard PPO, the agent interacts with the environment to find the optimal policy π^* with the parameter θ^{π^*} that maximizes the reward:

$$\mathcal{L}\left(S, a; \theta^{\pi}\right) = \mathbb{E}\left[\frac{\pi_{\theta^{\pi}}(S, a)}{\pi_{\theta^{old}}(S, a)} A^{\pi}(S, a)\right], \tag{23}$$

where the policy distribution of the actor network is represented by $\pi_{\theta^{\pi}}(S, a)$ and that of the old actor network by $\pi_{\theta^{old}}(S, a)$. Here, $A^{\pi}(S, a)$ is the advantage function. To prevent excessive policy updates, standard PPO uses the following clipping surrogate method:

$$\mathcal{L}^{\text{clip}}\left(S, a; \theta^{\pi}\right) = \mathbb{E}\left[\min\left(\frac{\pi_{\theta^{\pi}}(S, a)}{\pi_{\theta^{old}}(S, a)} A^{\pi}(S, a), \text{clip}\left(\frac{\pi_{\theta^{\pi}}(S, a)}{\pi_{\theta^{old}}(S, a)}, 1-\epsilon, 1+\epsilon\right) A^{\pi}(S, a)\right)\right], \tag{24}$$

where the clipping parameter is ϵ. The formula for the advantage estimate A^{π} is $A^{\pi} = R_t + \lambda V(S_{t+1}) - V(S_t)$ where the estimated value of the state S_t is denoted by $V(S_t)$ and the observed return is represented by R_t. The parameters are updated as follows: $\theta^{t+1} = \underset{\theta_\pi}{\text{argmax}} \mathbb{E}\left[\mathcal{L}^{\text{clip}}\left(s, a; \theta_\pi\right)\right]$. The policy is learned using a mini-batch L_t.

B. Modified Proximal Policy Optimization

By adding an adaptive clipping parameter ϵ_t and making changes to the value function update, modified PPO expands on PPO. This parameter, which is obtained from the Kullback-Leibler (KL) divergence between the current policy and the old policy, is dynamically changed based on the behaviour of the policy:

$$\epsilon_t = \epsilon \,\text{sign}\left(\text{KL}\left(\pi_{\theta_{\text{old}}} \| \pi_{\theta_t}\right) - \delta\right), \tag{25}$$

where δ is a small constant and $\left(\text{KL}\left(\pi_{\theta_{\text{old}}} \| \pi_{\theta_t}\right)\right.$ denotes the KL divergence between the previous and the present policies. The fixed threshold ϵ in the surrogate objective function of the traditional PPO is replaced by this adaptive clipping parameter. Adapting ϵ_t dynamically in response to the KL divergence, modified PPO makes sure that policy updates are under control and consistent with the behaviour of the present policy. By allowing for bigger updates when the policy is closer to the original and avoiding excessive modifications when it deviates greatly from it, the adaptive clipping technique enables precise and stable policy updates. The optimization process's stability and convergence are enhanced by this dynamic modification. Until convergence, the algorithm iterates over episodes and steps, updating networks and refining the policy using both standard and modified PPO. The detailed description of the PPO and modified PPO implementation are similar to those outlined in [39].

C. Computational Complexity

We provide the complexity analysis for standard PPO and modified PPO algorithms in **Table II**. Note that both the algorithms possess the similar input layer, the first hidden layer, and the second hidden layer as $9N + 9$, l_1, and l_2, respectively, where N represents the number of RIS elements. Moreover, similar dimensions are considered for the actor-network and critic network of both the algorithms. However, actor and the critic networks in the output layer are different i.e., N and 1, respectively. Consequently, the complexity of the standard PPO algorithm is estimated as $O[2((9N+9)l_1 + l_1 l_2) + l_2(N+1) + \epsilon]$ whereas the complexity analysis for modified PPO is given by $O[2((9N+9)l_1 + l_1 l_2) + l_2(N+1) + \epsilon_t]$. Here, ϵ represents the clip factor of standard PPO and ϵ_t is the clip factor for the modified PPO.

TABLE II: Complexity Analysis of standard PPO and modified PPO [39]

Algorithm	**Complexity**
Standard PPO	$O[2((9N+9)l_1 + l_1 l_2) + l_2(N+1) + \epsilon]$
Modified PPO	$O[2((9N+9)l_1 + l_1 l_2) + l_2(N+1) + \epsilon_t]$

VI. Proposed solution to implement DUDR

The DUDR system is designed to set up a comprehensive healthcare monitoring framework that involves setting up a database, collecting patient health data, evaluating emergency levels based on predefined criteria, and alerting healthcare providers if an emergency is detected. The system follows these steps:

1) Database Setup: Create a real-time database to store patient health data, including heart rate, blood pressure,

temperature, oxygen saturation, respiratory rate, glucose level, ECG readings, and consciousness level.

2) Data Collection: Generate random coordinates for users and edge servers, calculate distances, evaluate signal strengths, and store this data along with simulated health parameters in the database.
3) Emergency Factor Calculation: Calculate an emergency factor based on health data by assessing various health indicators like heart rate, blood pressure, and temperature to determine patient urgency.
4) Evaluation of Scores: Categorize patients into high, low, or no emergency levels by calculating scores for direct and RIS signal paths and sorting patients accordingly.
5) Alert Notification: Send email alerts to healthcare providers for high or low emergency patients by formatting patient data into an HTML email and using SMTP.
6) Visualization with Streamlit: Use Streamlit to visualize patient data and emergency status in an interactive web interface, allowing healthcare providers to view real-time data and monitor patient health metrics and emergency alerts.

This implementation ensures a robust and responsive healthcare monitoring system, enhancing personalized patient care.

VII. Simulation Results

We consider latency and overhead as the primary performance metrics due to their vital impact on real-time responsiveness and resource efficiency in healthcare networks. The simulation parameters are outlined as follows [10]. The number of IoT nodes (K) is set to 100, and the number of ECNs (I) is 10. The number of RISs is 20. The weights $(w_{i,k}^1, w_{i,k}^2, w_{i,k}^3, w_{i,k}^4)$ are $[0.6, 0.5, 0.2, 0.8]$. The CPU frequency of the ECN (Ω^{ECN}) is 4 GHz. The data size ($D_{i,k}$) varies randomly between 300 and 1200 KB. The required CPU cycles for a task ($C_{i,k}$) are 0.110^6 cycles. The CPU frequency range of an IoT node (Ω^{IoT}) is 1 GHz. The maximum transmission power ($P^{\max}$) is 23 dBm. The maximum latency ($L_{\max}^{\text{ECN}}, L_{\max}^{\text{IoT}}$) is 100 ms. The number of RIS elements (N) is 32. The noise power ($\sigma_{i,k}^2$) is 10^{-12} W. The total battery capacity (E^{total}) is 2500 mAH, and the maximum residual energy of an IoT node ($E_{i,k}^{\text{res}}$) is 500 mAH. The bandwidth (B) is 10 MHz. Additionally, the DRL setup parameters are as follows [39]: the learning rate for the critic-network is 0.0002, the soft update coefficient is 0.0005, the learning rate for the actor-network is 0.0001, the mini-batch size is 64, the discount factor is 0.9, the number of neurons for the two hidden layers are $[512, 512]$, the replay buffer capacity is $1,000,000$, and variance of the action noise is 0.1.

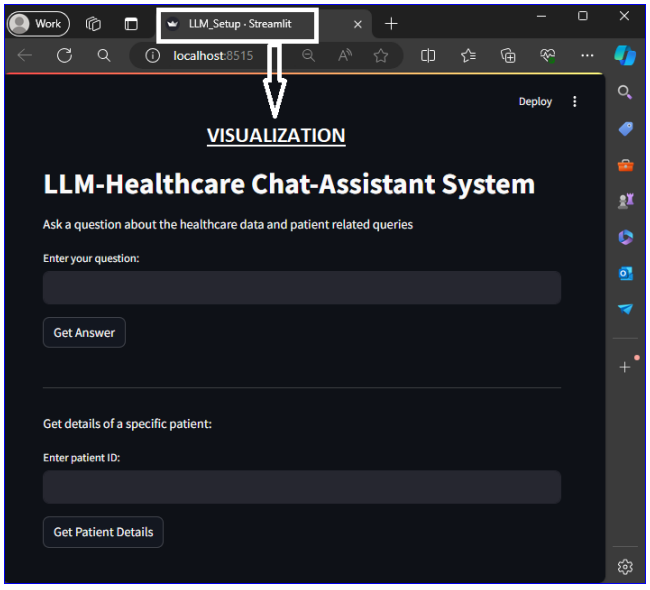


Fig. 4: LLM-based Chat-assistant system.

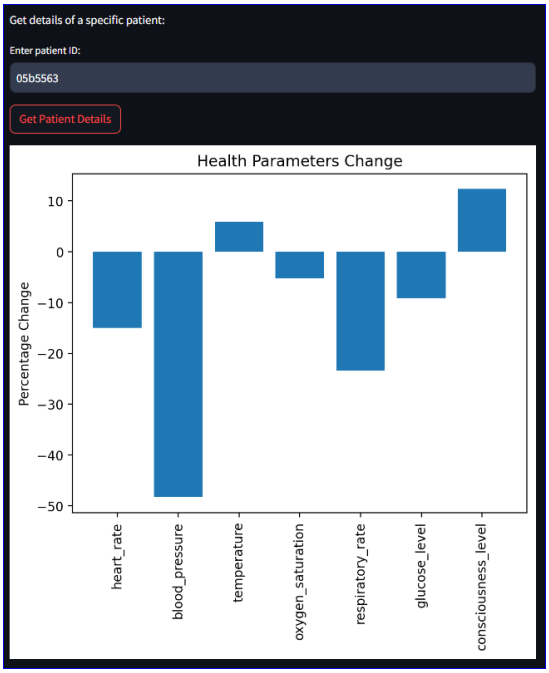


Fig. 5: Comparative performance analysis of the specific patient based on LLM stored data and real-time data.

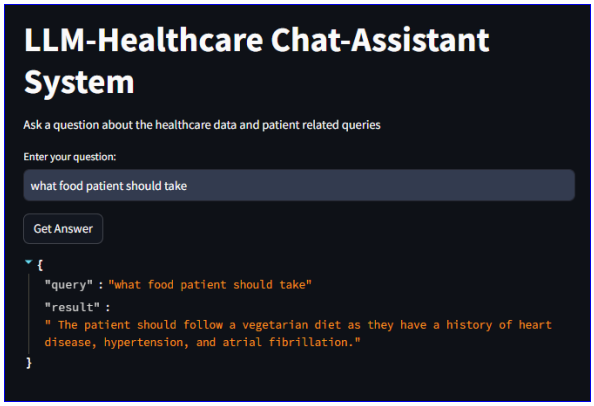


Fig. 6: LLM-based Chat-assistant addressing queries related to patients based on real-time health data.

The comprehensive functionality and integration of the chat assistant system, showcasing its architecture, data analysis capabilities, real-time patient interaction, health monitoring over time, and emergency alert notifications are illustrated by the following figures. Fig. 4 depicts the architecture of the

TABLE III: High-Emergency Patient Alert Details

ID	Emergency Level	Heart Rate (bpm)	Blood Pressure (mmHg)	Temperature (°C)	Oxygen Saturation (%)	Respiratory Rate (breaths/min)	Glucose Level (mg/dL)
02fb2b2	High	106.4688177166527	175.20673459726896	39.42061214045992	96.3631347289573	24.41505696808657	97.21422519852825
00a0258	High	100.22138548881156	120.29109318133378	39.011219967723544	98.94477790224133	23.14558327980297	90.25056001258336

ID	ECG	Consciousness Level (GCS)	Respiratory Distress Symptoms	Location	Google Maps	Timestamp
02fb2b2	Premature Ventricular Contractions	4.952926671804939	None	(97.37754581190126, 20.016399880138672)	Google Maps	2024-05-28 13:28:49
00a0258	Bradycardia	4.81753731671595	None	(23.490463758913382, 23.50230336291763)	Google Maps	2024-05-28 13:28:47

TABLE IV: Low-Emergency Patient Alert Details

ID	Emergency Level	Heart Rate (bpm)	Blood Pressure (mmHg)	Temperature (°C)	Oxygen Saturation (%)	Respiratory Rate (breaths/min)	Glucose Level (mg/dL)
03fcef1	Low	99.4190714460207	113.19362771001719	36.806600266350735	91.02150424909478	17.081330473756574	94.8485430392117
0000000	Low	102.28149856938138	155.13993896209055	37.61902949513211	98.13440088600638	15.430118653565344	91.00634380972996

ID	ECG	Consciousness Level (GCS)	Respiratory Distress Symptoms	Location	Google Maps	Timestamp
03fcef1	Premature Ventricular Contractions	4.93297072362235	None	(67.57490050170526, 26.462463561717453)	Google Maps	2024-05-28 13:36:38
0000000	Ventricular Fibrillation	14.005003470243134	None	(60.3673551870876106, 46.538227141152499)	Google Maps	2024-05-28 13:36:35

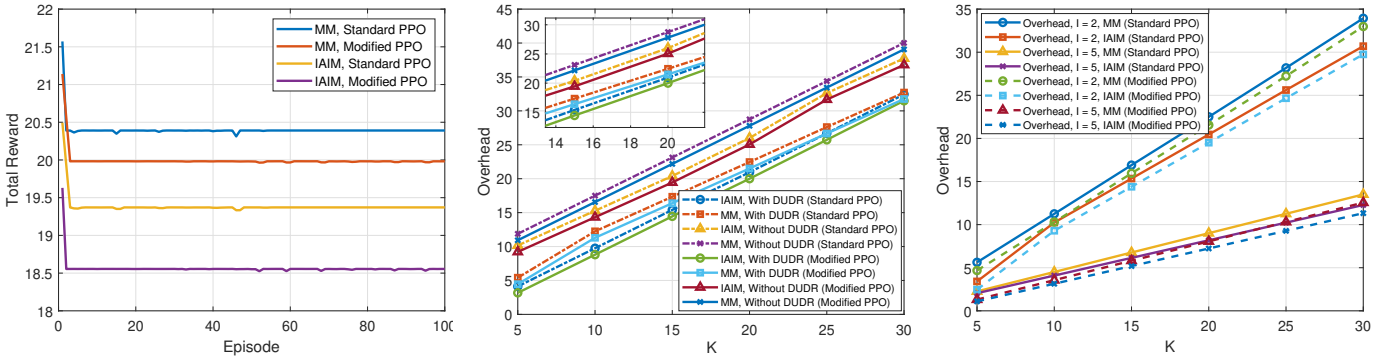
Fig. 7: An illustration of comparative and latest patient details.

chat assistant, showing how it integrates multiple data sources, including real-time patient data and stored information, to provide personalized healthcare support. Fig. 5 demonstrates the system's ability to analyze patient data from both stored records and real-time inputs, highlighting the effectiveness of the chat assistant in delivering accurate and timely health insights. Fig. 6 shows the chat assistant in action, responding to patient queries using up-to-date health metrics to provide relevant and precise advice. Fig. 7 presents a comparison between historical and current health data of patients, demonstrating how the system tracks changes over time to ensure accurate health monitoring. Table III and Table IV show the email alerts generated by the system for high and low-emergency patients, respectively, including health metrics and location information. These alerts ensure that healthcare providers are informed of all emergency levels, with a link to the patient's location for immediate action.

Fig. 8a presents a comparative analysis of the cumulative overhead for an IoT device when executing or offloading tasks using both IAIM and MM, leveraging standard PPO and a modified PPO algorithm. IAIM with the modified PPO algorithm significantly reduces the overhead by approximately 9.6% compared to MM with the standard PPO algorithm, demonstrating its superior efficiency. This performance enhancement is due to the AI model's dynamic adaptability, enabling real-time, data-driven decision-making by continuously adjusting to system and environmental conditions. Unlike MM, which operates on static assumptions, IAIM uses advanced algorithms to proactively adapt to changes, ensuring optimal performance. The modified PPO algorithm contributes to overhead reduction through several key improvements. Adaptive learning rates enable dynamic fine-tuning, allowing the algorithm to respond effectively to changing workloads and network conditions, minimizing unnecessary overhead. The improved exploration-exploitation balance ensures that the algorithm can explore new strategies while refining existing ones, optimizing the decision-making process. Additionally, the better convergence properties of the modified PPO algorithm help in quickly reaching optimal solutions, reducing the time and resources spent on suboptimal strategies. These advancements collectively enhance task execution and offloading efficiency in IoT devices, making IAIM with the modified PPO a more robust solution for minimizing overhead and improving overall system performance. The Dual-LLM architecture semantically interprets system state transitions and guides RIS configuration for adaptive, context-aware decision-making, thereby enhancing convergence and reducing overhead.

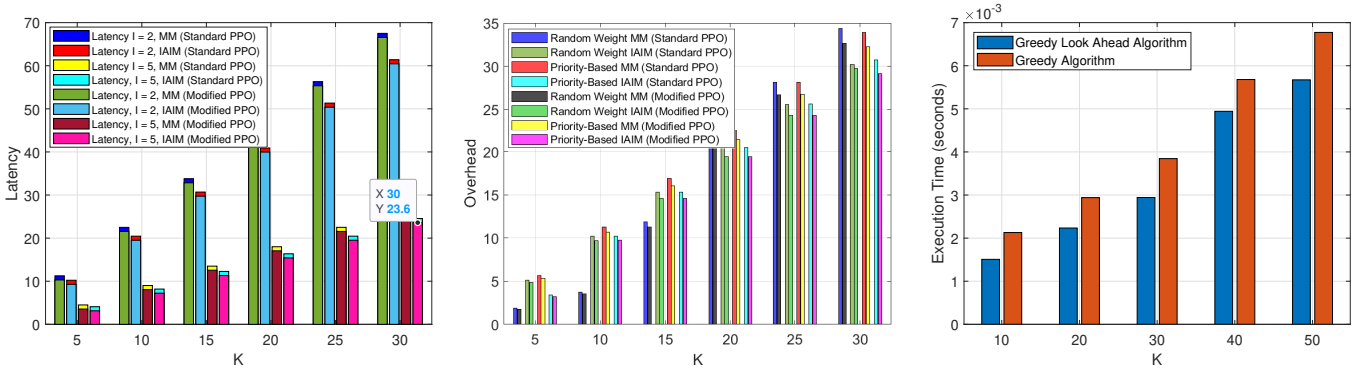
Fig. 8b compares the performance of IAIM and MM in IoT systems, highlighting the impact of DUDR on cumulative overhead across varying numbers of IoT nodes using standard PPO and modified PPO algorithms. The results consistently show that IAIM achieves lower overhead than MM in all scenarios, with DUDR further enhancing this effect, particularly as the number of IoT nodes increases. This trend is attributed to DUDR's ability to dynamically optimize data paths based on real-time conditions and node-specific demands, which reduces unnecessary transmissions and improves network efficiency. In contrast, MM, especially without DUDR, struggles with scalability and adaptability, leading to higher overheads

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2025.3587065



(a) Convergence. (b) DUDR mechanism. (c) Impact of number of ECNs.

Fig. 8: Optimization of network overhead through the implementation of standard PPO and modified PPO.



(a) Impact of number of ECNs on Latency. (b) Impact of weight selection. (c) Execution time of GA and GLAA.

Fig. 9: Optimization of network latency and overhead through the implementation of standard PPO and modified PPO.

and less effective data management. The combination of DUDR with the modified PPO algorithm further amplifies these benefits by leveraging adaptive learning rates, better exploration-exploitation balance, and improved convergence properties, ensuring even more effective overhead reduction. This difference underscores the value of IAIM combined with DUDR in managing large-scale IoT networks, ensuring minimal congestion and maximized operational efficiency. The The LLM framework complements DUDR by analyzing semantic patterns in traffic demands and reconfiguring RIS behavior to dynamically optimize propagation paths, reducing congestion.

Fig. 8c illustrates the effects of increasing IoT nodes and ECNs on cumulative overhead using standard and modified PPO. The data shows that while overhead rises with more IoT nodes, the increase is less pronounced under IAIM than with MM due to IAIM's adaptive management of data traffic. Adding ECNs significantly mitigates overhead increases, as IAIM efficiently utilizes these nodes to enhance network performance. By dynamically reallocating resources and optimizing data flows, IAIM handles higher node densities without a proportional increase in overhead. In contrast, MM's static management approach leads to steeper overhead increases with more IoT nodes due to its inability to effectively distribute workloads among ECNs, resulting in congestion and inefficiencies. The plot demonstrates the benefits of IAIM and the modified PPO algorithm, which maintain lower overhead across various ECN setups. The modified PPO algorithm enhances IAIM's capabilities by optimizing data paths and resource management, using adaptive learning rates, improved exploration-exploitation balance, and better convergence properties. This allows IAIM to dynamically adjust to network conditions, efficiently managing resources, and resulting in more stable and lower overhead as the network scales up. The Dual-LLM system augments scalability by semantically analyzing workload patterns and directing RIS reconfiguration to distribute traffic efficiently across ECNs.

Fig. 9a presents the impact of increasing the number of IoT nodes and ECNs on latency, using standard and modified PPO algorithms. As the number of IoT nodes grows, latency increases; however, IAIM significantly moderates this rise compared to MM due to its advanced real-time adaptive mechanisms that optimize data traffic management and processing efficiency. IAIM consistently reduced latency by approximately 30-40% compared to MM for different number of ECNs. Moreover, IAIM with modified PPO reduced latency by 52.5% compared to IAIM with standard PPO when using five ECNs. Strategic deployment of ECNs under IAIM effectively balances network load, mitigating delays caused by processing and transmission bottlenecks. Continuous real-time adjustments in resource allocation and data flow optimization enable IAIM to sustain prompt responsiveness even as the

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2025.3587065

number of IoT nodes increases. In contrast, MM demonstrates limited scalability and adaptability, leading to higher latency with more IoT nodes due to its static resource management approach. This analysis highlights IAIM's distinct advantages, especially when enhanced by the modified PPO algorithm, which further refine IAIM's capabilities through adaptive learning rates, improved exploration-exploitation balance, and superior convergence properties. LLM-enhanced IAIM enables the RIS to respond to latency variations in real time, adapting beam patterns to minimize delays under increasing IoT node densities.

Fig. 9b plots cumulative overhead against the number of IoT nodes, comparing IAIM and MM approaches using either random weight selection (RWS) or priority-based weight selection (PWS) with standard and modified PPO algorithms. The results indicate that PWS markedly reduces overhead as the number of IoT nodes increases compared to RWS, especially in IAIM. The DUDR mechanism in IAIM prioritizes urgent data flows, optimizing resource allocation by focusing on critical tasks and minimizing attention to non-emergency data. In contrast, RWS allocates resources indiscriminately, leading to irregular overheads and inefficiencies as networks scale. This approach results in increased latency due to the system managing a growing volume of less critical data. Prioritizing weights allows the system to serve urgent cases first, preventing non-emergency patients from consuming excessive resources. The graph underscores the benefits of priority-based selection, which minimizes overhead and improves overall network performance by strategically managing data flows and resource usage. This approach is particularly effective in IAI systems, which handle complex decision-making and adapt to changing conditions. Advanced algorithms in IAIM dynamically adjust priorities based on real-time data, showcasing a significant advantage over manual models in large-scale IoT environments. This prioritization ensures critical tasks receive necessary resources promptly, maintaining an efficient and responsive network. By understanding task urgency and patient criticality, the LLM guides RIS to prioritize high-importance links, enabling faster processing of critical data under PWS.

Fig. 9c compares the execution time of the greedy and GLAAs for optimal User-ECN and User-RIS-ECN selection. The figure shows that the GLAA takes less time than the basic GA. This is counterintuitive at first glance, as the look-ahead algorithm involves more steps, including additional evaluations of potential future states. However, the efficiency of the GLAA can be attributed to its intermediate filtering step. By comparing the user-to-RIS score with the current best score before proceeding to further evaluations, the algorithm effectively reduces the number of unnecessary computations. This filtering mechanism ensures that only promising paths are considered in detail, thereby optimizing the overall execution process. In contrast, the basic GA evaluates all potential paths without such preemptive filtering, leading to a higher computational load and longer execution times. Thus, the figure highlights the efficiency of the look-ahead strategy in balancing detailed evaluation with computational resource management, resulting in faster overall execution. The Dual-LLM predicts high-value user–RIS link combinations early in the selection process, allowing the GLAA to focus on semantically relevant candidates and minimize execution time.

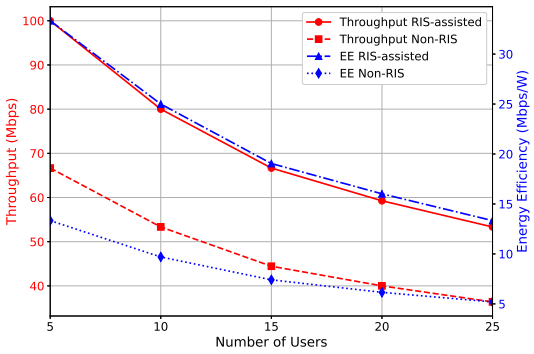


Fig. 10: Throughput and Energy Efficiency versus Number of Users.

Fig. 10 illustrates the performance trends of throughput and energy efficiency for both RIS-assisted and Non-RIS configurations under varying user densities. Throughput is derived as the ratio of transmitted bits to end-to-end latency, while energy efficiency is measured in terms of throughput per unit power consumption. The IAIM framework employed in this study utilizes a Dual-LLM architecture, optimized through a modified PPO algorithm, to enable adaptive and context-aware scheduling decisions. The results reveal that the RIS-assisted system consistently delivers higher throughput across all user loads when compared to the Non-RIS baseline. Despite the expected reduction in throughput with increasing user density, the RIS-enabled approach demonstrates a slower decline and maintains a higher performance margin. Specifically, the RIS-assisted framework achieves up to 1.5 times higher throughput under light traffic conditions and sustains 1.46 times improvement even under heavy user loads. In terms of energy efficiency, the RIS-based system exhibits significantly better performance over the entire range of user densities. The RIS configuration attains more than 2.5 times improvement in energy efficiency compared to the Non-RIS counterpart. This gain remains consistent even as the number of users increases, highlighting the sustainability benefits of RIS deployment in power-constrained scenarios. These enhancements are attributed to the synergy between the reconfigurable nature of RIS and the intelligent decision-making capability of the Dual-LLM agent. The RIS infrastructure enables real-time adaptation of the wireless channel, while the LLM-driven policy introduces semantic reasoning into the user scheduling process. The PPO-based optimization further ensures robust convergence and effective learning in dynamic environments, affirming the proposed system's effectiveness in enabling scalable and energy-efficient wireless healthcare communication.

Fig. 11 compares the proposed Dual-LLM-based RIS model with a conventional static RIS model in terms of average latency and efficiency score across increasing numbers of users. A static model refers to a system with fixed RIS phase configurations and routing policies, which remain unchanged regardless of user load, emergency status, or network dynamics [25]. It treats all communication uniformly and lacks any form of real-time adaptation. In contrast, the proposed system incor-

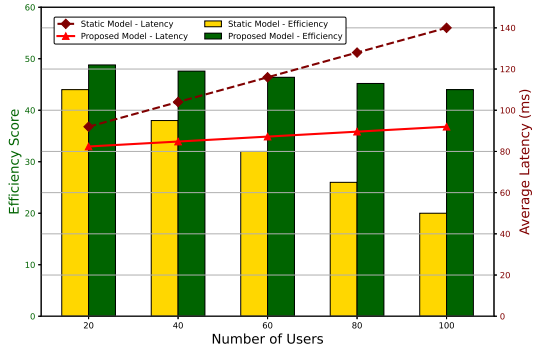


Fig. 11: Average latency and Efficiency score versus Number of Users.

porates a priority-based routing mechanism that differentiates between low- and high-emergency patients. Low-emergency patients are handled by the RAG module, which semantically interprets medical context and queues non-critical data. High-emergency patients, on the other hand, are routed directly via dynamically reconfigured RIS paths and accelerated through MEC servers. This results in adaptive resource allocation, where RIS beamforming, bandwidth, and computing slots are distributed in real time based on patient urgency and network conditions. The average latency is computed as the mean end-to-end transmission delay across all active user sessions, considering both routing decisions and RIS-induced path updates. For each user, this includes transmission, propagation, processing, and RIS reconfiguration delays, weighted according to emergency priority. The efficiency score $\eta(t)$ is derived from the system's total overhead formulation and is defined as $\eta(t) = \frac{1}{\sum_{i,k}\left[\beta_{i,k}(t)\cdot O_{i,k}^{\mathrm{ECN}}(t)+(1-\beta_{i,k}(t))\cdot O_{i,k}^{\mathrm{IoT}}(t)\right]}$. This efficiency score reflects how well the system manages time-sensitive and energy-constrained communication, effectively balancing emergency response quality and resource utilization. As the number of users increases, the proposed model demonstrates significantly lower latency and higher efficiency by ensuring that emergency data receives prioritized access while maintaining balanced system load. The static model, lacking such differentiation and dynamic routing capability, suffers from performance degradation under the same conditions.

VIII. Encryption and Data Protection Mechanisms

While the primary focus of this work is on the design, optimization, and performance evaluation of a novel Dual-LLM framework integrated with RIS for healthcare networks, data privacy, security, and regulatory compliance remain critical for any practical deployment. Future extensions of this framework will explore the integration of lightweight encryption protocols such as TLS 1.3 and elliptic curve cryptography (ECC) to ensure secure data transmission across IoT devices, edge computing nodes, RIS controllers, and LLM inference servers. Anonymized data processing can be enabled through homomorphic encryption, while federated learning paradigms will be considered to support decentralized model training without sharing raw patient data. For access control and auditability, blockchain-based smart contracts offer a promising direction by providing immutable logs and fine-grained policy enforcement. These strategies are essential to align with healthcare regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), which mandate strict protections on sensitive patient information. As this paper presents the first known integration of Dual-LLMs and RIS in healthcare systems, the current scope is limited to validating the core system model and its efficiency. Building upon this foundation, the incorporation of privacy-preserving mechanisms and compliance-aware security layers will be central to future development and real-world deployment of the proposed architecture.

IX. Conclusions and Future Work

This study explored the application of LLMs in advanced wireless networks for smart healthcare by integrating LLM-based IAI to optimize DUDR and RIS within a MEC environment. The proposed DUDR mechanism adapts routing decisions based on real-time user-specific data, thereby improving network efficiency and ensuring consistent, high-quality data transmission tailored to healthcare demands. The framework included an LLM-based chat assistant for personalized real-time healthcare assistance and an optimization framework to minimize network overhead. Using a DRL-driven modified PPO algorithm, the approach significantly improved data transmission efficiency and reduced latency. Simulations showed that IAIM and modified PPO reduced overhead by 9.6% compared to MM and standard PPO. IAIM consistently reduced latency by approximately 30 – 40% compared to MM for different number of ECNs. Moreover, IAIM with modified PPO reduced latency by 52.5% compared to IAIM with standard PPO when using five ECNs. The GLAA and PWS mechanisms dynamically optimized data paths and resource allocation, minimizing unnecessary transmissions and enhancing network efficiency. This ensured minimal congestion and maximized operational efficiency, demonstrating the value of IAIM and DUDR in managing large-scale IoT networks. Future work will also address current limitations by integrating secure learning frameworks such as federated learning and homomorphic encryption to ensure compliance with HIPAA and GDPR. Lightweight LLM variants and model compression will be employed to reduce inference latency at the edge. Explainable AI and domain-specific fine-tuning will enhance output reliability and trustworthiness. Real-world deployment with commercial internet of medical things (IoMT) and RIS platforms will validate scalability and interoperability. Beyond healthcare, the framework can be extended to other latency-critical domains such as industrial IoT, intelligent transportation systems, and disaster recovery networks, where dynamic edge intelligence and optimized routing are equally essential.

References

[1] A. Yarali, "From 5G to 6G: Technologies, architecture, AI, and security," in *Future 6G Networks*, 2023, pp. 185–201.

[2] T. Nguyen, H. Nguyen, A. Ijaz, S. Sheikhi, A. V. Vasilakos, and P. Kostakos, "Large language models in 6G security: challenges and opportunities," *arXiv preprint arXiv:2403.12239*, 2024.

[3] J. Sauvola *et al.*, "LLM and GPT technologies in 6G networks," in *Proc. in 6G Symposium*, 2024.

[4] S. Tarkoma *et al.*, "Societal, ethical, and economic aspects of LLMs in 6G networks," *University of Helsinki Faculty of Science Communications*, Apr. 2024.

[5] L. Karaçay *et al.*, "Guarding 6G use cases: a deep dive into AI/ML threats in all-senses meeting," *Annals of Telecommunications*, Apr. 2024.

[6] A. V. Vasilakos *et al.*, "Synergy between LLMs and blockchain technology for 6G security," *Annals of Telecommunications*, 2024.

[7] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," vol. 109, no. 7, July 2021, pp. 1166–1199.

[8] J. Shao, J. Tong, Q. Wu, W. Guo, Z. Li, Z. Lin, and J. Zhang, "WirelessLLM: Empowering large language models towards wireless intelligence," 2024. [Online]. Available: https://arxiv.org/abs/2405.17053

[9] W. Lee and J. Park, "LLM-empowered resource allocation in wireless communications systems," 2024. [Online]. Available: https://arxiv.org/abs/2408.02944

[10] R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, S. Sun, X. Shen, and H. V. Poor, "Interactive AI with retrieval-augmented generation for next generation networking," *arXiv preprint arXiv:2401.11391*, 2024.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *arXiv preprint arXiv:2005.11401*, 2021.

[12] M. Kamali and L. Petre, "Comparing routing protocols," in *Proc. in ICECCS*, 2015, pp. 206–209.

[13] L. Athota, V. K. Shukla, N. Pandey, and A. Rana, "Chatbot for healthcare system using artificial intelligence," in *Proc. in IEEE ICRITO*, 2020, pp. 978–983.

[14] P. Ramjee, B. Sachdeva, S. Golechha, S. Kulkarni, G. Fulari, K. Murali, and M. Jain, "CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients," *arXiv preprint arXiv:2402.04620*, 2024.

[15] A. Tayal and A. Tyagi, "Dynamic contexts for generating suggestion questions in rag based conversational systems," *Proc. in Companion ACM Web Conf.*, vol. WWW '24, no. 1, p. 3651905, May 2024. [Online]. Available: http://dx.doi.org/10.1145/3589335.3651905

[16] Y. Zhu, C. Ren, S. Xie, S. Liu, H. Ji, Z. Wang, and C. Pan, "REALM: RAG-driven enhancement of multimodal electronic health records analysis via large language models," *arXiv preprint*, 2024.

[17] T. Nadarzynski, O. Miles, A. Cowie, and D. Ridge, "Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study," *Digital Health*, vol. 5, p. 2055207619871808, 2019.

[18] S.-G. Park, A. Kim, T. Yoon, C. Kamyod, and C. G. Kim, "A study of generative large language model for healthcare," in *Proc. 7th InCIT.,*, Nov. 2023, pp. 397–400.

[19] H.-W. Hu, Y.-c. Lin, C.-H. Chia, E. Chuang, and Y. Cheng Ru, "Leveraging large language models for generating personalized care recommendations in dementia," in *Proc. IEEE iWEM*, July 2024, pp. 1–4.

[20] M. A. Hossain, W. Liu, and N. Ansari, "Computation-efficient offloading and power control for MEC in IoT networks by meta-reinforcement learning," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16 722–16 730, May 2024.

[21] Y. Liu, M. Peng, G. Shou, Y. Chen, and S. Chen, "Toward edge intelligence: Multiaccess edge computing for 5G and internet of things," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6722–6747, 2020.

[22] T. Zhao, L. He, X. Huang, and F. Li, "DRL-based secure video offloading in MEC-enabled IoT networks," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18 710–18 724, 2022.

[23] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surv. Tutor.*, vol. PP, no. 99, pp. 1–1, 2017.

[24] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. PP, no. 99, pp. 1–1, 2017.

[25] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, no. 99, pp. 5896–5907, 2017.

[26] P. Zhao, H. Tian, C. Qin, and G. Nie, "Energy-saving offloading by jointly allocating radio and computational resources for mobile edge computing," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.

[27] O. Muoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4738–4755, 2015.

[28] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, 2016.

[29] L. Yang, J. Cao, H. Cheng, and Y. Ji, "Multi-user computation partitioning for latency sensitive mobile cloud applications," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2253–2266, 2015.

[30] L. Ni, J. Zhang, C. Jiang, C. Yan, and K. Yu, "Resource allocation strategy in fog computing based on priced timed petri nets," *IEEE Internet Things J.*, vol. PP, no. 99, pp. 1–1, 2017.

[31] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, 2016.

[32] S. T. Hong and H. Kim, "Qoe-aware computation offloading scheduling to capture energy-latency tradeoff in mobile clouds," in *Proc. IEEE SECON*, 2016, pp. 1–9.

[33] Y. Ni, Y. Liu, H. Zhao, H. Cao, N. Kumar, and P. Nehra, "Outage performance analysis of RIS-aided D2D networks for healthcare application," in *Proc. IEEE GLOBECOM Conference*, 2022, pp. 3047–3052.

[34] M. Mercuri, E. Arnieri, R. De Marco, P. Veltri, F. Crupi, and L. Boccia, "Reconfigurable intelligent surface-aided indoor radar monitoring: A feasibility study," *IEEE J. Electromagn. RF Microw. Med. Biol.*, vol. 7, no. 4, pp. 354–364, 2023.

[35] A. Fontes, I. de L. Ribeiro, K. Muhammad, A. H. Gandomi, G. Gay, and V. H. C. de Albuquerque, "AI-empowered data offloading in MEC-enabled IoV networks," 2022. [Online]. Available: https://arxiv.org/abs/2204.10282

[36] Z. Yang, M. Chen, X. Liu, Y. Liu, Y. Chen, S. Cui, and H. V. Poor, "AI-driven UAV-NOMA-MEC in next generation wireless networks," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 66–73, Oct. 2021.

[37] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.,*, vol. 18, no. 8, pp. 4157–4170, 2019.

[38] K. Aziz, S. Tarapiah, S. H. Ismail, and S. Atalla, "Smart real-time healthcare monitoring and tracking system using GSM/GPS technologies," in *Proc. in 3rd MEC ICBDSC*, 2016, pp. 1–7.

[39] S. Kurma, M. Katwe, K. Singh, C. Pan, S. Mumtaz, and C.-P. Li, "RIS-empowered MEC for URLLC systems with digital-twin-driven architecture," *IEEE Trans. Commun.*, vol. 72, no. 4, pp. 1983–1997, Apr. 2024.