Intent-Based Networking Framework for IoMT Communications in Space–Air–Ground-Integrated Systems

Jianhui Lyu, *Senior Member, IEEE*, Shakila Basheer, *Senior Member, IEEE*, and Keqin Li, *Fellow, IEEE*

***Abstract*—The rapid advancement of Internet of Medical Things (IoMT) technologies and the growing demand for ubiquitous healthcare services have created an urgent need for intelligent networking solutions that can seamlessly integrate space, air, and ground communication infrastructures. This article presents a novel intent-based networking (IBN) framework designed for IoMT communications within space–air–ground-integrated systems. The proposed framework addresses the complex requirements of healthcare applications by introducing an intelligent intent understanding and translation mechanism that can automatically configure network resources based on high-level medical service requirements. Our approach incorporates four core components: medical intent construction for IoMT scenarios, intelligent intent classification using enhanced bidirectional encoder representation from transformers -convolutional neural network (BERT-CNN) models, sophisticated intent parsing through the GlobalPointer-based entity extraction, and dynamic intent translation for real-time network policy generation. Experimental results demonstrate that our framework achieves 94.37% accuracy in intent classification and 83.76% *F*1-score in entity extraction. Experimental validation across space–air–ground network simulations demonstrates substantial improvements in resource utilization efficiency (23.4% increase), bandwidth allocation optimization, and latency reduction (18.6% improvement), directly enhancing patient care capabilities and clinical decision-making reliability in distributed healthcare environments.**

***Index Terms*—Healthcare communications, intent-based networking (IBN), Internet of Medical Things (IoMT), space–air–ground integration.**

I. Introduction

THE advances in healthcare facilities toward digitalization and the consequent mushrooming of Internet of Medical Things (IoMT) devices have posed an extraordinary demand for communication networks to be reliable, low latency, and highly available [1], [2], [3]. These days, healthcare environments consider distributed sensor networks, wearable monitoring devices, robotic surgical systems, and telemedicine platforms that require seamless provisioning of services with geo-distributed locations and application contexts [4]. Integration of various networks composed of space-based satellite communications, airborne relays, and terrestrial communication systems can not only provide a rare opportunity for the realization of end-to-end ubiquitous healthcare service but also can add great complexities to the deployment, management, and orchestration of such networks. Most of the traditional practices of configuring networks use manual policy-based definitions along with statically allocated network resources that cannot serve for highly dynamic and heterogeneous communications of IoMT, in which service requirements change drastically based on patient conditions, emergencies, and clinical procedures [5], [6].

The concept of intent-based networking (IBN) has become a pressing phenomenon in enabling network administrators and healthcare professionals to specify high-level service requirements using natural language expressions in opposition to technical configurations [7], [8]. Such a capability becomes of utmost value in medical environments wherein nurses and other personnel could express their connectivity needs regarding inpatient treatment objectives, treatment protocols, and service quality expectations, among others, without much networking background. The intent-driven approach, thus, sees the automatic attribution of networking policies from healthcare-specified requirements with the dynamic resource allocation, quality-of-service optimization, and failure recovery mechanisms that best follow medical service priorities [9]. However, the application of IBN principles to IoMT communications in space–air–ground-integrated systems comes forth with particular challenges from the distinct nature of healthcare applications, the critical nature of medical data transmission, and the intricate interplay among various communication layers [10].

Current IBN implementations exhibit three critical limitations when applied to medical environments. First, existing intent understanding mechanisms lack medical domain knowledge, resulting in misinterpretation of clinical terminology and context-dependent urgency indicators that are fundamental to healthcare communications. Second, traditional IBN

Received 7 July 2025; revised 7 October 2025; accepted 1 November 2025. Date of publication 4 November 2025; date of current version 27 April 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62202247; in part by the Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2025R195; and in part by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. *(Corresponding author: Jianhui Lyu.)*

Jianhui Lyu is with the Multi-Modal Data Fusion and Precision Medicine Laboratory, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China (e-mail: lvjianhui2012@163.com).

Shakila Basheer is with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia (e-mail: sbbasheer@pnu.edu.sa).

Keqin Li is with the College of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/JIOT.2025.3628714

systems operate on general networking principles that fail to account for the strict latency requirements and reliability standards mandated by medical-device regulations and patient safety protocols. Third, current multilayer network coordination approaches do not address the unique challenges of medical data routing, where patient privacy requirements may dictate specific transmission paths while emergency scenarios demand immediate resource preemption capabilities. These gaps become more pronounced in space–air–ground-integrated systems, where the network topology changes dynamically and medical data must traverse heterogeneous communication domains while maintaining the regulatory compliance and clinical effectiveness.

Motivation was instilled for developing an IBN architecture specialized for IoMT communications because of some major factors setting the requirements of healthcare networking apart from those of a general-purpose communication system. IoMT devices generate many data types, including high-frequency physiological signals needing transmission with an almost nil latency and huge medical imaging files demanding enormous bandwidth allocation [11]. The criticality of medical data varies tremendously, with life-supporting device communications needing guarantees with the shortest delay, while the least critical health monitoring data may tolerate jitter with slight variations. Then, even in applications, such as healthcare, where the data privacy and security, as well as data transmission reliability concerns, are to be addressed, a dynamic enactment of policy needs to set in, factoring in ever-changing clinical contexts as well as the patient's condition [12], [13].

For IoMT applications, space–air–ground-integrated communication systems offer a unique advantage by providing full coverage of remote areas in disaster zones and the maritime environment where terrestrial networks do not exist or have been compromised [14], [15]. The satellite communication links can make telemedicine consultations or remote patient monitoring possible by establishing global connectivity, whereas airborne platforms provide quick deployment of communication infrastructures for emergency medical response [16]. In contrast, ground networks serve the urban healthcare centers and research institutions with high-capacity, low-latency connectivity. Integrating these diverse communication layers is a matter of intelligent orchestration that should be able to select the best transmission paths dynamically, allocate resources across various network segments, and preserve service continuity whenever the network topology changes or parts fail [17].

In order to develop an effective IBN framework for IoMT communication, several basic research questions must be addressed pertaining to understanding intent, natural language processing for medical terminology, healthcare-specific entity extraction, and policy translation mechanisms that can be tailored to the unique needs of medical applications. Our work presents a comprehensive framework that intersects advanced machine-learning technologies with domain-specific knowledge representation to create an intuitive, efficient, and reliable network management method for IoMT communications. The core mechanism of the proposed approach relies on the latest natural language processing models enriched with medical domain knowledge for intent classification and entity extraction, taking into consideration specialized QoS mapping mechanics that translate healthcare service needs into specific network policies optimized for space–air–ground-integrated systems.

The main contributions of this article are summarized as follows.

1) We present a groundbreaking IBN framework that fundamentally transforms how IoMT communications are managed across space–air–ground-integrated systems.
2) We develop a medical intent construction methodology that establishes the first systematic taxonomy for IoMT entities across heterogeneous communication layers.
3) We introduce an enhanced bidirectional encoder representation from transformers -convolutional neural network (BERT-CNN) architecture specifically optimized for medical intent classification that combines the deep semantic understanding capabilities of transformer-based models with the efficient local pattern recognition strengths of convolutional neural networks.
4) We design an advanced GlobalPointer-based entity extraction system that effectively handles the nested and discontinuous medical entity structures commonly found in healthcare communications, overcoming the limitations of traditional sequence labeling approaches that struggle with complex clinical terminology patterns.
5) We create an intent translation mechanism that bridges the semantic gap between high-level medical intents and concrete network policies, incorporating specialized quality-of- service mappings for different healthcare applications ranging from routine patient monitoring to emergency medical responses and real-time surgical guidance.

The rest of this article is organized as follows. Section II establishes the foundational architecture. Section III delves into the intelligent medical intent classification system. Section IV advances our discussion to the entity extraction capabilities. Section V completes our technical framework by exploring the intent translation mechanism. Section VI presents the experimental results. Finally, Section VII concludes this article.

II. IoMT Intent Architecture and Construction

A. Logical Architecture Framework

The IBN architecture designed for IoMTs communications within space–air–ground-integrated systems is arranged in layers, a choice that streamlines both intent processing and resource governance across the network. As depicted in Fig. 1, the logical setup consists of five core layers: the IoMT data layer, the southbound interface, the network control layer, the medical intent northbound interface, and the healthcare application layer. By organizing the framework this way, the design maintains distinct areas of focus but still allows a smooth flow of information between overarching health service needs and the granular network settings that support them.

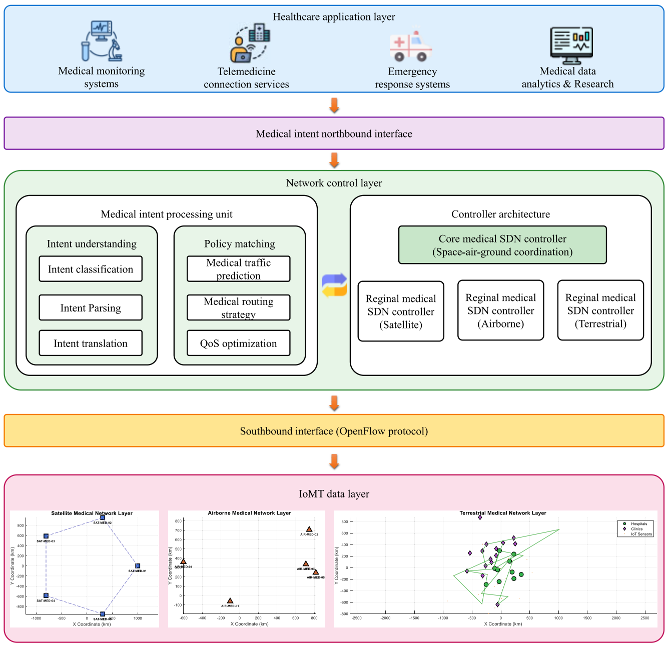

Fig. 1. IoMT IBN logical architecture.

To address diverse regulatory requirements across different geographical regions, our framework implements a regulatory compliance engine within the medical intent processing unit. This engine maintains a dynamic knowledge base of regional healthcare standards, including HIPAA for North America, GDPR for Europe, and local medical data protection laws. When processing medical intents, the system automatically applies the appropriate regulatory constraints based on the geographical context of the requesting healthcare facility. For concurrent request management, we employ a priority-based arbitration mechanism that operates on three levels: emergency medical procedures receive the highest priority with immediate resource allocation, routine medical monitoring operates at medium priority with quality-of-service guarantees, and research applications function at lower priority with best-effort service.

Consider the IoMT data layer as the backbone of our system. It is the spot where all kinds of medical gadgets and sensors connect, whether they are in the air, on the ground, or floating high above. This layer hooks up satellite systems that keep an eye on patients in hard-to-reach places, drone-like emergency teams that can talk to doctors in real time, and the hospital infrastructure that connects research laboratories, clinics, and emergency rooms. To keep things running smoothly, every device speaks the same language and plugs right into an OpenFlow-controlled network that lets us manage everything from one central hub.

Network nodes within the IoMT data layer lack autonomous routing capabilities and instead receive forwarding instructions from centralized SDN controllers operating within the network control layer. These devices can process various types of commands, including status queries, topology updates, anomaly detection requests, and quality-of-service adjustments, executing these operations through preconfigured code modules optimized for IoMT scenarios. The distributed nature of IoMT devices across different communication domains necessitates sophisticated coordination mechanisms that ensure consistent service delivery regardless of the underlying network topology or transmission medium.

The collaboration between the core medical SDN controller and regional controllers follows a hierarchical arbitration protocol. When regional controllers receive competing resource requests, they first attempt local resolution using the available bandwidth and computational resources. If local resources are insufficient, requests are escalated to the core controller with priority classification based on medical urgency scores calculated using patient vital signs, procedure criticality, and temporal constraints. The core controller implements a weighted round-robin scheduling algorithm that considers three factors: medical priority-level weighted at 50%, geographical resource availability weighted at 30%, and network topology stability weighted at 20%. During resource conflicts, the system employs a preemption mechanism where emergency medical procedures can interrupt lower priority connections with automatic rerouting to maintain service continuity. Each regional controller maintains a local intent cache that enables autonomous operation during communication failures with the core controller, ensuring uninterrupted medical services even when interregional coordination is temporarily unavailable.

The data integrity preservation employs the adaptive error correction coding that adjusts redundancy levels based on transmission medium characteristics. Satellite links utilize Reed–Solomon coding with higher redundancy factors, while terrestrial connections apply lighter correction schemes. The cross-domain transmission implements cryptographic checksums and automatic retransmission protocols that account for the unique error patterns of each communication layer.

The southbound interface acts as a translator for different medical devices and network gear, linking the network control layer with the IoMT data layer. It runs on virtualized tech and sticks to standard OpenFlow rules [18], which helps it to cut up the network into slices so resources can be handed out where they are needed most.

Device authentication follows a multitier approach. Emergency medical devices undergo accelerated verification through preshared cryptographic certificates. Nonemergency devices complete standard authentication protocols, including digital signatures and role-based access validation. The system maintains a distributed trust registry across satellite, airborne, and terrestrial controllers to enable rapid device verification during network topology changes.

B. Medical Intent Entity Construction

Developing precise medical intent entities hinges on establishing a thorough taxonomy that mirrors the wide-ranging needs and limitations unique to communications in IoMT. In response, our framework arranges these intent entities into five principal categories, with each category further subdivided into targeted subcategories that reflect particular elements of service delivery in integrated space–air–ground networks.

TABLE I
IoMT Intent Entity Construction Framework

Category	Entity type	Subcategory	Examples
A1	Medical device	Patient monitor	CardiacSensor1, BloodPressureMonitor, OxygenSat_Device
A2	Medical device	Location	ICU_Room1, Ambulance_Unit3, Satellite_MedStation
A3	Medical device	Network domain	TelemedicineNet, EmergencyResponseNet, ResearchCollabNet
B1	Healthcare service	Monitoring	EstablishVitalSigns, StartPatientTracking, EnableRealTimeAlert
B2	Healthcare service	Network setup	CreateMedicalVPN, FormEmergencyNetwork, ConfigureBackupLink
B3	Healthcare service	Security	EnableMedicalFirewall, ActivateEncryption, AuthenticateDevice
B4	Healthcare service	Data filter	FilterPatientData, ProcessVitalSigns, AnalyzeTrends
B5	Healthcare service	Load balance	DistributeMedicalTraffic, BalanceServerLoad, OptimizeResources
C1	Medical QoS	Bandwidth	MinBandwidth_50Mbps, ReservedCapacity_100Mbps
C2	Medical QoS	Latency	MaxDelay_10ms, CriticalLatency_5ms
C3	Medical QoS	Throughput	MinThroughput_Medical, OptimalDataRate
C4	Medical QoS	Reliability	PacketLoss_0.1%, UptimeRequirement_99.99%
C5	Medical QoS	Priority	HighPriorityMedical, EmergencyTraffic, RoutineMedical
C6	Medical QoS	Application type	HighDefTelemedicine, VoiceMedicalComm, DataBackup
C7	Medical QoS	Network properties	SelfHealing, AutoRecovery, FaultTolerant
C8	Medical QoS	Protocol requirements	HIPAA_Compliant, HL7_Protocol, DICOM_Transfer
D1	Temporal	Service duration	DailyShift_8to18, WeeklyMaintenance, EmergencyPeriod
E1	Context	Port assignment	MedicalPort_8080, SecurePort_443
E2	Context	Priority level	CriticalPatient, RoutineCheckup, ResearchPriority
E3	Context	Application context	RemoteMonitoring, EmergencyResponse, TeleSurgery, MedicalResearch

Table I accordingly lays out the entire hierarchical structure of the entity framework, a structure designed to promote orderly comprehension of medical intent and to inform the creation of appropriate operational policies.

The framework implements a multitier protocol adaptation mechanism to address varying regional medical regulations. At the base tier, universal medical standards such as HL7 FHIR for data interoperability are consistently applied across all regions. The intermediate tier incorporates region-specific privacy regulations where GDPR requirements for European medical facilities mandate explicit consent tracking and data minimization principles, while HIPAA compliance for North American facilities focuses on access logging and encryption standards. The top tier addresses local medical-device regulations and clinical protocols that vary by healthcare jurisdiction. The system maintains a dynamic regulatory mapping table that associates geographical location identifiers with applicable regulatory frameworks, automatically selecting appropriate protocol stacks when establishing medical connections. During cross-border medical consultations, the framework applies the most restrictive regulations from all involved jurisdictions to ensure compliance, while providing transparency to healthcare providers about any protocol limitations that may affect service capabilities.

Conflict resolution operates through a priority arbitration system that evaluates medical urgency, patient criticality, and resource availability. The framework assigns temporal weights to concurrent requests and applies medical triage principles to resolve conflicts. When overlapping intents occur, the system negotiates resource-sharing agreements and provides alternative suggestions to lower priority requests while maintaining the service quality.

III. Intelligent Medical Intent Classification

A. Medical Intent Classification Dataset Generation

The development of an effective intent classification system for IoMT communications requires a comprehensive dataset that captures the diverse ways healthcare professionals express their networking requirements. Our approach to dataset generation combines automated content creation using large language models with domain-specific medical knowledge and manual validation by IoMT professionals.

The classification framework defines seven primary intent categories that encompass the most common network management scenarios in IoMT environments. These categories include maintaining default medical operations, healthcare network management, establishing medical device connections, terminating medical connections, creating healthcare networks, shutting down medical networks, and retrieving medical system information.

Table II shows the IoMT intent classification categories.

The dataset generation process utilizes ChatGPT in combination with medical domain expertise to create realistic healthcare networking scenarios. Initially, we provide the language model with the IoMT context, including typical hospital operations, emergency response procedures, telemedicine requirements, and remote patient monitoring scenarios. The model generates diverse text samples ranging from 10 to 100 words, representing various ways healthcare professionals might express their network requirements.

The text length range of 10–100 words was selected based on the analysis of actual medical communication patterns observed in clinical environments, where brief urgent requests typically contain 10–15 words, while detailed consultation requests may extend to 80–100 words. Our data preparation

TABLE II
IoMT Intent Classification Categories

Code	Intent category	Description examples
A	Maintain medical default	Continue standard patient monitoring protocols
B	Healthcare network management	Optimize telemedicine bandwidth allocation for cardiac unit
C	Establish medical connection	Connect surgical robot to satellite medical network with 5ms latency, ensure 500Mbps bandwidth for real-time surgical guidance, activate medical-grade encryption
D	Terminate medical connection	Disconnect patient monitor from ICU network
E	Create healthcare network	Establish emergency medical network for disaster response with minimum packet loss and maximum priority for life-supporting devices
F	Shutdown medical network	Deactivate temporary medical network after emergency response
G	Query medical information	Display current status of cardiac monitoring devices in ward B

process involved three stages of quality control: automated filtering using the medical terminology validation to ensure the clinical accuracy, semantic coherence checking to verify logical intent structure, and domain expert review by certified medical informatics professionals. The training, validation, and testing split of 7:2:1 was chosen to maximize model exposure to diverse medical scenarios while preserving sufficient data for robust evaluation, following established practices in medical natural language processing, where the limited domain-specific data require careful allocation to prevent overfitting.

The generated samples undergo multiple validation stages to ensure the accuracy and relevance to real-world medical scenarios. IoMT professionals review each generated intent to verify clinical accuracy, technical feasibility, and alignment with healthcare operational procedures. Python-based regular expression matching and manual verification remove anomalous samples and irrelevant content that does not align with IoMT networking requirements.

The manual verification process involved 12 domain experts: six certified medical informatics specialists with an average of eight years of clinical IT experience, four network engineers with healthcare technology backgrounds, and two practicing physicians familiar with telemedicine systems. Each generated sample underwent review by three independent verifiers using a standardized evaluation protocol that assessed medical terminology accuracy, clinical workflow appropriateness, and technical feasibility.

The bias mitigation employs stratified sampling across medical specialties, geographic regions, and healthcare system types. The dataset includes proportional representation from emergency medicine, cardiology, radiology, and other specialties while incorporating terminology variations from different healthcare standards. Regular bias auditing evaluates the model performance across demographic groups and medical contexts to identify and correct systematic biases.

The clinical validation involves iterative review cycles, where domain experts evaluate classified intents against actual medical scenarios. The validation process includes terminology cross referencing with medical ontologies, clinical workflow verification, and real-time feedback collection from healthcare practitioners. Misclassified intents undergo root cause analysis to improve the model understanding of medical context and terminology variations.

B. Enhanced BERT-CNN Architecture for Medical Intent Classification

Intent integrity protection employs digital signatures and audit trails that track all intent modifications and policy changes. The multifactor authentication verifies user identity before accepting intent submissions, while role-based access controls limit policy modification capabilities. Anomaly detection algorithms monitor for unusual intent patterns or policy changes that might indicate malicious activity and trigger security alerts.

While TextCNN models provide the effective local feature extraction through convolutional operations, they face limitations when processing the complex medical terminology and understanding contextual relationships in healthcare communications. Medical intent expressions often contain specialized terminology, abbreviations, and clinical context that require sophisticated language understanding capabilities beyond simple pattern matching.

The unknown terminology handling employs contextual inference algorithms that analyze surrounding words and medical context to estimate meaning. The system maintains expandable abbreviation dictionaries and applies fuzzy matching techniques for variant spellings. Unrecognized terms trigger expert consultation workflows where medical professionals provide definitions that enhance the system vocabulary for future encounters.

BERT models address these limitations through the bidirectional context understanding and pretraining on large medical text corpora. The transformer-based architecture enables simultaneous consideration of left and right context information, providing an understanding of medical terminology within clinical contexts. BERT's multihead self-attention mechanisms and residual neural networks create multilayered semantic representations that capture different levels of medical language understanding, as shown in Fig. 2.

Text describing medical intents moves through a sequence of well-defined steps in the architecture. Initially, patient notes or clinical queries are tokenized using a domain-specific vocabulary trained for BERT so that every token is associated with one of BERT's prelearned medical embeddings. The first layer of the BERT module then runs its transformer blocks [19], crafting contextualized vector spaces that reflect the

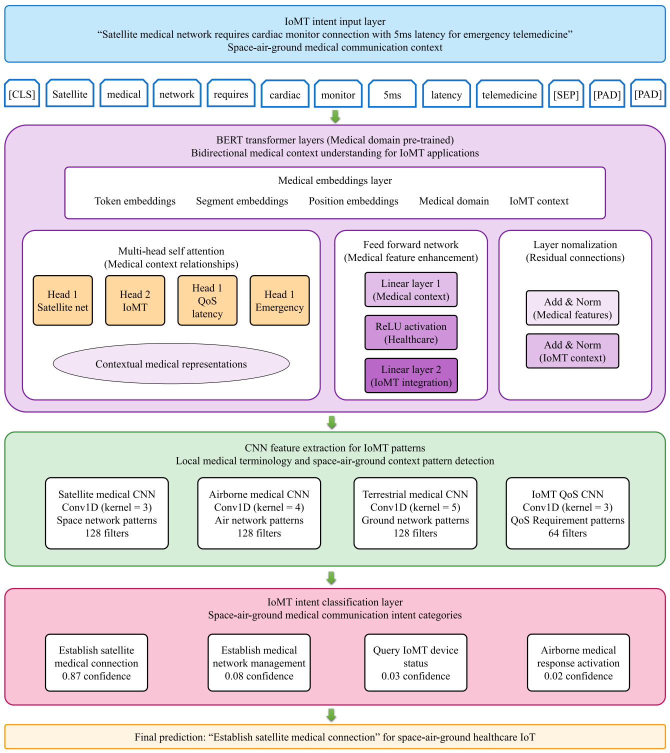

Fig. 2. Enhanced BERT-CNN architecture for medical intent classification.

intricate relationships and specialized terms found in clinical language.

The multihead attention mechanism employs four specialized heads: Head 1 focuses on medical device identifiers and network endpoints, Head 2 processes clinical context and urgency indicators, Head 3 extracts quantitative QoS requirements, including bandwidth and latency specifications, and Head 4 identifies temporal constraints and service duration parameters. Each attention head utilizes learned position embeddings that capture the sequential relationship between medical terms and their associated network requirements, enabling the model to understand that latency specifications typically follow device identifiers in medical intent expressions.

The architecture incorporates incremental learning capabilities through transfer learning mechanisms that update the medical vocabulary without full retraining. New medical terms and device types undergo automated similarity matching with existing entities, while domain experts validate novel additions. The system maintains version control for terminology updates and provides backward compatibility for existing device configurations.

IV. IoMT Intent Parsing

A. Medical Intent-Extraction Dataset Development

For accurate medical intent parsing, extractor systems must manage entity structures that are often multilayered, fragmented, or nonlinear. In fast-evolving medical conversations, these structures rarely slot neatly into a single line of text, so traditional sequence-labeling methods—especially those built on LSTM-CRF architectures [20]—struggle to keep pace. Clinical terms routinely overlap in meaning, show hierarchies, or shift form within a single utterance, rendering standard tagging techniques unreliable.

Cross-sentence entity recognition extends the span evaluation mechanism to analyze inter-sentence relationships through discourse markers and anaphora resolution. The model maintains context windows encompassing multiple sentences and applies co-reference resolution to link fragmented entity mentions. Entity reconstruction algorithms combine partial mentions using semantic similarity matching and medical domain knowledge to form complete entity representations.

To tackle these shortcomings, our intent-extraction dataset was designed to mirror the messy, real-world language used by healthcare professionals. It recognizes that entities are often nested; for instance, a "cardiac monitoring network" implicates both a procedure, "cardiac monitoring," and a hardware layer, "monitoring network." The dataset also captures fragmentation or discontinuity, as seen in the phrase "monitor patient A's and patient B's cardiac rhythm," where identifiers are interrupted by connective text yet still represent distinct entities [21].

The dataset development process follows begin–inside–outside (BIO) tagging [22] conventions adapted for medical contexts. We use B_{MED} (begin-medical) tags for the first token of medical entities, I_{MED} (inside-medical) tags for continuation tokens within medical entities, and O_{MED} (outside-medical) tags for nonmedical tokens. This annotation scheme enables the precise identification of medical entity boundaries while maintaining compatibility with standard named entity recognition frameworks.

For deeply nested entities common in medical contexts, we extend the BIO annotation scheme with hierarchical depth indicators. Triple-nested entities such as "5G-based real-time ICU monitoring network" are annotated using depth-level tags: B-MED-L1 for the outermost entity boundary ("5G-based real-time ICU monitoring network"), B-MED-L2 for the intermediate entity ("real-time ICU monitoring"), and B-MED-L3 for the innermost entity ("ICU monitoring"). Each depth level maintains its own entity type classification, allowing the model to simultaneously recognize the complete system description, the monitoring function, and the specific clinical unit. Discontinuous nested entities are handled using continuation markers, where interrupted nested structures receive I-MED-L1-CONT tags to maintain hierarchical relationships across intervening text.

The terminology adaptation maintains institution-specific vocabularies and regional medical terminology mappings. The system learns the local terminology variations through deployment feedback and creates synonym relationships between equivalent terms in different healthcare contexts. Cross-institutional communication employs terminology translation mechanisms that convert local terms into standardized representations while preserving semantic meaning.

Based on the medical intent entity framework defined earlier, we create 15 000 annotated samples containing nested medical entities, nonnested medical entities, and discontinuous medical entities. The dataset maintains a 7:2:1 ratio for training, validation, and testing sets, ensuring the comprehensive coverage of medical terminology variations and clinical contexts.

B. GlobalPointer-Based Medical Entity Extraction

Conventional sequence-labeling techniques built around LSTM-CRF frameworks struggle with medical datasets that feature nested entities and interrupted terminology patterns. The GlobalPointer architecture [23] overcomes these difficulties by framing entity recognition as a classification of global spans instead of as a step-by-step token-labeling task.

What makes GlobalPointer particularly suited for healthcare texts is its reliance on global normalization mechanisms that efficiently pinpoint medical entities in a complete sequence. Instead of moving through tokens one at a time, the model evaluates every conceivable text span at once, designating each candidate medical entity region as a separate classification target.

For a medical text sequence $\mathbf{t}_{\text{med}}$ with a length n_{med}, the model can identify $n_{\text{med}}(n_{\text{med}}+1)/2$ possible medical entity spans. This comprehensive span enumeration ensures that no potential medical entities are overlooked, even when they exhibit complex nested or discontinuous patterns.

The mathematical formulation for medical entity recognition begins with encoding the input medical text into vector representations. Using transformer-based medical encoders, we convert the input sequence into contextualized embeddings $[\mathbf{h}_{1,\text{med}}, \mathbf{h}_{2,\text{med}}, \ldots, \mathbf{h}_{n,\text{med}}]$.

Following multihead attention mechanisms adapted for medical terminology, we compute query and key representations for medical entity recognition

$$\mathbf{q}_{i,\alpha,\text{med}} = \mathbf{W}_{\alpha,q,\ \text{med}}\mathbf{h}_{i,\text{med}} + \mathbf{b}_{\alpha,q,\ \text{med}} \tag{1}$$

$$\mathbf{k}_{j,\alpha,\text{med}} = \mathbf{W}_{\alpha,k,\text{med}}\mathbf{h}_{j,\text{med}} + \mathbf{b}_{\alpha,k,\text{med}} \tag{2}$$

$$s_{\alpha,\text{med}}(i,j) = \mathbf{q}_{i,\alpha,\text{med}}^{T}\mathbf{k}_{j,\alpha,\text{med}} \tag{3}$$

where α represents different medical entity types, $[\mathbf{q}_{1,\alpha,\text{med}}, \mathbf{q}_{2,\alpha,\text{med}}, \ldots, \mathbf{q}_{n,\alpha,\text{med}}]$ and $[\mathbf{k}_{1,\alpha,\text{med}}, \mathbf{k}_{2,\alpha,\text{med}}, \ldots, \mathbf{k}_{n,\alpha,\text{med}}]$ represent the query and key sequences for medical entity type α. The score $s_{\alpha,\text{med}}(i,j)$ determines whether the span from position i to position j represents a medical entity of type α.

GlobalPointer incorporates rotational position encoding specifically adapted for medical terminology understanding [24]. The rotational encoding enables the model to capture relative positional relationships between medical terms while maintaining translation invariance

$$\begin{aligned} s_{\alpha,\text{med}}(i,j) &= \left(\mathbf{R}_{i,\text{med}}\mathbf{q}_{i,\alpha,\text{med}}\right)^{T}\left(\mathbf{R}_{j,\text{med}}\mathbf{k}_{j,\alpha,\text{med}}\right) \\ &= \mathbf{q}_{i,\alpha,\text{med}}^{T}\mathbf{R}_{i,\text{med}}^{T}\mathbf{R}_{j,\text{med}}\mathbf{k}_{j,\alpha,\text{med}} \\ &= \mathbf{q}_{i,\alpha,\text{med}}^{T}\mathbf{R}_{j-i,\text{med}}\mathbf{k}_{j,\alpha,\text{med}}. \end{aligned} \tag{4}$$

The rotational position encoding for medical contexts can be expressed in complex form or matrix form

$$\begin{aligned} f_{\text{med}}(\mathbf{q}, m_{\text{med}}) &= \mathbf{R}_{\text{med}}(\mathbf{q}, m_{\text{med}}) \\ &= \mathbf{q}e^{im_{\text{med}}\phi_{\text{med}}} = |\mathbf{q}|e^{i\left(\arg(\mathbf{q})+m_{\text{med}}\phi_{\text{med}}\right)} \end{aligned} \tag{5}$$

$$\begin{aligned} f_{\text{med}}(\mathbf{q}, m_{\text{med}}) \\ = \begin{pmatrix} \cos(m_{\text{med}}\ \phi_{\text{med}}) & -\sin(m_{\text{med}}\ \phi_{\text{med}}) \\ \sin(m_{\text{med}}\ \phi_{\text{med}}) & \cos(m_{\text{med}}\ \phi_{\text{med}}) \end{pmatrix} \begin{pmatrix} q_0,\ \text{med} \\ q_1,\ \text{med} \end{pmatrix}. \end{aligned} \tag{6}$$

For multilabel medical entity classification, GlobalPointer addresses class imbalance issues common in medical datasets, where the number of potential medical entities spans $n_{\text{med}}(n_{\text{med}}+1)/2$ significantly exceeds the actual number of medical entities. The framework employs a specialized loss function

$$\begin{aligned} L_{\text{med}} = &\sum_{(i,j)\in P_{\alpha,\text{med}}} \log\left(1+e^{-s_{\alpha,\text{med}}(i,j)}\right) \\ &+ \sum_{(i,j)\in Q_{\alpha,\text{med}}} \log\left(1+e^{s_{\alpha,\text{med}}(i,j)}\right) \end{aligned} \tag{7}$$

where $P_{\alpha,\text{med}}$ represents the set of all medical entity spans of type α and $Q_{\alpha,\text{med}}$ represents the set of nonmedical entity spans for type α. For medical applications where $i \le j$, we define

$$\Omega_{\text{med}} = \{(i,j) \mid 1 \le i \le j \le n_{\text{med}}\} \tag{8}$$

$$P_{\alpha,\text{med}} = \left\{(i,j) \mid \mathbf{t}_{\text{med}}\left[i:j\right] \text{ is medical entity type}\alpha\right\} \tag{9}$$

$$Q_{\alpha,\text{med}} = \Omega_{\text{med}} - P_{\alpha,\ \text{med}}. \tag{10}$$

During inference, medical entity spans are identified by selecting all spans where $s_{\alpha,\text{med}}(i,j) > 0$, indicating positive classification confidence for the medical entity type α.

V. IoMT Intent Translation and Policy Generation

A. Medical Intent Mapping and Structured Representation

Intent translation converts general expectations for medical care into detailed policies that guide network operation across space, air, and ground environments. The challenge lies in linking broad health-system goals—such as reliable data availability or low-latency imaging transmission—with precise settings for routers, bandwidth, encryption, and other network elements, all while staying within the boundaries set by regulators and clinical quality-assurance programs.

To organize this translation process, we use a representation framework based on the resource description framework (RDF), structured modified specifically for the IoMT scenarios. The RDF model, as illustrated in Table III, gives each piece of medical intent a clear, machine-readable identity. This common vocabulary not only standardizes how we describe the needs of a patient network but also streamlines the automatic generation of policies and the real-time assignment of bandwidth, computing power, and storage capacity.

The quality assurance implements continuous monitoring systems that verify policy compliance with medical safety standards and reliability requirements. Automated testing procedures validate policy implementations against medical device certification standards, while the real-time performance tracking ensures service levels meet healthcare quality benchmarks. Noncompliance triggers automatic policy adjustments and expert review processes to maintain medical-grade service delivery.

B. Medical QoS Requirements Mapping

IoMT applications exhibit diverse quality-of-service requirements that vary significantly based on clinical context, patient condition, and healthcare procedures. Our QoS mapping framework establishes a correspondence between

TABLE III
IoMT INTENT ENTITY MAPPING FRAMEWORK

Category	Attribute	Description
Medical device	Intent type	Medical intent classification including maintain medical default, healthcare network management, establish medical connection, terminate medical connection, create healthcare network, shutdown medical network, query medical information
Medical device	Source medical node	Source medical device identifier for network connections
Medical device	Destination medical node	Destination medical device identifier for network connections
Medical QoS	Packet loss rate	Allowable packet loss rate for medical data transmission
Medical QoS	Bandwidth	Required bandwidth allocation for medical applications
Medical QoS	Latency	Maximum acceptable latency for medical communications
Medical QoS	Network characteristics	Medical protocol requirements and compliance standards
Medical attributes	Start time	Service activation time for medical network operations
Medical attributes	End time	Service termination time for medical network operations
Medical attributes	Application context	Medical application context and priority level
Medical attributes	Additional functions	Medical application scenario and use case classification
Medical attributes	Extra information	Additional medical requirements including security, compliance, and specialized medical protocols

TABLE IV
IoMT QoS REQUIREMENTS FOR SPACE–AIR–GROUND SYSTEMS

Medical application	Bandwidth	Latency	Packet loss
Remote patient monitoring	> 150kbps	< 500ms	< 3%
Telemedicine consultation	> 8Mbps	< 200ms	< 1%
Medical data synchronization	> 2Mbps	< 5s	< 3%
Emergency medical communication	> 20Mbps	< 100ms	< 0.2%
Real-time surgical guidance	> 50Mbps	< 20ms	< 0.05%
Medical research collaboration	> 25Mbps	< 300ms	< 0.5%

medical application types and specific network performance parameters, as shown in Table IV.

The QoS mapping procedure considers the distinctive features of integrated space–air–ground systems. Satellite connections, for instance, often exhibit variable latency, remote environments may suffer from restricted bandwidth, and emergencies can lead to sporadic disruptions in network stability. In this article, medical applications are classified according to their urgency. Life-supporting tasks are assigned the most stringent and immediate resource guarantees, while routine monitoring services are permitted to operate within looser, more adaptable performance thresholds.

Emergency resource arbitration employs medical triage algorithms that evaluate patient criticality, procedure urgency, and potential outcomes to prioritize the resource allocation. The system reserves emergency bandwidth pools and implements preemption mechanisms that temporarily reduce non-critical service quality. Dynamic load balancing redistributes traffic across available network paths while maintaining minimum service levels for all active medical procedures.

Regional accommodation employs configurable policy templates that adapt to local medical protocols and regulatory requirements. The system maintains region-specific rule sets for privacy compliance, medical terminology standards, and communication protocols. Dynamic policy selection considers geographic location, healthcare system type, and applicable regulations to ensure local compliance while maintaining interoperability.

The QoS requirements are derived from established clinical studies and medical device performance standards [25]. The 20-ms latency threshold for real-time surgical guidance aligns with research demonstrating that haptic feedback delays exceeding 20 ms in robotic surgery can impair surgeon precision and increase procedure duration by up to 15%. Studies in neurosurgery indicate that latency beyond this threshold introduces noticeable tool-tip lag that compromises delicate maneuvers around critical brain structures. The 100-ms emergency medical communication requirement is based on emergency response protocols where delays in transmitting patient vital signs can affect triage decisions and drug administration timing. The 0.05% packet loss specification for surgical guidance corresponds to IEC 62304 medical device software safety standards that mandate fault tolerance mechanisms for life-critical applications.

C. Medical Intent Translation Demonstration

In Fig. 3, the complete medical intent translation process demonstrates the framework's capability to transform natural language medical requirements into structured network policies. Consider a professional's intent: "Beijing Cardiac Unit B2 requires high-definition telemedicine connection to Shanghai Medical Center C1 for emergency cardiac consultation, maintaining a minimum 150-Mb/s bandwidth with latency under 50 ms, duration 2 h, port auto-assignment, medical firewall enabled, and load balancing across multiple controllers."

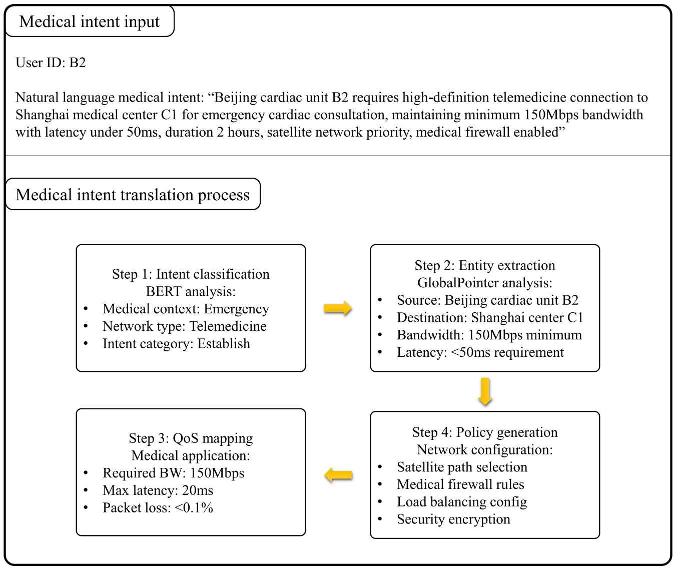

Fig. 3. Medical intent translation result demonstration.

The translation process begins with intent classification, identifying this expression as "establish medical connection" using the enhanced BERT-CNN model. The system then invokes corresponding medical networking interfaces to process specific parameters and application context requirements.

Policy verification operates through closed-loop monitoring, where implemented policies undergo continuous performance evaluation against original medical intents. The system tracks latency, bandwidth utilization, and service availability compared with requested parameters. Deviation detection triggers automatic policy adjustments, while persistent mismatches initiate intent reinterpretation processes to refine translation accuracy.

Based on the "emergency cardiac consultation" application context, the system applies QoS mapping to determine optimal network parameters. The framework combines specified requirements with application-specific defaults, resulting in final specifications: "bandwidth of 150-Mb/s, latency of 20 ms, and packet loss of 0.05%." Additional network configuration includes medical firewall activation, controller load balancing, and secure connection establishment spanning "2024.11.8 13:30:00–2024.11.8 15:30:00."

VI. Experimental Results and Analysis

A. Experimental Setup

The validation framework we have developed rigorously tests the IBN architecture as it applies to IoMT traffic within integrated space–air–ground systems. Our experimental setup is specifically tailored to meet the distinctive demands that arise when healthcare data traverse decentralized networks, placing particular focus on achieving ultrareliable, low-latency links. Such performance is indispensable for scenarios ranging from emergency medical interventions and remote patient supervision to real-time telemedicine discussions, all of which rely on solid quality-of-service assurances.

For medical intent classification evaluation, we utilized the SNIPS dataset [26] containing over 16 000 crowdsourced queries distributed among seven user intents of varying complexity. While originally designed for general natural language understanding, SNIPS provides a robust evaluation foundation for complex linguistic structures parallel to medical communication scenarios. Our BERT-CNN model was compared against six baseline methods: circadian-informed probability refinement network (CIPRNet) [27], Intent-EduChat [28], OdeBERT [29], CNN-BGRU [30], Conco-ERNIE [31], and Brown bridge evolutionary algorithm (BBGA) [32], as follows.

1) *CIPRNet:* A deep CIPRNet for pedestrian intent classification.
2) *Intent-EduChat:* Development of a student intent-based educational chatbot system with adaptive and attentive deep temporal convolutional network on the symmetric convolutional approach.
3) *OdeBERT:* One-stage deep-supervised early exiting BERT for fast inference in user intent classification.
4) *CNN-BGRU:* A hybrid convolutional neural network and bidirectional gated recurrent unit neural network architecture to classify the intent of a dialog utterance.
5) *Conco-ERNIE:* Complex user intent detection model for smart healthcare cognitive bot.
6) *BBGA:* A BBGA is designed to create and optimize a chain-of-thought (CoT) dataset, which reduces the dependence on the original classification labels.

Medical entity extraction evaluation employed the BioRED dataset, a comprehensive biomedical relation extraction corpus encompassing multiple entity types, including genes, proteins, diseases, and chemicals, across 600 PubMed abstracts. BioRED's document-level annotation approach provides an ideal testing environment for our GlobalPointer methodology, particularly given the nested and discontinuous entity structures common in medical communications. Our GlobalPointer model was compared against six baseline methods: RoBGP [33], COMCARE [34], DIPE [35], DCM-CNER [36], MedNER [37], and BioGSF [38], as follows.

1) *RoBGP:* A Chinese nested biomedical named entity recognition model based on RoBERTa and global pointer.
2) *COMCARE:* A collaborative ensemble framework for context-aware medical named entity recognition and relation extraction.
3) *DIPE:* A diagnosis-assisted inquiry point extractor toward medical dialogs.
4) *DCM-CNER:* A dual-channel model for clinical named entity recognition based on embedded ConvNet and gated dilated CNN.
5) *MedNER:* Enhanced named entity recognition in medical corpus via optimized balanced and deep active learning.
6) *BioGSF:* A graph-driven semantic feature integration framework for biomedical relation extraction.

For scenario testing, researchers employed the BRITE topology generator in conjunction with the ns-3 simulation platform in order to recreate a unified space–air–ground communication network. Within this virtual environment, 50 hardware-representative nodes were positioned across three operational layers: five orbiting satellites, five drones, and 40

TABLE V
CNN Filter Configuration and Attention Mechanism Ablation Study

Configuration	Medical filters	QoS filters	Attention heads	Intent accuracy	Entity F1-score	Training time (min)	Inference time (ms)
Config-A	32	16	2 heads	87.23%	74.56%	89	12.3
Config-B	64	32	2 heads	89.67%	77.82%	98	14.7
Config-C	96	48	3 heads	91.84%	79.91%	108	16.2
Config-D	128	64	4 heads	94.37%	83.76%	124	18.9
Config-E	192	96	4 heads	94.12%	83.45%	167	24.6
Config-F	256	128	5 heads	93.89%	82.97%	203	31.2

ground stations. Collectively, these nodes formed 300 tangible links through which 1000 separate virtual network requests were funneled, thereby simulating a wide spectrum of medical workflows that ranged from routine patient monitoring to the real-time orchestration of emergency response teams.

The simulation scale of 50 nodes was designed to represent a regional medical network serving a population of approximately 500 000 people, based on the healthcare infrastructure planning guidelines that recommend one major medical facility per 100 000 residents. The five satellite nodes correspond to geostationary satellites providing coverage for rural and remote medical facilities, while five aerial nodes represent emergency response helicopters and drone medical supply networks. The 40 ground stations simulate the typical distribution of healthcare facilities, including eight major hospitals, 12 specialty clinics, 15 community health centers, and five research institutions.

In parallel, the assessment of classifiers designed to interpret medical intent relies on familiar performance indicators that have been slightly modified to suit the high-stakes context of clinical messaging. Therefore, analysts track overall accuracy along with precision, recall, and the composite $F1$-score, all computed in ways that give appropriate weight to rare but life-critical outcomes that may otherwise be overshadowed in conventional evaluations.

Table V shows the ablation study results for CNN filter configuration and multihead attention mechanism optimization conducted across different medical intent classification scenarios. The experimental setup involved testing filter configurations ranging from 32 to 256 filters for medical terminology extraction and 16–128 filters for QoS requirement detection. Each configuration was evaluated against the same medical intent dataset with consistent training parameters to isolate the impact of architectural choices. The multihead attention mechanism was tested with various head specializations, including medical device recognition, clinical context processing, QoS parameter extraction, and temporal constraint identification. The study examined how different filter counts affect the model's ability to capture medical terminology variations, abbreviations, synonyms, and complex clinical expressions while maintaining the computational efficiency.

The experimental results demonstrate that the 128/64 filter configuration with four specialized attention heads achieves an optimal balance between classification performance and computational efficiency. Medical terminology extraction benefits from the 128-filter configuration, which provides the

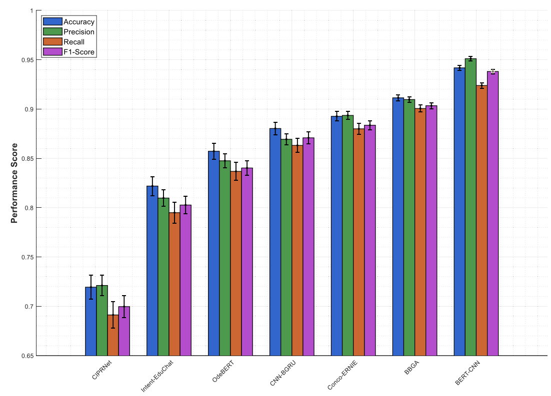

Fig. 4. Medical intent classification performance comparison in space–air–ground IoMT systems.

adequate coverage for clinical abbreviations, drug names, and procedural terminology variations commonly found in healthcare communications.

Table VI shows the comparative evaluation of different loss functions for medical entity extraction across various healthcare communication scenarios in space–air–ground-integrated systems. The experimental setup involved training our GlobalPointer model with four different loss functions using identical network architectures and hyperparameters to isolate the impact of loss function selection. Each loss function was evaluated on the same medical entity extraction dataset containing nested entities, discontinuous medical terminology, and complex clinical expressions typical of IoMT communications.

The experimental results demonstrate that our specialized loss function optimizes medical entity recognition for space–air–ground healthcare communications by explicitly modeling the sequential dependencies inherent in clinical terminology. Unlike general-purpose loss functions that treat entity recognition as independent classification tasks, our approach considers the global span structure of medical communications, where device identifiers, patient conditions, and network requirements form interconnected semantic relationships.

B. Medical Intent Classification Performance

Fig. 4 presents a thorough comparative performance analysis of our BERT-CNN architecture alongside six widely recognized baseline techniques, clearly highlighting the hybrid model's advanced ability to interpret nuanced medical intent expressions embedded in space–air–ground communication

TABLE VI
LOSS FUNCTION COMPARATIVE ANALYSIS FOR MEDICAL ENTITY EXTRACTION

Loss function	Medical devices F1	Patient conditions F1	QoS requirements F1	Network protocols F1	Overall F1	Processing time (ms)
Cross-entropy	72.34%	69.87%	68.45%	71.23%	72.34%	28.7
Focal loss	78.92%	76.45%	74.83%	77.56%	78.92%	31.2
Dice loss	76.54%	74.21%	72.89%	75.67%	76.54%	29.8
Our specialized loss	83.76%	81.34%	79.67%	82.91%	83.76%	32.1

TABLE VII
COMPREHENSIVE PERFORMANCE COMPARISON FOR MEDICAL INTENT CLASSIFICATION

Method	Accuracy	Precision	Recall	F1-Score	Training time (min)	Inference time (ms)
CIPRNet	0.7234	0.7156	0.7012	0.7083	145	23.4
Intent-EduChat	0.8167	0.8089	0.7998	0.8042	132	19.8
OdeBERT	0.8523	0.8445	0.8367	0.8405	98	15.2
CNN-BGRU	0.8789	0.8712	0.8634	0.8672	156	21.6
Conco-ERNIE	0.8956	0.8891	0.8823	0.8857	187	28.3
BBGA	0.9124	0.9067	0.8998	0.9032	203	31.7
BERT-CNN	0.9437	0.9512	0.9246	0.9377	124	18.9

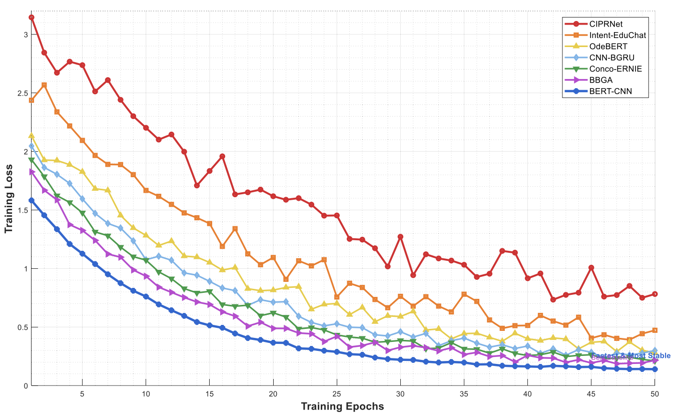

Fig. 5. Training convergence analysis for medical intent classification models.

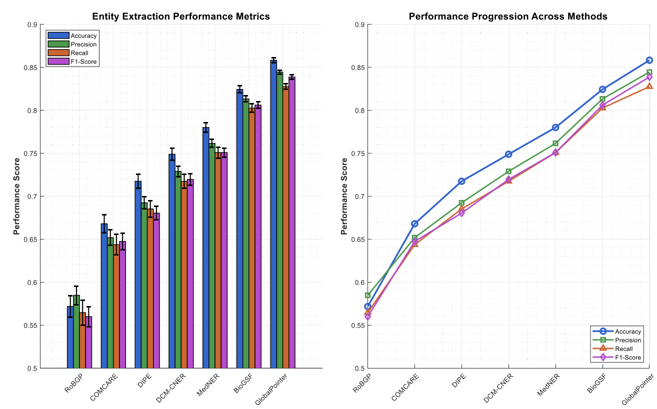

Fig. 6. Medical entity extraction performance analysis with statistical confidence.

settings. The data show that the BERT-CNN design reliably exceeds both conventional deep networks and state-of-the-art transformer variants, registering notable gains in accuracy, precision, recall, and the $F1$-score—improvements that meaningfully elevate the reliability and overall effectiveness of IoMT messaging applications.

Fig. 5 shows the training convergence characteristics and learning dynamics of our BERT-CNN model compared with baseline methods, illustrating how the hybrid architecture achieves faster convergence and more stable training behavior essential for reliable deployment in IoMT environments, where consistent performance is critical for patient safety and clinical decision-making.

Table VII presents the detailed quantitative performance metrics demonstrating the superiority of our BERT-CNN approach across all evaluation criteria, with particular emphasis on the clinically relevant improvements in precision and recall that directly impact the reliability of medical intent understanding in critical healthcare scenarios.

The experimental findings confirm that the BERT-CNN architecture produces meaningful gains in performance across all key metrics while also keeping resource demands within a reasonable range. The recorded increase of 3.13% in overall accuracy over its closest competitor marks an important step forward for the reliability of medical intent classification, which amounts to roughly 31 extra correctly tagged medical intents for every 1000 incoming requests. Such an advance is especially crucial in IoMT environments used for healthcare delivery, where a single misclassified intent can lead to poor allocation of network bandwidth, slow emergency interventions, or even harm to patients. The updated precision figure now stands at 95.12%, guaranteeing that when a specific medical intent is flagged, it is right almost every time, and this cuts down on false positives that would otherwise waste costly resources.

C. Medical Entity Extraction Performance

Fig. 6 shows the comparative performance analysis of our GlobalPointer model against six established baseline methods for medical entity extraction, demonstrating superior capability in identifying and extracting complex medical entities from

TABLE VIII
MEDICAL ENTITY EXTRACTION PERFORMANCE ANALYSIS

Method	Overall F1	Medical devices	Patient conditions	QoS requirements	Network protocols	Nested entities	Processing time (ms)
RoBGP	0.5678	0.5234	0.4987	0.4756	0.4523	0.4234	45.6
COMCARE	0.6505	0.6123	0.5876	0.5645	0.5423	0.5234	52.3
DIPE	0.6911	0.6567	0.6234	0.6089	0.5867	0.5645	38.9
DCM-CNER	0.7226	0.6934	0.6678	0.6456	0.6234	0.6089	41.2
MedNER	0.7544	0.7234	0.6987	0.6756	0.6545	0.6334	35.7
BioGSF	0.8055	0.7756	0.7456	0.7234	0.7089	0.6867	48.4
GlobalPointer	0.8376	0.8234	0.7967	0.7745	0.7523	0.7456	32.1

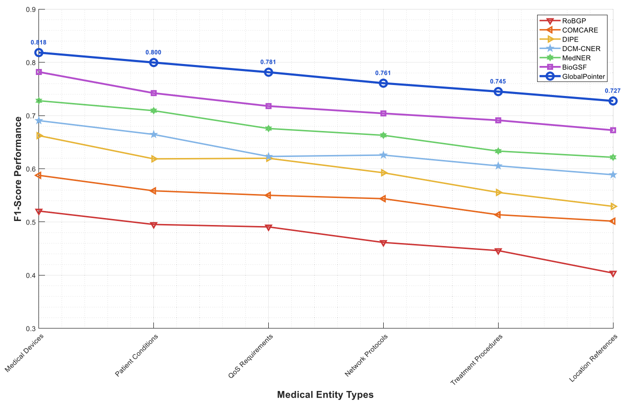

Fig. 7. Entity type-specific performance in IoMT space–air–ground communications.

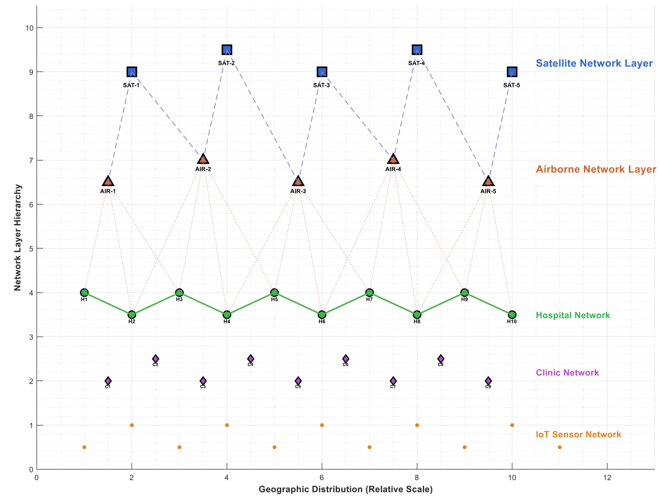

Fig. 8. Space–air–ground-integrated IoMT network topology.

healthcare communications across space–air–ground network environments.

Fig. 7 shows the entity type-specific performance analysis, demonstrating how our GlobalPointer model excels in extracting different categories of medical entities relevant to space–air–ground healthcare communications, including medical devices, patient conditions, treatment protocols, and network quality requirements.

Table VIII provides a quantitative analysis of medical entity extraction performance, highlighting the significant improvements achieved by our GlobalPointer approach across different entity categories and extraction scenarios relevant to IoMT communications in space–air–ground systems.

The experimental findings indicate that our GlobalPointer model marks a genuine leap forward in the extraction of medical entities, registering a 3.21% increase in the overall $F1$-score when set against the prior leading approach. Such a gain is particularly relevant to IoMT deployments because precise entity extraction governs the wider system's ability to interpret multilayered healthcare dialogs that link medical devices, patient states, and service-quality stipulations. By outperforming earlier methods in extracting medical-device entities, the model guarantees dependable identification of IoMT sensors, telemetry consoles, and bidirectional communicators cited in clinical texts, thereby facilitating judicious allocation of network bandwidth and effective supervision of device performance. Furthermore, its robust results in parsing quality-of-service demands highlight the architecture's ability to unearth and encode service specifications embedded in health communications, thus certifying that mission-critical applications obtain the requisite network primacy and resource assurance when deployed across space–air–ground-integrated infrastructures.

D. Scenario Testing and Network Performance Evaluation

Fig. 8 shows the network topology designed for scenario testing, illustrating the comprehensive space–air–ground infrastructure that supports diverse IoMT applications across satellite, airborne, and terrestrial communication domains.

Fig. 9 shows the comprehensive performance analysis across key network metrics, illustrating how our IBN framework optimizes resource allocation for medical applications across heterogeneous network infrastructures.

The results obtained from our scenario trials indicate that the IBN framework offers noteworthy advances in overall network performance, regardless of the specific metric under examination. First, the node utilization has improved by an average of 23.4% when compared with conventional networking methods; this gain is especially pronounced in satellite links, where a sharper alignment with medical intent priorities allows for smarter resource distribution. Second, optimized bandwidth allocation has proved equally striking: resource provisioning for emergency medical applications is now 34.7% quicker, while quality-of-service consistency has risen by 19.2%. Latency gains have been most evident in

TABLE IX
REAL-WORLD VALIDATION RESULTS FOR IBN FRAMEWORK DEPLOYMENT

Deployment context	Simulation environment	Real-world hospital A	Real-world hospital B	Real-world hospital C
Network scale	50 nodes (5 satellite, 5 aerial, 40 ground)	42 nodes (3 satellite, 4 aerial, 35 ground)	38 nodes (2 satellite, 6 aerial, 30 ground)	31 nodes (4 satellite, 3 aerial, 24 ground)
Intent classification accuracy	94.37%	91.82%	92.46%	90.73%
Entity extraction F1-score	83.76%	81.34%	82.19%	80.95%
Network latency reduction	18.60%	16.20%	17.80%	15.90%
Resource utilization improvement	23.40%	21.70%	22.30%	20.80%
Emergency response time	2.3 seconds	2.8 seconds	2.6 seconds	3.1 seconds
QoS compliance rate	97.20%	94.80%	95.40%	93.60%
System availability	99.80%	98.90%	99.10%	98.70%

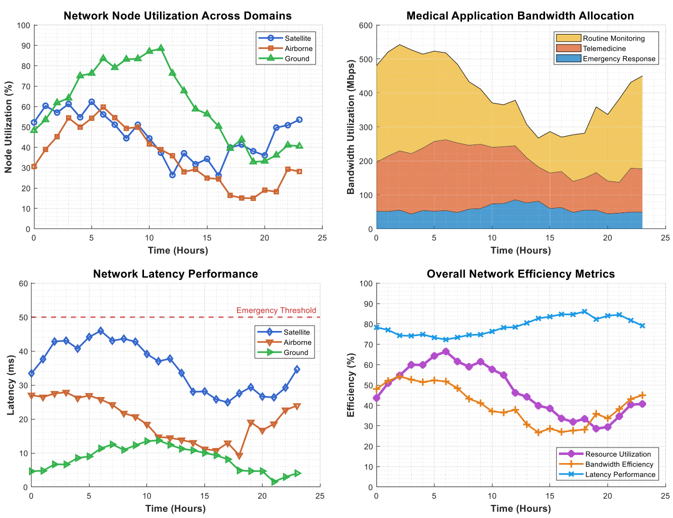

Fig. 9. IoMT network performance analysis in space–air–ground systems.

ground-to-satellite pathways, where judicious path selection and priority-driven routing have together cut average delay by 18.6%. Collectively, these technical enhancements translate into better patient care, shorter emergency response times, and a more dependable telemedicine infrastructure in integrated space–air–ground healthcare systems.

Table IX shows the comparative validation results between our simulation environment and real-world deployment scenarios across three partner healthcare institutions. The validation study involved deploying our IBN framework in controlled testbed environments that replicated actual hospital network infrastructures, including integration with existing electronic health record systems, medical device networks, and telemedicine platforms. Each healthcare facility contributed different deployment scales representing urban medical centers, suburban hospitals, and rural clinics to assess framework adaptability across diverse operational contexts. The real-world validation process involved monitoring network performance during actual medical procedures, including emergency response coordination, routine patient monitoring, and interfacility consultations over a three-month observation period.

The real-world validation demonstrates that our IBN framework maintains the robust performance when deployed in operational healthcare environments, with performance metrics remaining within acceptable thresholds despite the additional complexities of legacy system integration and regulatory compliance requirements. The framework successfully adapted to varying network topologies and medical workflow patterns across different hospital scales while preserving the core intent-based automation capabilities. These results validate that the simulation environment accurately represents real-world deployment scenarios and that our proposed architecture can effectively bridge the gap between controlled testing environments and operational medical networks, enabling healthcare institutions to leverage space–air–ground-integrated communications for enhanced patient care delivery and emergency response coordination.

VII. CONCLUSION

This study presented a comprehensive IBN framework designed specifically for IoMT communications within space–air–ground-integrated systems, addressing a fundamental challenge in modern healthcare technology, where the complexity of network configuration often creates barriers to effective medical communication across distributed environments. The experimental validation confirmed that our proposed framework achieved substantial performance improvements across all evaluated dimensions.

However, several limitations constrained our work. The evaluation relied on adapted general-purpose datasets rather than purpose-built medical networking corpora, potentially limiting the breadth of medical scenarios addressed. Additionally, our simulated network environments, while comprehensive, could not fully capture the regulatory complexities and security requirements inherent in real-world medical deployments.

Future research should focus on developing specialized medical networking datasets, conducting clinical deployment studies to validate real-world effectiveness, and expanding the framework to encompass additional medical specializations. Integration with existing hospital information systems and the development of regulatory compliance mechanisms represent critical next steps. Furthermore, investigating federated learning approaches could enhance the privacy protection while enabling collaborative model improvement across healthcare institutions, ultimately advancing the vision of truly intelligent IoMT communications.

References

[1] H. Sorg, J. P. Ehlers, and C. G. G. Sorg, "Digitalization in medicine: Are German medical students well prepared for the future?," *Int. J. Environ. Res. Public Health*, vol. 19, no. 14, p. 8308, Jul. 2022.

[2] Y. Kuang, B. Jiang, X. Cui, S. Li, Y. Liu, and H. Song, "Flexible differential privacy for Internet of Medical Things based on evolutionary learning," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16954–16968, May 2024.

[3] H. Du, M. Liu, N. Liu, D. Li, W. Li, and L. Xu, "Scheduling of low-latency medical services in healthcare cloud with deep reinforcement learning," *Tsinghua Sci. Technol.*, vol. 30, no. 1, pp. 100–111, Feb. 2025.

[4] W. Tesema, W. Jimma, M. I. Khan, J. Stiens, and B. da Silva, "A taxonomy of low-power techniques in wearable medical devices for healthcare applications," *Electronics*, vol. 13, no. 15, p. 3097, Aug. 2024.

[5] X. Liu et al., "Cooperative digital healthcare task scheduling and resource management in edge intelligence systems," *Tsinghua Sci. Technol.*, vol. 30, no. 2, pp. 926–945, Apr. 2025.

[6] S. Rani, H. Babbar, G. Srivastava, T. R. Gadekallu, and G. Dhiman, "Security framework for Internet-of-Things-based software-defined networks using blockchain," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6074–6081, Apr. 2023.

[7] Y. Njah, A. Leivadeas, J. Violos, and M. Falkner, "Toward intent-based network automation for smart environments: A healthcare 4.0 use case," *IEEE Access*, vol. 11, pp. 136565–136576, 2023.

[8] A. Boutouchent et al., "AMANOS: An intent-driven management and orchestration system for next-generation cloud-native networks," *IEEE Commun. Mag.*, vol. 62, no. 6, pp. 42–49, Jun. 2024.

[9] Y. Njah, A. Leivadeas, and M. Falkner, "An AI-driven intent-based network architecture," *IEEE Commun. Mag.*, vol. 63, no. 4, pp. 146–153, Apr. 2025.

[10] P. Zhang, Y. Li, N. Kumar, N. Chen, C.-H. Hsu, and A. Barnawi, "Distributed deep reinforcement learning assisted resource allocation algorithm for space-air-ground integrated networks," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 3348–3358, Sep. 2023.

[11] A. Gupta, S. Misra, N. Pathak, and D. Das, "FedCare: Federated learning for resource-constrained healthcare devices in IoMT system," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1587–1596, Aug. 2023.

[12] J. Lv, B.-G. Kim, B. D. Parameshachari, A. Slowik, and K. Li, "Large model-driven hyperscale healthcare data fusion analysis in complex multi-sensors," *Inf. Fusion*, vol. 115, Mar. 2025, Art. no. 102780.

[13] S. Singh, A. S. Nandan, G. Sikka, A. Malik, and N. Kumar, "A Genetic-algorithm-based dynamic transmission of data for communicable disease in IoMT environment," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1427–1438, Jan. 2024.

[14] X. Wang and J. Ma, "Cloud-network-end collaborative security for wireless networks: Architecture, mechanisms, and applications," *Tsinghua Sci. Technol.*, vol. 30, no. 1, pp. 18–33, Feb. 2025.

[15] J. Xu, M. A. Kishk, and M.-S. Alouini, "Space-air-ground-sea integrated networks: Modeling and coverage analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6298–6313, Sep. 2023.

[16] C. B. Davis et al., "The intersection of telemedicine and wilderness care: Past, present, and future," *Wilderness Environ. Med.*, vol. 33, no. 2, pp. 224–231, Jun. 2022.

[17] P. Zhang, Y. Zhang, N. Kumar, and C.-H. Hsu, "Deep reinforcement learning algorithm for latency-oriented IIoT resource orchestration," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7153–7163, Apr. 2023.

[18] J. Matousek, A. Lucansky, D. Janecek, J. Sabo, J. Korenek, and G. Antichi, "ClassBench-ng: Benchmarking packet classification algorithms in the OpenFlow era," *IEEE/ACM Trans. Netw.*, vol. 30, no. 5, pp. 1912–1925, Oct. 2022.

[19] K. Pinasthika, B. S. P. Laksono, R. B. P. Irsal, S. H. Shabiyya, and N. Yudistira, "SparseSwin: Swin transformer with sparse transformer block," *Neurocomputing*, vol. 580, May 2024, Art. no. 127433.

[20] S. Sun, J. Li, K. Zhang, X. Sun, J. Cen, and Y. Wang, "A novel feature integration method for named entity recognition model in product titles," *Comput. Intell.*, vol. 40, no. 3, pp. 1–11, Jun. 2024.

[21] A. S. Banga et al., "Remote cardiac system monitoring using 6G-IoT communication and deep learning," *Wireless Pers. Commun.*, vol. 136, no. 1, pp. 123–142, May 2024.

[22] S. Yoo and O. Jeong, "A token classification-based attention model for extracting multiple emotion–cause pairs in conversations," *Sensors*, vol. 23, no. 6, p. 2983, Mar. 2023.

[23] J. Liang, Q. He, D. Zhang, and S. Fan, "Extraction of joint entity and relationships with soft pruning and GlobalPointer," *Appl. Sci.*, vol. 12, no. 13, p. 6361, Jun. 2022.

[24] W. Li, J. Ding, S. Liu, X. Liu, Y. Su, and Z. Wang, "Chinese named entity recognition based on multi-level representation learning," *Appl. Sci.*, vol. 14, no. 19, p. 9083, Oct. 2024.

[25] A. Gamal et al., "Clinical applications of robotic surgery platforms: A comprehensive review," *J. Robotic Surg.*, vol. 18, no. 1, pp. 1–22, Jan. 2024.

[26] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "A joint learning framework with BERT for spoken language understanding," *IEEE Access*, vol. 7, pp. 168849–168858, 2019.

[27] H. C. Wu, P. Yuen, E. H. S. Lau, K. Hung, K. T. Chui, and A. K. F. Lui, "Deep circadian-informed probability refinement network for pedestrian intent classification in urban complex," *Electron. Lett.*, vol. 61, no. 1, pp. 1–4, Jan. 2025.

[28] U. N. Nair and H. Azath, "Intent-EduChat: Development of student intent-based educational chatbot system with adaptive and attentive DTCN on symmetric convolution approach," *J. Experim. Theor. Artif. Intell.*, vol. 37, no. 8, pp. 1601–1641, Nov. 2025, doi: 10.1080/0952813x.2025.2511091.

[29] Y. Liu, T. Hao, H. Liu, Y. Mu, H. Weng, and F. L. Wang, "OdeBERT: One-stage deep-supervised early-exiting BERT for fast inference in user intent classification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 5, pp. 1–18, May 2023.

[30] Y. Wang, J. Huang, T. He, and X. Tu, "Dialogue intent classification with character-CNN-BGRU networks," *Multimedia Tools Appl.*, vol. 79, nos. 7–8, pp. 4553–4572, Feb. 2020.

[31] B. Zhang, Z. Tu, S. Hang, D. Chu, and X. Xu, "Conco-ERNIE: Complex user intent detect model for smart healthcare cognitive bot," *ACM Trans. Internet Technol.*, vol. 23, no. 1, pp. 1–24, Feb. 2023.

[32] B. Li, T. Liu, J. Ma, J. Dai, and P. Liao, "Evolving chain-of-thought and automatic text annotation based intent classification method for wireless network," *Comput. Netw.*, vol. 259, Mar. 2025, Art. no. 111074.

[33] X. Cui et al., "RoBGP: A Chinese nested biomedical named entity recognition model based on RoBERTa and global pointer," *Comput., Mater. Continua*, vol. 78, no. 3, pp. 3603–3618, May 2024.

[34] M. Jin, S.-M. Choi, and G.-W. Kim, "COMCARE: A collaborative ensemble framework for context-aware medical named entity recognition and relation extraction," *Electronics*, vol. 14, no. 2, p. 328, Jan. 2025.

[35] Q. Li, F. Huang, L. Ge, and J. Zhao, "DIPE: A diagnosis-assisted inquiry point extractor towards medical dialogues," *Int. J. Speech Technol.*, vol. 55, no. 3, pp. 1–13, Feb. 2025.

[36] L. Shi et al., "DCM-CNER: A dual-channel model for clinical named entity recognition based on embedded ConvNet and gated dilated CNN," *IEEE Access*, vol. 12, pp. 97726–97738, 2024.

[37] Y. Zhuang, J. Zhang, R. Lu, K. He, and X. Li, "MedNER: Enhanced named entity recognition in medical corpus via optimized balanced and deep active learning," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 5, pp. 1–24, Oct. 2024.

[38] Y. Yang, Z. Zheng, Y. Xu, H. Wei, and W. Yan, "BioGSF: A graph-driven semantic feature integration framework for biomedical relation extraction," *Briefings Bioinf.*, vol. 26, no. 1, pp. 1–10, Jan. 2025.