

Deep Neural Networks for Real-Time Medical Streaming in 6G Cross-Modal Semantic Communication Systems

Xiaohong Lv, Jing Wang, Shalli Rani, Ayush Dogra, Adam Slowik, Daohua Pan, Huamao Jiang, and Yanhong Feng*

Abstract: In the world of Six-Generation 6G network, real-time medical streaming plays an important part in providing fast and accurate services. A new method known as cross-modal semantic communication helps in sending meaningful information to different types of data. This paper presents a deep neural network based system that uses 6G cross-modal semantic communication systems to handle medical streaming in real time. The proposed system includes a semantic encoder, a semantic decoder, and a method capable of measuring similarity meaning across various data types. The semantic encoder extracts important features like text, sound, and pictures from different types of medical data, and combines them to create an integrated information. After this, the semantic decoder redesigns the data as per the required format. Using Siamese and pseudo-Siamese networks, the cross-modal semantic similarity evaluation technique compares the meaning of the original and redesigned data across different types of data, resulting in improved encoding and decoding processes. Experimental results show that the proposed framework excels in semantic similarity and real-time performance compared to traditional communication systems. This deep neural networks based encoding and decoding framework enables efficient and effective real-time medical streaming in 6G cross-modal semantic communication systems.

Key words: Six-Generation (6G); cross-modal semantic communication; deep neural networks; real-time medical streaming data

1 Introduction

The social networking has entered our life with the

coming of Sixth Generation (6G) communication system, going farther from human communication to connect all kinds of devices and services together^[1–3].

- Xiaohong Lv and Yanhong Feng are with Department of Imaging, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China. E-mail: lvxh@jzmu.edu.cn; fengyh@jzmu.edu.cn.
- Jing Wang is with School of Computer Science and Engineering, Southeast University, Nanjing 210018, China. E-mail: wangjing91@seu.edu.cn.
- Shalli Rani and Ayush Dogra are with Institute of Engineering and Technology, Chitkara University, Punjab 140401, India. E-mail: shalli.rani@chitkara.edu.in; ayush.dogra@chitkara.edu.in.
- Adam Slowik is with Department of Electronics and Computer Science, Koszalin University of Technology, Koszalin 75-453, Poland. E-mail: adam.slowik@tu.koszalin.pl.
- Daohua Pan is with College of Computer Science, Harbin Institute of Technology, Harbin 150006, China. E-mail: pandaohua@hit.edu.cn.
- Huamao Jiang is with Multi-modal Data Fusion and Precision Medicine Laboratory, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China. E-mail: jianghm@jzmu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 2024-07-04; revised: 2024-09-10; accepted: 2024-11-08

With regards to non-cooperative scenarios, intelligent cognitive communication is an important part of semantic communication. By harnessing the power of artificial intelligence, cognitive communication aims to automatically identify and adapt to various communication parameters, such as modulation schemes, coding styles, and channel conditions^[4, 5]. The rapid advancements in deep learning techniques have transformed intelligent cognitive communication from a theoretical concept to a feasible engineering solution. Deep learning based intelligent cognitive communication seeks to establish a sophisticated mapping between signals and parameters, encompassing a wide range of research areas^[6–9].

However, in real-world wireless communication channels, the signals undergo many unfavorable effects which can degrade severely the functionalities and robustness of cognitive systems. Multiple paths propagation is a common problem in wireless environments, the signal transmitted from a transmitter to the receiver arrives at receiver through more than one path by reflection, diffraction, and scattering^[10, 11]. These different signal paths lead to copies of the signal that arrive at the receiver with differences in amplitude, phase, and time delays, resulting in a fading and distortion. Multipath propagation creates a time delay that spreads the received signal in the time domain, causing symbol boundaries to be misaligned and produce inter-symbol interference. One of the challenges in wireless communication systems is phase noise, which is produced by non-ideal circuit elements acting in operation mode exercise^[12]. This phase noise can cause an unwanted frequency shift and thus degrade the quality of our signal.

In recent years, deep learning techniques have become a potential way to advance intelligent wireless communications^[13, 14]. Deep learning models have the ability to automatically learn and extract meaningful feature representations from raw signal data, in contrast to traditional methods which mostly rely on human constructed features and domain-specific knowledge. Owing to this capability, deep learning models can perform better and generalize more broadly across a variety of application scenarios, as it enables the models to identify complex patterns and relationships within the signals. Researchers can create intelligent wireless communication systems that adjust to changing channel conditions, maximise resource allocation, and improve overall quality of service by

utilising the power of deep neural networks. Further, the automated feature learning capability of deep learning models removes the necessity for laborious manual feature engineering, which is frequently time-consuming and highly skilled.

The 6G cross-modal semantic communication system faces numerous obstacles that are specific to real-time medical streaming. Text from electronic health records, voice from interactions between doctors and patients, and pictures from different medical scans are all examples of the multi-modal nature of medical streaming data^[15–17]. Important semantic information is carried by each modality and must be accurately and efficiently delivered across the communication system. However, real-time medical streaming may have strict criteria that standard communication systems, which are primarily built for raw data delivery, must meet. These requirements include high reliability to ensure accurate data delivery, low latency to enable rapid transmission of critical medical information, and semantic preservation to maintain the context and meaning of medical information across different modalities. The medical data come in various forms and the way we need to process and send plays an important role. In 6G communication, solving this problem is crucial so that real-time medical decision can be made efficiently and treatment can be done.

In the era of 6G, real-time medical communications face significant challenges due to the need for efficient and reliable transmission of multi-modal medical data while preserving semantic information. Traditional communication systems struggle to address these challenges effectively. The proposed framework aims to overcome these limitations by enabling intelligent, adaptive, and semantically-aware medical data transmission in 6G networks. By leveraging deep learning techniques and cross-modal semantic fusion, the proposed framework ensures the efficient and reliable transmission of multi-modal medical data, such as text, audio, and images, while preserving critical semantic information. The framework's ability to adapt to varying channel conditions and its integration with the 6G network architecture further enhance its performance. It can revolutionize real-time medical communications, enabling timely and accurate medical decision-making, improving patient outcomes, and paving the way for advanced telemedicine applications in the 6G era.

In this research paper, a method capable of encoding

and decoding real-time medical data in 6G cross-modal semantic communication systems using deep neural networks is introduced. The main contributions of this paper are stated below.

(1) A semantic encoder that can extract important features from medical data is developed. This data from various sources are then combined into a single representation. It is done using a Stack Long Short-Term Memory (SLSTM) network for semantic fusion. It uses a specific neural network for each type of data (transformer for text, Convolutional Neural Networks (CNNs) for images, and Gated Recurrent Units (GRUs) for sounds).

(2) Based on the obtained semantic representation and the user's requirement, a semantic decoder is created that can redesign the required data type. The decoder uses attention-based Generative Adversarial Networks (GANs) for image reconstruction, task-driven neural networks like transformer for text generation and Tacotron and WaveNet for speech synthesis.

(3) To compare meaning between the original and redesigned data, a cross-modal semantic similarity evaluation method using Siamese and pseudo-Siamese networks is introduced that reduces the loss of meaning by optimizing the encoding and decoding processes using backpropagation.

While previous works have applied neural networks to individual modalities, our framework introduces a unified architecture for the joint processing of multiple medical data modalities, develops a novel cross-modal attention mechanism for semantic feature fusion, and proposes an adaptive channel coding scheme that considers semantic content and channel conditions.

The rest of the paper is organized as follows. Section 2 is about related work. Section 3 presents the system model for real-time medical streaming data in 6G cross-modal semantic communication systems. Section 4 describes the proposed deep neural networks based encoding and decoding framework, including the semantic encoder, semantic decoder, and cross-modal semantic similarity evaluation mechanism. Section 5 provides experimental results and analysis. Finally, Section 6 concludes the paper.

2 Related Work

In recent years, there has been a growing interest in intelligent cognitive communication and deep learning

techniques for wireless communication systems. Deep learning techniques have been extensively studied for various aspects of wireless communications, including physical layer design, channel estimation, and signal processing. Schmidhammer et al.^[18] investigated the use of deep learning for device-free localization in wideband wireless networks. Their approach leverages multipath information to improve localization accuracy using deep neural networks. Semantic communication focuses on the transmission of meaningful information rather than raw data. Xie et al.^[19] proposed a semantic communication system utilizing deep learning techniques, enhancing system capacity while reducing semantic errors by focusing on the comprehension of sentences instead of conventional bit or symbol errors found in traditional communication methods.

Cross-modal communication involves the interaction and integration of multiple data modalities, such as text, speech, and images. Wang et al.^[20] introduced a novel category of highly efficient deep joint source-channel coding methodologies intended for facilitating end-to-end video transmission across wireless channels, designated as Deep Video Semantic Transmission (DVST). Han et al.^[21] proposed an innovative end-to-end Deep Learning (eeDL) centric transceiver, wherein semantic information is discerned and encoded from input speech spectrums at the transmitter, while the decoded semantic information yields corresponding transcriptions at the receiver. Xie et al.^[22] proposed a semantic communication framework grounded on Federated Learning (FL), termed Federated Learning based Semantic Communication (FLSC), tailored for multitasking distributed image transmission encompassing Internet of Things (IoTs) devices. Yao et al.^[23] introduced a semi-neural framework characterized by an iterative architecture, denoted as Iterative Semantic Joint Source-Channel Coding (IS-JSCC). However, they have limitations in handling multi-modal medical data effectively. While some approaches focus on specific modalities, others cannot preserve semantic information accurately across different modalities. Additionally, their adaptability to varying channel conditions is limited, which is crucial for real-time medical communications. These approaches also face challenges seamlessly integrating with the 6G network architecture. Additionally, Jiang et al.^[24] proposed a

large AI model based multi-modal semantic communication framework. Jiang et al.^[25] proposed a large AI model-based semantic communication framework specifically designed for image data, applying the segment anything model-based knowledge base that can split the original image into different semantic segments by universal semantic knowledge.

Semantic communication and deep learning have shown promise in various applications in the medical domain. Kang et al.^[26] designed an energy-efficient task-oriented semantic communication framework with a triple-based scene graph for image information. Dai and Wang^[27] proposed an end-to-end framework that emulates the process of healthcare, including a deep recognition module and action evaluation module. These studies demonstrate the potential of semantic communication in improving medical data transmission and interpretation.

Despite the significant contributions of these studies to the advancement of intelligent cognitive communication and deep learning techniques in wireless networks, there is still a need for further research on cross-modal semantic communication in the context of 6G systems.

3 System Model

This study proposes a medical neural network based encoding and decoding framework for real-time semantic communication (namely MEDNETS) in 6G systems. We consider a healthcare scenario in which medical streaming data is from various sources (such as patients, electronic health records (EHRs), medical devices, and healthcare professionals), which need to be efficiently communicated in real time to support critical medical decision-making and treatment

delivery. The proposed system model, as depicted in Fig. 1, consists of a transmitter, a wireless channel, and a receiver, each incorporating deep neural networks for cross-modal semantic encoding, decoding, and evaluation.

On the transmitter side, multi-modal medical streaming data (S^Q) includes text (S^T), audio (S^A), images (S^I), and video (S^V), where $Q \in \{T, A, I, V\}$ denotes the modality index, and T , A , I , and V represent the text, audio, image, and video modalities, respectively. Each modality carries unique information that is important for fully understanding and analyzing a patient's medical condition. For example, medical reports, doctors' prescriptions, and EHRs can be regarded as text data; doctor-patient conversations, medical dictations, and verbal instructions constitute audio data; and X-ray, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) scans are image data. Video data may include surgical videos, endoscopic examinations, and telemedicine consultations. Multi-modal medical streaming data are efficiently sent using a semantic encoder at the transmitter side. In order to create a single semantic representation, the semantic encoder seeks to extract the key semantic components from each modality and carries out cross-modal semantic fusion. Let f^Q denote the semantic feature extraction function for modality Q , and g^Q denote the cross-modal semantic fusion function. The extracted semantic features for each modality can be represented as

$$\mathbf{z}^T = f^T(S^T; \theta^T) \quad (1)$$

$$\mathbf{z}^A = f^A(S^A; \theta^A) \quad (2)$$

$$\mathbf{z}^I = f^I(S^I; \theta^I) \quad (3)$$

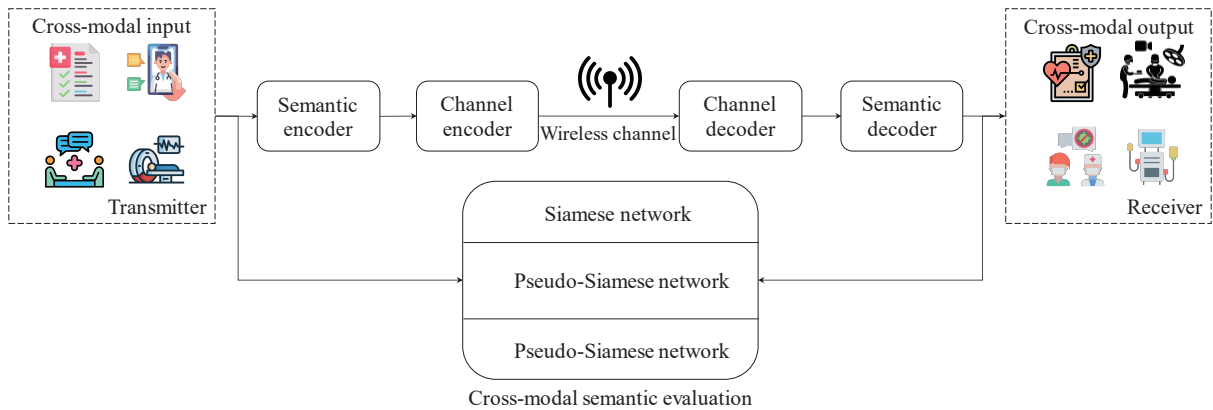


Fig. 1 Overall model.

$$\mathbf{Z}^V = \mathbf{f}^V(S^V; \theta^V) \quad (4)$$

In Eq. (1), $\mathbf{Z}^T$ denotes the extracted text semantic features, $\mathbf{f}^T(\cdot)$ is the text semantic feature extraction function, S^T is the input text data, and θ^T denotes the learnable parameters of the text feature extraction function; in Eq. (2), $\mathbf{Z}^A$ denotes the extracted audio semantic features, $\mathbf{f}^A(\cdot)$ is the audio semantic feature extraction function, S^A is the input audio data, and θ^A denotes the learnable parameters of the audio feature extraction function; in Eq. (3), $\mathbf{Z}^I$ denotes the extracted image semantic features, $\mathbf{f}^I(\cdot)$ is the image semantic feature extraction function, S^I is the input image data, and θ^I denotes the learnable parameters of the image feature extraction function; and in Eq. (4), $\mathbf{Z}^V$ denotes the extracted video semantic features, $\mathbf{f}^V(\cdot)$ is the video semantic feature extraction function, S^V is the input video data, and θ^V denotes the learnable parameters of the video feature extraction function.

To enhance interoperability, MEDNETS incorporates a flexible data preprocessing module that can handle various input formats from different medical devices. This module utilizes adaptive parsing algorithms and standardized data transformation techniques to convert diverse medical data into a unified format suitable for semantic encoding. Additionally, MEDNETS employs a protocol abstraction layer that allows seamless integration with different communication protocols used in medical devices, ensuring compatibility across a wide range of healthcare systems.

To capture the inter-modal dependencies and correlations, the extracted semantic features from different modalities are fused using the cross-modal semantic fusion function $\mathbf{g}^Q$. The fused semantic representation can be expressed as

$$\mathbf{Z} = \mathbf{g}^Q(\mathbf{Z}^T, \mathbf{Z}^A, \mathbf{Z}^I, \mathbf{Z}^V; \phi^Q) \quad (5)$$

where $\mathbf{Z}$ represents the fused semantic representation, and ϕ^Q denotes the learnable parameters of the cross-modal semantic fusion function.

After that, a channel encoder processes the fused semantic representation to produce a signal that is compatible with the wireless channel. The goal of the channel encoder is to minimise the loss of semantic information while adjusting the semantic representation to the wireless channel's properties, such as fading, interference, and noise. The channel encoding process can be formulated as

$$\mathbf{x} = \mathbf{c}(\mathbf{Z}; \psi) \quad (6)$$

where $\mathbf{x}$ is the channel-encoded signal, $\mathbf{c}(\cdot)$ is the channel encoding function, and ψ is the learnable parameters of the channel encoder.

The wireless channel, which can be represented as a noisy and fading channel, is then used to transmit the encoded signal. The signal that is received at the receiving end might be stated as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \quad (7)$$

where $\mathbf{H}$ represents the channel matrix, capturing the effects of fading, interference, and other channel impairments, and $\mathbf{n}$ represents the Additive White Gaussian Noise (AWGN) with zero mean and variance σ_n^2 .

To extract the transmitted semantic representation from the received signal, a channel decoder is used on the receiving end. One way to formulate the channel decoding procedure is

$$\hat{\mathbf{Z}} = \mathbf{c}^{-1}(\mathbf{y}; \psi') \quad (8)$$

where $\hat{\mathbf{Z}}$ is recovered semantic representation, $\mathbf{c}^{-1}(\cdot)$ is channel decoding function, and ψ' is the learnable parameters of the channel decoder.

Based on the user's real-time medical communication needs, a semantic decoder on the receiver's side reconstructs the required data type. The semantic decoder uses the recovered semantic representation $\hat{\mathbf{Z}}$ as input to rebuild the multi-modal medical streaming data. The formula for semantic decoding process is

$$\hat{\mathbf{s}}^Q = \mathbf{f}^{Q,-1}(\hat{\mathbf{Z}}; \theta^{Q'}) \quad (9)$$

where $\hat{\mathbf{s}}^Q$ denotes the reconstructed multi-modal medical streaming data for modality Q , $\mathbf{f}^{Q,-1}(\cdot)$ is the inverse function of the semantic feature extraction, and $\theta^{Q'}$ is the learnable parameter of the semantic decoder.

For the accuracy and reliability of the redesigned multi-modal medical streaming data, a mechanism known as a cross-modal semantic evaluation is also included. It uses Siamese and pseudo-Siamese networks to compare the original and redesigned data meaning. While comparing similar type of data (text to text/audio to audio), the Siamese network is used and for different types (text to image/audio to video), the pseudo-Siamese network is used. The Siamese network consists of two matching subnetworks, $\mathbf{h}^Q(\cdot)$ with shared parameters $\varphi^{Q[28, 29]}$. The equation to compute the semantic similarity between the original and redesigned data of the same modality is

$$\mathcal{L}_{\text{Siamese}}(S^Q, \hat{S}^Q) = \frac{h^Q(S^Q; \varphi^Q) \cdot h^Q(\hat{S}^Q; \varphi^Q)}{|h^Q(S^Q; \varphi^Q)| \cdot |h^Q(\hat{S}^Q; \varphi^Q)|} \quad (10)$$

where $\mathcal{L}_{\text{Siamese}}(S^Q, \hat{S}^Q)$ represents the Siamese loss for modality Q , and $|\cdot|$ denotes the Euclidean norm.

Let $Q_1, Q_2 \in \{T, A, I, V\}$ with $Q_1 \neq Q_2$ denote two different modalities. The pseudo-Siamese network consists of two different subnetworks, $h^{Q_1}(\cdot)$ and $h^{Q_2}(\cdot)$, with parameters φ^{Q_1} and φ^{Q_2} , respectively^[30]. The semantic similarity between the original data of modality Q_1 and the redesigned data of modality Q_2 is calculated using Eq. (11),

$$\begin{aligned} \mathcal{L}_{\text{Pseudo-Siamese}}(S^{Q_1}, \hat{S}^{Q_2}) \\ = \frac{h^{Q_1}(S^{Q_1}; \varphi^{Q_1}) \cdot h^{Q_2}(\hat{S}^{Q_2}; \varphi^{Q_2})}{|h^{Q_1}(S^{Q_1}; \varphi^{Q_1})| \cdot |h^{Q_2}(\hat{S}^{Q_2}; \varphi^{Q_2})|} \end{aligned} \quad (11)$$

where $\mathcal{L}_{\text{Pseudo-Siamese}}(S^{Q_1}, \hat{S}^{Q_2})$ represents the pseudo-Siamese loss between modalities Q_1 and Q_2 .

The overall objective of the proposed system model is to minimize the semantic loss between the original and redesigned multi-modal medical data being streamed keeping in mind the real-time communication requirements. The semantic loss is a weighted combination of the Siamese and pseudo-Siamese losses across all modalities,

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{Siamese}} + \lambda_2 \cdot \mathcal{L}_{\text{Pseudo-Siamese}} \quad (12)$$

where $\mathcal{L}_{\text{total}}$ represents the total loss function that the model aims to minimize during training, λ_1 represents a weighting factor for the Siamese network loss, and λ_2 represents a weighting factor for the pseudo-Siamese network loss. $\mathcal{L}_{\text{Siamese}} = \sum_Q \mathcal{L}_{\text{Siamese}}(S^Q, \hat{S}^Q)$ and $\mathcal{L}_{\text{Pseudo-Siamese}} = \sum_{Q_1 \neq Q_2} \mathcal{L}_{\text{Pseudo-Siamese}}(S^{Q_1}, \hat{S}^{Q_2})$ are the total Siamese and pseudo-Siamese losses aggregated across all modalities.

To meet the real-time communication requirements, the proposed system model aims to minimize the end-to-end latency while maintaining acceptable semantic similarity. Let T_{total} denote the total end-to-end latency. It can be calculated as

$$T_{\text{total}} = T_{\text{enc}} + T_{\text{tx}} + T_{\text{dec}} \quad (13)$$

The encoding latency T_{enc} is dependent on the computational complexity of the semantic encoder and the channel encoder, the decoding latency T_{dec} is dependent on computational complexity of the channel decoder, and the semantic decoder. The transmission delay T_{tx} depends on wireless channel settings like Signal-to-Noise Ratio (SNR) and bandwidth.

The trade-off between the semantic similarity and end-to-end delay is reduced in our proposed system by combining semantic-channel coding with Adaptive Modulation and Coding (AMC). Semantic-channel coding builds semantic decoder, channel encoder, and semantic encoder together, considering both the semantic information and the features of the wireless channel. AMC maximise the transmission efficiency while preserving the quality of service by regulating the modulation and coding which are channel dependent.

Additionally, in order to facilitate cooperative learning and knowledge exchange across various healthcare organisations while maintaining data privacy, the suggested system model integrates FL approaches. FL allows the distributed training of deep neural networks across multiple healthcare entities without the need for centralized data collection. Using its own data, each healthcare organization trains its model; only the model updates are shared with a central server. The revised global model is distributed back to the healthcare entities by the central server, aggregating the revisions to the model. This method allows for collaborative learning to enhance the overall effectiveness of the cross-modal semantic communication system while maintaining data privacy, because the raw medical streaming data are never shared with other healthcare organizations.

The overall workflow of the MEDNETS framework can be summarized in the following seven steps.

(1) The system collects multi-modal medical streaming data and performs semantic encoding at the transmitter side.

(2) The system performs cross-modal semantic fusion to obtain a unified semantic representation.

(3) The system performs channel encoding and transmits the encoded signal over the wireless channel.

(4) The system performs channel decoding and semantic decoding at the receiver side to reconstruct the desired modality.

(5) The system performs cross-modal semantic evaluation using Siamese and pseudo-Siamese networks to assess the semantic similarity between the original and reconstructed medical streaming data.

(6) The system optimizes the trade-off between semantic similarity and end-to-end delay through joint semantic-channel coding and adaptive modulation and coding techniques.

(7) The system applies federated learning strategies

that protect patient privacy while facilitating cooperative learning and knowledge exchange across several healthcare organizations.

To enable quick medical decision-making and treatment while keeping the meaning of the multi-modal medical streaming data intact, the proposed model aims to provide efficient and reliable real-time medical streaming in 6G networks.

The deep learning algorithms in the suggested framework can easily be integrated into a 6G network, allowing medical streaming data to be processed smartly and adaptively at the edge and cloud levels. This integration enables real-time analysis, decision-making, and feedback, which are crucial for time-sensitive medical applications. 6G networks^[31] also support huge Multiple-Input Multiple-Output (MIMO), which can significantly boost the capacity and reliability of wireless communication systems. The proposed framework can use these advanced multiple access methods to efficiently transport multi-modal medical streaming data over 6G networks. For example, Non-Orthogonal Multiple Access (NOMA) principles can be integrated into the semantic encoding and decoding processes, allowing multiple medical streaming data to be transmitted simultaneously over the same frequency. It is calculated as

$$r = \sqrt{P_1}h_1x_1 + \sqrt{P_2}h_2x_2 + n_0 \quad (14)$$

where r is the received signal, P_1 and P_2 are the power allocation coefficients for the medical data streams, h_1 and h_2 are the channel coefficients, x_1 and x_2 are the transmitted medical data symbols, and n_0 is the additive noise.

The proposed framework enhances the spatial diversity and multiplexing using extensive MIMO offered by 6G communication network. The framework improves the signal quality, increase capacity, and enhance reliability of transmitted data by installing numerous antennas at both the base station and medical devices. The channel capacity is calculated as

$$C = \log_2 \det \left(\mathbf{I}_{N_r} + \frac{\rho}{N_t} \mathbf{H} \mathbf{H}^H \right) \quad (15)$$

where C is the channel capacity, N_r and N_t are the number of receiving and transmitting antennas, respectively, ρ is the SNR, $\mathbf{I}_{N_r}$ is the $N_r \times N_r$ identity matrix, $\det(\cdot)$ denotes the matrix determinant, and $(\cdot)^H$ is the Hermitian (conjugate transpose) operator.

The proposed framework increases the capabilities of

6G to perform semantic encoding, decoding, and similarity evaluation, closer to medical devices and users. It enables real-time analysis and decision-making, reducing latency and bandwidth of the system. Also, while ensuring data privacy, the framework enables combined learning and knowledge sharing among various healthcare organizations.

The MEDNETS framework seamlessly integrates with the 6G network architecture to enable intelligent and adaptive medical data processing. By deploying the deep learning algorithms used in MEDNETS on edge computing nodes and cloud servers, the framework leverages the distributed computing capabilities of 6G networks. This allows for real-time analysis, decision-making, and feedback, critical for time-sensitive medical applications. The MEDNETS framework also takes advantage of the advanced capabilities of 6G networks, such as AI integration, to enhance the efficiency and reliability of medical data transmission. Incorporating advanced multiple access techniques, such as NOMA, enables the simultaneous transmission of multiple medical data streams over the same frequency resources, improving spectral efficiency. Furthermore, the MEDNETS framework exploits the massive MIMO capabilities of 6G networks to achieve high spatial diversity and multiplexing gains, resulting in improved signal quality, increased capacity, and higher reliability for medical data transmission. By leveraging these advanced features of 6G networks, the MEDNETS framework enables efficient, reliable, and semantically-aware transmission of multi-modal medical data in real-time, paving the way for transformative medical applications, such as telemedicine, remote patient monitoring, and personalized healthcare delivery.

Subsequently, we present a clear mathematical formulation of the cross-modal semantic communication problem and the optimization objectives of the MEDNETS framework. Let $\mathbf{X} = \{\mathbf{X}^T, \mathbf{X}^A, \mathbf{X}^I\}$ denote the input multi-modal medical data, where $\mathbf{X}^T$, $\mathbf{X}^A$, and $\mathbf{X}^I$ represent the text, audio, and image modalities, respectively. The channel conditions are represented by the channel matrix $\mathbf{H}$ and the noise vector $\mathbf{n}$. The reconstructed multi-modal medical data is denoted as $\hat{\mathbf{X}} = \{\hat{\mathbf{X}}^T, \hat{\mathbf{X}}^A, \hat{\mathbf{X}}^I\}$, where $\mathbf{X}^m$ and $\hat{\mathbf{X}}^m$ denote the original and reconstructed data of modality m , with $m \in \mathcal{M} = \{T, A, I\}$ denoting the modality index. Let $\mathcal{I}$ denote the index set of training samples.

The optimization objectives of the MEDNETS framework are as follows:

(1) Minimize the semantic loss between the original and reconstructed medical data:

$$\mathcal{L}_{\text{semantic}} = \sum_{m \in \mathcal{M}} \mathcal{D}(\mathbf{X}^m, \hat{\mathbf{X}}^m) \quad (16)$$

where $\mathcal{M} = \{T, A, I\}$ denotes the set of modalities and $\mathcal{D}(\cdot, \cdot)$ is a semantic distance measure.

(2) Maximize the cross-modal semantic similarity:

$$\mathcal{S}_{\text{cross-modal}} = \sum_{\substack{m_1, m_2 \in \mathcal{M} \\ m_1 \neq m_2}} \mathcal{C}(\mathbf{Z}^{m_1}, \mathbf{Z}^{m_2}) \quad (17)$$

where $\mathbf{Z}^{m_1}$ and $\mathbf{Z}^{m_2}$ is the semantic representations of modalities m_1 and m_2 , respectively, $m_1, m_2 \in \mathcal{M}$ with $m_1 \neq m_2$ are two different modality indices, and $\mathcal{C}(\cdot, \cdot)$ is a cross-modal similarity measure.

(3) Minimize the end-to-end latency:

$$\mathcal{L}_{\text{latency}} = t_{\text{encoding}} + t_{\text{transmission}} + t_{\text{decoding}} \quad (18)$$

where t_{encoding} , $t_{\text{transmission}}$, and t_{decoding} are the latencies of the encoding, transmission, and decoding processes, respectively.

The overall optimization problem can be formulated as

$$\min_{\theta} (\mathcal{L}_{\text{semantic}} - \lambda \mathcal{S}_{\text{cross-modal}} + \gamma \mathcal{L}_{\text{latency}}) \quad (19)$$

it subject to the channel constraints and the available bandwidth, where θ represents the learnable parameters of the MEDNETS framework, and λ and γ are the balancing coefficients. All three terms are normalized to be dimensionless before aggregation, and λ and γ are balancing coefficients that trade off semantic fidelity, cross-modal similarity, and latency.

As a result, the proposed system combines the advanced features of 6G networks for real-time medical streaming. By using 6G's smartness, the framework provides secure, reliable, and efficient transfer of multi-modal medical data. It paves the path for different medical applications like personalized healthcare, telemedicine, and remote patient monitoring, improving patient health and transforming the healthcare sector.

4 Cross-Modal Semantic Communication System Design

4.1 Cross-modal semantic encoding

The cross-modal semantic encoding integrates the real-time semantic information from different types of medical data. As shown in Fig. 2, cross-modal semantic fusion and modality-specific semantic extraction are the two main components of the encoding module.

To address the dynamic nature of medical knowledge, MEDNETS incorporates a continuous learning module that periodically updates its semantic representations based on new clinical guidelines and research findings. This module utilizes a combination of online learning algorithms and federated learning techniques to incrementally refine the model's understanding of medical concepts and their relationships. Additionally, a knowledge graph integration component allows MEDNETS to incorporate structured medical knowledge, enabling it to reason over evolving clinical guidelines and adapt its semantic processing accordingly.

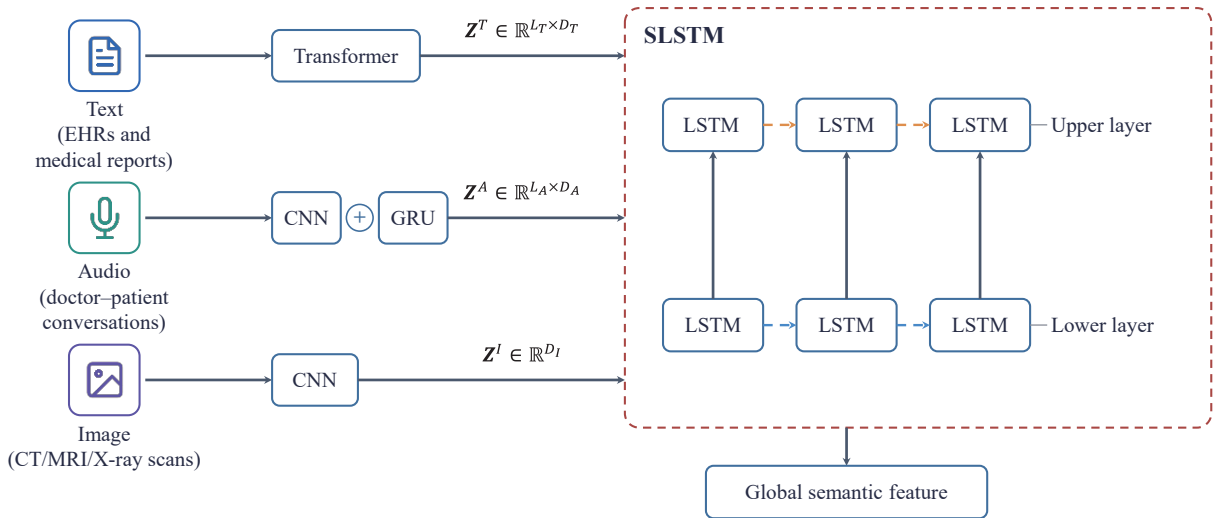


Fig. 2 Cross-modal semantic encoding module.

The suggested system considers three main modalities in the context of real-time medical streaming: text (S^T), audio (S^A), and images (S^I). Every modality has distinct semantic information that adds to our knowledge of a patient's overall health. We use deep neural networks to handle the unique features of each type of data to extract semantic features.

For textual medical streaming data, like EHRs and medical reports, we employ a transformer network^[32]. The transformer is pre-trained on large-scale medical corpora and fine-tuned on task-specific datasets to handle the domain-specific vocabulary and incorporate medical knowledge, allowing the network to learn contextualized representations that capture the intricacies of medical language. The self-attention mechanism in the transformer is particularly advantageous for encoding medical text, as it can capture long-range dependencies and handle variable-length sequences. The transformer can identify and encode important medical concepts, relationships, and semantics by attending to different parts of the input sequence. For example, it can effectively capture the association between a disease and its symptoms or the temporal progression of a patient's condition. These contextualized representations enable accurate cross-modal semantic communication in the medical domain, facilitating the integration of textual information with other modalities such as audio and images.

We use a combination of CNN and GRU to extract semantic features from audio medical data (like doctor-patient conversations and medical dictations). Initially, preprocessing techniques like spectral subtraction are applied to reduce noise and enhance the audio signal quality. Once pre-processed, the audio signal is transformed into a time-frequency representation, like a spectrogram, which captures both the temporal and spectral characteristics of the audio. The spectrogram is sent to CNN to detect local spectral patterns and extract higher-level features. Then, the output from the CNN is processed through a series of GRU layers to model temporal relationships and extract semantic features. The output of the GRU layers is a sequence of semantic feature vectors $\mathbf{Z}^T \in \mathbb{R}^{L_T \times D_T}$, where L_T is the length of the audio sequence and D_T is the dimensionality of the semantic feature space.

We extract semantic features from medical imaging data, such as CT, MRI, and X-ray scans, using a deep CNN architecture. CNNs have proven to be remarkably effective in various computer vision applications, such

as the analysis of medical images^[33]. The input image S^I is first preprocessed to normalize the pixel values and resize the image to a fixed resolution. The preprocessed image is then passed through a series of convolutional layers, which learn hierarchical features by applying convolutional filters at different scales and resolutions. The output of the convolutional layers is flattened and passed through fully connected layers to obtain a compact semantic feature representation $\mathbf{Z}^I \in \mathbb{R}^{D_I}$, where D_I is the dimensionality of the semantic feature space.

In the cross-modal semantic encoding module, the attention mechanism is incorporated into the CNN and GRU networks for audio data processing to enhance feature extraction and focus on the most relevant information. Self-attention layers are employed after each convolutional layer in the CNN, allowing the network to capture dependencies between different time-frequency regions in the spectrogram and enabling CNN to learn more discriminative and informative features. Similarly, the GRU layers are augmented with attention mechanisms to selectively focus on the most relevant temporal segments of the audio data. The attention weights are learned during training, enabling the GRU to prioritize the most informative parts of the audio sequence for semantic encoding. Integrating attention mechanisms in the CNN and GRU networks improves the extracted features' quality and the audio data's overall semantic representation, leading to more accurate and efficient cross-modal semantic communication in the MEDNETS framework.

Once the semantic features are extracted from each modality, the next step is to fuse these features to obtain a unified semantic representation. The cross-modal semantic fusion component aims to capture the interactions and dependencies among different modalities and generate a comprehensive semantic representation that encapsulates the relevant information for real-time medical decision-making.

We employ an SLSTM network, as seen in the right portion of Fig. 2, to accomplish efficient cross-modal semantic fusion. The LSTM network^[34], of which the SLSTM network is a version, has demonstrated excellent performance in modeling sequential data and capturing long-term dependencies. The SLSTM network processes input sequences hierarchically by stacking multiple LSTM layers on top of one another. MEDNETS incorporates a data reconciliation module

that employs probabilistic graphical models and Bayesian inference techniques to address potential conflicts or inconsistencies in multi-modal medical data. This module analyzes the semantic relationships and temporal dependencies across different data modalities to identify and resolve inconsistencies. When conflicts are detected, the system applies a weighted voting mechanism based on the reliability and relevance of each data source. In cases where automatic resolution is impossible, MEDNETS flags the inconsistencies for human expert review, ensuring the integrity of the processed medical information.

The extracted semantic features from each modality, $\mathbf{Z}^T$, $\mathbf{Z}^A$, and $\mathbf{Z}^I$, are first concatenated along the feature dimension to form a multi-modal semantic feature matrix $\mathbf{Z} \in \mathbb{R}^{L \times D}$, where L is the maximum sequence length among the modalities and $D = D_T + D_A + D_I$ is the total dimensionality of the concatenated semantic feature space. If the sequences have different lengths, zero padding is applied to ensure a consistent sequence length.

The multi-modal semantic feature matrix $\mathbf{Z}$ is then fed into the SLSTM network. At each time step t , the SLSTM network takes the semantic feature vector $\mathbf{z}_t \in \mathbb{R}^D$ as input and computes the hidden state $\mathbf{h}_t \in \mathbb{R}^H$ and the cell state $\mathbf{c}_t \in \mathbb{R}^H$, where H is the number of hidden units in each LSTM layer. The computation at each time step can be formulated as

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ii}\mathbf{z}_t + \mathbf{b}_{ii} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_{hi}) \quad (20)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{if}\mathbf{z}_t + \mathbf{b}_{if} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_{hf}) \quad (21)$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{ig}\mathbf{z}_t + \mathbf{b}_{ig} + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_{hg}) \quad (22)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{io}\mathbf{z}_t + \mathbf{b}_{io} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_{ho}) \quad (23)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (24)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (25)$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{g}_t$, and $\mathbf{o}_t$ are the input gate, forget gate, cell gate, and output gate, respectively. $\mathbf{W}_{ii}$, $\mathbf{W}_{if}$, $\mathbf{W}_{ig}$, $\mathbf{W}_{io}$, $\mathbf{W}_{hi}$, $\mathbf{W}_{hf}$, $\mathbf{W}_{hg}$, and $\mathbf{W}_{ho}$ are the weight matrices, and $\mathbf{b}_{ii}$, $\mathbf{b}_{if}$, $\mathbf{b}_{ig}$, $\mathbf{b}_{io}$, $\mathbf{b}_{hi}$, $\mathbf{b}_{hf}$, $\mathbf{b}_{hg}$, and $\mathbf{b}_{ho}$ are the bias vectors. $\mathbf{W}_{qh}$ and $\mathbf{b}_q$ are the output weight matrix and bias vector, respectively. $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and hyperbolic tangent activation functions, respectively, and $\odot$ denotes element-wise multiplication.

The SLSTM network processes the multi-modal semantic feature matrix $\mathbf{Z}$ sequentially, updating the

hidden state and cell state at each time step based on the current input and the previous states. The SLSTM network's layered architecture enables hierarchical semantic feature processing, capturing both high-level and low-level dependencies between the modalities.

The fused cross-modal semantic representation is represented by the final hidden state of the SLSTM network $\mathbf{h}_L$ which is obtained after processing the entire sequence. This gathers semantic data from all types of medical data and their interactions, providing a comprehensive understanding of the patient's medical condition.

For a specific requirement in real-time, an extra neural network layers are included that further processes this fused cross-modal semantic representation to refine and design it. This refined representation supports therapy recommendations, patient monitoring, and medical decision making. Therefore, this process allows for accurate and fast medical decision-making and treatment by understanding the health status of the patient.

4.2 Cross-modal semantic decoding

The cross-modal semantic decoding is another module that plays an important crucial role in real-time medical communication. It redesigns the required data type as per the need of the user and the received semantic representation. Its purpose is to generate a meaningful and understandable outputs that can help in assisting medical professionals in making quick and accurate decisions. This type of decoding consists of networks dedicated for each data type and a task-oriented LSTM decoder (Fig. 3).

The task-driven LSTM decoder takes the received semantic representation $\hat{\mathbf{Z}}$ as input. It generates a sequence of modality-specific semantic features $\hat{\mathbf{z}}_t^Q$ based on the user's task requirements. The LSTM decoder is designed to capture the temporal dependencies and context information within the semantic representation, enabling accurate reconstruction of the desired modality.

The LSTM decoder operates sequentially, processing the received semantic representation $\hat{\mathbf{Z}}$ step by step. At each time step t , the LSTM decoder takes the current semantic feature $\hat{\mathbf{z}}_t$ and the previous hidden state $\mathbf{h}_{t-1}$ as inputs and generates the current hidden state $\mathbf{h}_t$ and the output semantic feature $\hat{\mathbf{z}}_t^Q$,

$$\hat{\mathbf{z}}_t^Q = \mathbf{W}_{qh}\mathbf{h}_t + \mathbf{b}_q \quad (26)$$

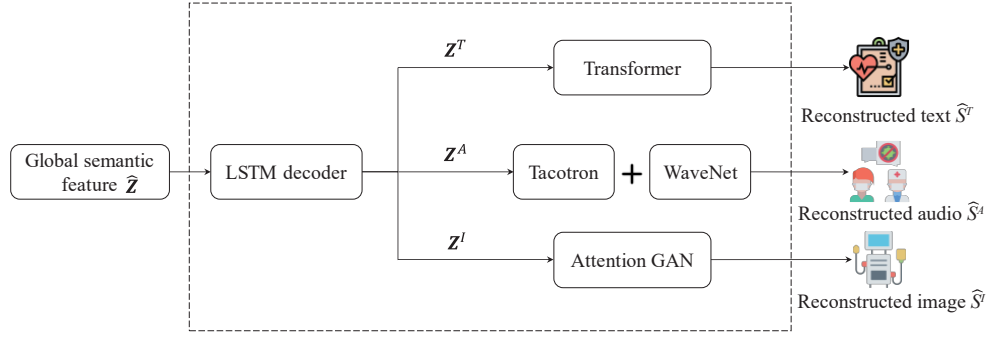


Fig. 3 Cross-modal semantic decoding model.

The LSTM decoder's task-driven design enables semantic information to be generated according to the user's data requirements. For example, in real-time medical communication, if the user requests a summary of text related data of the patient's, the LSTM decoder semantic features that are relevant to that data have been produced. This guarantees that the information generated is specific to the user's request.

However, the LSTM decoder will produce semantic characteristics unique to the picture modality if the user needs a visual representation of the patient's medical images.

To reconstruct the desired modality, the modality-specific semantic information $\hat{z}^Q$ produced by the LSTM decoder is input into the appropriate modality-specific reconstruction networks. Reconstruction networks is intended to produce meaningful and interpretable outputs for medical decision-making in real-time by mapping the semantic features back to the original data space.

An encoder based on transformer is used to reconstitute textual data. The transformer decoder creates the reconstructed text sequence $\hat{S}^T$ by using the modality-specific semantic characteristics $\hat{z}^T$ as input. The use of transformer decoder helps to effectively capture the long-distance relationships while simultaneously producing logical and situation-appropriate text. It is made up of many self-attention processes and many layers of feed-forward neural networks.

Medical specialists can use the resulting text sequence $\hat{S}^T$ to obtain essential information, such as therapy recommendations, medical reports, and patient histories. As long as to recreate the data related to audio an integration of WaveNet-based vocoder and Tacotron-based decoder has been made in the use.

The input $\hat{S}^A$ using the mel-spectrogram helps in recovering the audio using the WaveNet vocoder. Also, WaveNet vocoder learns the probability to guarantee high-quality audio synthesis. From the reconstructed audio, medical personnel can extract pertinent information, such as spoken instructions, physician-patient talks, and medical dictations. This audio enhances the efficiency and speed of medical decision-making by supporting telemedicine, distant consultations, and real-time patient monitoring. An attention-based generative adversarial network is used to recreate picture data. To generate realistic and semantically relevant medical images, this network includes a discriminator and a generator^[35].

The reconstructed image is denoted by $\hat{S}^I$, which is generated by the generator network using the modality-specific semantic features $\hat{z}^I$ as input.

The generator network uses an attention mechanism to focus on the important regions of the image during this process. By enabling the generator to gather the most pertinent and instructive portions of the image, this attention process improves the overall appeal of the rebuilt image.

The discriminator network, on the other hand, has been trained to distinguish between real images and fake images. It is the discriminator's responsibility to provide input to the generator, so that it can generate images that are more accurate and semantically consistent. The adversarial training process can be formulated as a minimax game between the generator G and the discriminator D ,

$$\min_G \max_D V(D, G) = E(S^I) \sim p_{\text{data}}(S^I) [\log D(S^I)] + E(\hat{z}^I) \sim p_{\hat{z}^I}(\hat{z}^I) [\log(1 - D(G(\hat{z}^I)))] \quad (27)$$

where $V(D, G)$ is the adversarial value function of the minimax game, $E(\cdot)$ denotes the expectation operator,

$p_{\text{data}}(S^I)$ is the distribution of real medical images, $p_{\hat{z}^I}(\hat{z}^I)$ is the distribution of the modality-specific semantic features input to the generator, and G and D denote the generator and discriminator networks, respectively. The generator aims to minimize the objective function, while the discriminator aims to maximize it.

Healthcare providers can get visual information on a patient's state, including anatomical structures, pathological anomalies, and therapy progress, from the reconstructed medical images $\hat{S}^I$. The reconstructed images can be used for remote patient monitoring, therapy planning, and diagnosis in real-time medical communication scenarios. The attention mechanism in GAN ensures that reconstructed images highlight the most important and informative regions, aiding accurate and timely medical decision-making.

Convolutional focus modules or individual attention layers can be utilised to build the system of attention in a GAN. By calculating weights for attention according to the spatial relationships within the image, self-attention layers help the network concentrate on important regions. Additionally, convolutional attention modules build attention maps—which highlight important areas of the image—by using convolutional processes. Using the attention weights, it creates during training, the generator knows to highlight informative visual regions. The discriminator network in the attention-grabbing GAN is designed taking these attention mappings into account as well. By merging attention maps with reconstructed images, the discriminator assesses the accuracy and uniformity of attention-guided reconstructions. Based on the reconstructions' level of realism and semantic coherence, it gives the generator feedback.

Adversarial loss and reconstructive loss are combined to train the attention GAN. Adversarial loss encourages the generator to create realistic images that trick the discriminator, whereas reconstruction loss guarantees that the reconstructed images closely resemble the initial pictures in terms of semantic content and pixel-level resemblance. The generator and discriminator are iteratively changed during training. The generator minimises both reconstructive loss and adversarial loss, while the discriminator seeks to maximise adversarial loss by separating genuine images from fraudulent ones. Until the produced images sufficiently realistic and have conceptual coherence with the source images, this iterative process is continued.

In the cross-modal semantic decoding module, the attention GAN is trained using a combination of adversarial loss and reconstruction loss functions. The generator network learns to reconstruct realistic and semantically meaningful images from the received semantic representations, while the discriminator network learns to distinguish between the reconstructed and real images. The attention mechanism is incorporated into the generator network to focus on the most relevant regions of the image during the reconstruction process. Attention weights are learned through convolutional attention modules or self-attention layers, guiding the generator to prioritize the salient features and spatial dependencies in the image. The adversarial training process optimizes the generator and discriminator networks alternately, minimizing the adversarial loss and reconstruction loss. This ensures that the reconstructed images are not only visually realistic but also semantically consistent with the original images. The attention mechanism enhances the visual fidelity and preserves the semantic content of the reconstructed images, resulting in improved quality and interpretability for medical image communication.

In real-time medical circumstances, medical practitioners can access aesthetically comprehensible and spatially pertinent pictures using the attention-based reconstruction of images network. These rebuilt images are used for diagnosis, treatment planning, and progress monitoring, among other uses. By making sure that important and instructive portions of the images are emphasised, the attention mechanism promotes precise and efficient decision-making. The cross-modal semantic decoding module, comprising task-driven LSTM decoder and modality-specific reconstruction networks, is the backbone of the immediate time medical communication system. Experts in medicine can easily understand and analyse the results produced by this module. Reconstructed modalities are guaranteed to be coherent, semantically consistent, and visually informative using reconstruction networks such as transformer-based decoder for text, attention GAN for images, and Tacotron-WaveNet for audio.

The cross-modal meaning decoding module provides medical practitioners access the necessary knowledge to make rapid choices in real-time clinical contexts through the application of deep instruction and attention approaches.

4.3 Cross-modal semantic similarity evaluation

In order to guarantee communication efficacy and dependability in real-time medical environments, it is critical to assess the degree to which artificial medical streaming data accurately reflects the original across a variety of data kinds. Siamese and pseudo-Siamese networks are used in the cross-modal semantic equivalent assessment approach to gauge how consistent and coherent the reconstructed material is in comparison to the original. Feedback to the cross-modal semantic encoding and decoding modules must be provided using this method. It aids in modifying network settings intended to improve system efficiency. The assessment approach guarantees the accuracy, significance, and relevance of the information provided for medical decision-making by limiting the semantic loss between the original and reproduced data.

The original and reconstructed content of the same modality are compared semantically. Siamese networks consist of exactly two identical subnetworks with identical features and designs. In order to learn a metric space where comparable inputs are mapped closer together and diverse inputs farther away, siamese networks are built.

In real-time medical communication, Siamese networks independently evaluate the semantic similarity between the reconstructed and original input for each modality. For instance, the semantic similarity between the original and reconstructed text data is compared using a Siamese network with two identical transformer-based subnetworks. Similarly, a Siamese network with two identical CNN-GRU-based subnetworks is utilised to analyse the semantic similarity between the original and reconstructed audio data.

The original and reconstructed data are fed into the Siamese networks, which project them into a common embedding space. The objective is to minimise the difference in the embedding space between the reconstructed and original data, thus demonstrating semantic solid equivalence. The Equation of the Siamese loss function for assessing intra-modal similarity is in Eq. (28),

$$\mathcal{L}_{\text{Siamese}}(S^Q, \hat{S}^Q) = \frac{1}{N} \sum_{i=1}^N \left| f(S_i^Q) - f(\hat{S}_i^Q) \right|_2^2 \quad (28)$$

where $f(\cdot)$ denotes the shared embedding function of the two identical subnetworks, S_i^Q and $\hat{S}_i^Q$ denote the

i -th original and reconstructed sample of modality Q , respectively, N is the number of samples, and $\|\cdot\|_2^2$ denotes the squared Euclidean (ℓ_2) distance.

Establishing that the new information accurately reflects the original data is essential for real-time medical communication systems. By training to minimise the Siamese loss function, which promotes the embeddings of the actual and recreated data to be near together in the embedding space, Siamese networks play an important role. The Siamese network receives both the original and reconstructed inputs during training, and the corresponding Siamese loss is used with backpropagation to adjust the network's parameters.

In medical situations, Siamese networks offer a quantifiable means of evaluating the semantic resemblance between actual and rebuilt data for each modality. These data aid in confirming the precision and dependability of the medical information supplied. Additionally, the evaluation of semantic similarity between the original data from one modality and the rebuilt data from another is done using pseudo-Siamese networks. A subnetwork based on transformer network for text and an attention GAN-based network for images, for instance, might be part of a pseudo-Siamese network that compares the semantic similarity between the original text and the reconstructed image data. They are then mapped into a common embedding space by these networks. The objective is to minimize the distance between the embeddings of the original and reconstructed data in this space, thereby demonstrating strong semantic similarity across modalities. The pseudo-Siamese loss is formulated as

$$\mathcal{L}_{\text{Pseudo-Siamese}}(S^{Q_1}, \hat{S}^{Q_2}) = \frac{1}{N} \sum_{i=1}^N \left| f_1(S_i^{Q_1}) - f_2(\hat{S}_i^{Q_2}) \right|_2^2 \quad (29)$$

where $f_1(\cdot)$ and $f_2(\cdot)$ denote the embedding functions learned by the two distinct subnetworks of the pseudo-Siamese network for modalities Q_1 and Q_2 , respectively.

In order to tightly align the encapsulation of the initial data from one modality with recreated data from another form of communication in an embedding space, pseudo-Siamese networks have been trained to minimise the pseudo-Siamese loss function. The networks use their own subnetworks to handle the original and rebuilt data, compute the pseudo-Siamese loss, and use backpropagation to modify the network

parameters during training. These networks offer a quantifiable means of evaluating the degree of similarity between the data reconstructed from one modality and the initial data from another.

An attention mechanism can be added to the pseudo-Siamese networks, namely to the subnetwork that handles reconstructed picture data, to improve inter-modal similarity evaluation even more. When the network compares the image with the initial data from an alternative modality, this attention mechanism enables the network to concentrate on important and informative portions of the image. It is possible to use individual attention levels or computational attention modules, which are comparable to the attention process utilised in the picture reconstruction network. Using attention weights acquired during training, these mechanisms direct the pseudo-Siamese network to prioritize important regions of the rebuilt image that are contributing most to the semantic connection with the initial data from another modality. Calculating the attention-guided pseudo-Siamese network yields

$$\begin{aligned} \mathcal{L}_{\text{Attention-Pseudo-Siamese}}(S^{Q_1}, \hat{S}^{Q_2}) \\ = \frac{1}{N} \sum_{i=1}^N \left| f_1(S_i^{Q_1}) - f_2(\hat{S}_i^{Q_2}, A_i) \right|_2^2 \end{aligned} \quad (30)$$

where $\mathcal{L}_{\text{Attention-Pseudo-Siamese}}$ denotes the attention-guided pseudo-Siamese loss, $f_2(\cdot, \cdot)$ is the two-argument extension of $f_2(\cdot)$ in Eq. (29), which additionally takes the attention weights A_i of the i -th sample as input and is learned by the attention-guided subnetwork of the pseudo-Siamese network.

The overall loss function for the cross-modal semantic similarity evaluation mechanism is

$$\begin{aligned} \mathcal{L}_{\text{Total}} = \lambda_1 \mathcal{L}_{\text{Siamese}} + \lambda_2 \mathcal{L}_{\text{Pseudo-Siamese}} \\ + \lambda_3 \mathcal{L}_{\text{Attention-Pseudo-Siamese}} \end{aligned} \quad (31)$$

where λ_3 is the weighting factors for attention-guided pseudo-Siamese loss.

To minimise the overall loss function during training, techniques, such as adaptive learning rate optimisation algorithms and stochastic gradient descent, have been used. By adjusting network parameters by backpropagation across Siamese and pseudo-Siamese networks, semantic similarity evaluation performance is enhanced. The real and reconstructed data from various modalities are mapped by these networks into a single embedding space, where dissimilar and similar data are positioned farther apart and closer together, respectively. The original and rebuilt medical

streaming data are evaluated for semantic consistency and coherence using this learned embedding space.

Quantitative evaluations of the degree of similarity between the original and reconstructed medical streaming data are provided by the cross-modal semantic similarity evaluation procedure, both within and between modalities. The quality, dependability, and efficacy of real-time medical communication systems are assessed using these parameters. In the cross-modal semantic similarity evaluation module, the pseudo-Siamese network employs modality-specific architectures to compare different modalities, such as text and image. For textual data, a transformer-based subnetwork is used to capture the contextual dependencies and semantic information. For image data, an attention GAN-based subnetwork is employed to learn the visual features and spatial relationships. These subnetworks are trained to project different modalities into a shared embedding space, where semantically similar data from different modalities are closer. The pseudo-Siamese network learns to capture the semantic relationships between modalities by minimizing the distance between the embeddings of semantically related text and image pairs. The attention mechanism in the image subnetwork helps focus on the image's most relevant regions for cross-modal comparison. It learns to assign higher weights to the salient regions, contributing more to the semantic similarity with the textual data and enabling more accurate and fine-grained cross-modal evaluation.

The results obtained from the performance evaluation help to optimise network settings and provide input to improve the semantic related to cross-modal process of encoding and decoding modules. The various adjustments can help to preserve the meaning of semantic and improve the reconstruction of data accuracy by reducing semantic loss. This review process is critical for ensuring the accuracy and consistency of medical data exchanged during real-time communication settings. It maintains the dependability and quality of communication networks by constantly evaluating how well the original and recreated data match. Similarly, evaluation results provide instant feedback to healthcare providers on the correctness and reliability of received medical information, allowing them to make informed decisions and take appropriate action^[36, 37].

MEDNETS includes a modular Application Programming Interface (API) layer that supports

standard healthcare data exchange formats, such as HL7 FHIR and DICOM, to facilitate seamless integration with existing healthcare information systems. This API layer enables bidirectional communication between MEDNETS and EHR systems, picture archiving, and communication systems, other healthcare IT infrastructures. By supporting these standards, MEDNETS can be easily incorporated into clinical workflows, enhancing its practical value in real-world healthcare settings.

Therefore, optimising real-time medical systems requires the cross-modal semantic similarity evaluation mechanism, which have been integrated the pseudo-Siamese networks with attention mechanisms for inter-modal similarity and Siamese networks for intra-modal similarity. By preserving semantic coherence and consistency between the original and reconstructed medical streaming data, it enables medical professionals to use the communicated data to make well-informed judgements.

5 Experiment and Results Analysis

5.1 Parameters setting

The tests related to simulation has been carried out using PyTorch 1.10.0 and Python 3.8 on a PC equipped with an NVIDIA GeForce RTX 3090 GPU. For multi-modal tasks, we utilise the MIMIC-III dataset for audio–text pairs, the VQA-Med dataset for image-question-answer triplets, and the MIMIC-CXR dataset for image–text pairings. The PhysioNet/CinC Challenge 2016—Heart Sound Recordings dataset is used for audio signal reconstruction. The OpenI dataset is used to assess how well the MEDNETS framework manages textual data and different medical imaging modalities. Stochastic Gradient Descent (SGD), which starts learning at 0.001 and is continuously optimised, is used to train the network. There are ten batches in total. The quantity of epochs (training iterations) of 50 has been used for the experiment.

Seven benchmarks are used to assess the proposed MEDNETS framework: DVST^[20], eeDL^[21], FLSC^[22], conventional, error-free, single-modal, and IS-JSCC^[23].

The traditional method uses Huffman encoding for text source encoding, Turbo coding for text channel encoding, Joint Photographic Experts Group (JPEG) for image encoding, Low-Density Parity-Check (LDPC) for image channel encoding, and Adaptive

Multi-Rate Wideband (AMR-WB) for speech encoding. The error-free method transmits noise-free speech, text, and image modal signals to express the limit value of the transmission signal as an upper or lower bound. The single-modal method is a traditional single-modal semantic communication transmission that does not realize multi-modal semantic-level fusion transmission.

The traditional method serves as a baseline, representing the conventional approach to medical data transmission. The error-free method provides an upper bound on the performance, assuming ideal channel conditions. The single-modal method highlights the limitations of transmitting medical data without considering cross-modal semantic information. DVST, eeDL, FLSC, and IS-JSCC are state-of-the-art methods in cross-modal semantic communication that have shown promising results in various applications. These methods are chosen for their ability to handle multi-modal data, their performance preserving semantic information, and their adaptation to varying channel conditions. By comparing MEDNETS with these benchmark methods, we aim to demonstrate the superiority of our framework in terms of its ability to efficiently and reliably transmit multi-modal medical data, while preserving semantic information in real-time communication scenarios.

5.2 Results and performance analysis

5.2.1 Semantic text transmission

The word-level text similarity results for different methods are presented in Fig. 4. The proposed MEDNETS framework achieves the highest similarity scores across all SNR levels, outperforming the other benchmarks. At an SNR of 18 dB, the MEDNETS framework achieves a word-level text similarity of 0.95, demonstrating its superiority in preserving semantic information during transmission.

Figure 5 presents the results of the sentence-level text similarity. The MEDNETS framework consistently achieves the highest similarity scores, indicating its effectiveness in capturing and preserving the semantic meaning at the sentence level. At an SNR of 18 dB, the MEDNETS framework achieves a sentence-level text similarity of 0.98, further validating its superior performance.

5.2.2 Semantic speech transmission

Similarity in speech S_{speech} has been calculated by the equation in Eq. (32),

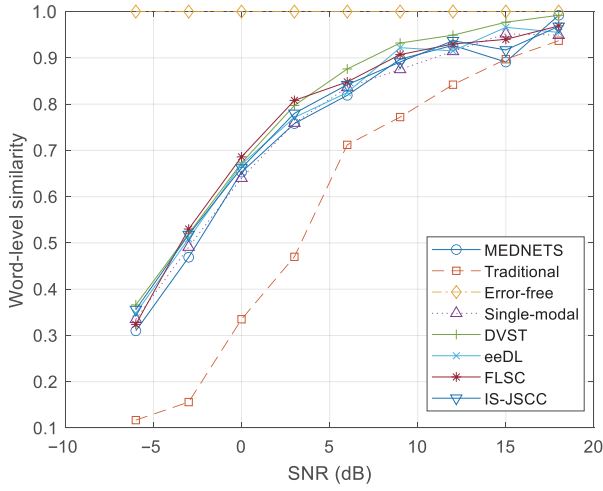


Fig. 4 Word-level text similarity.

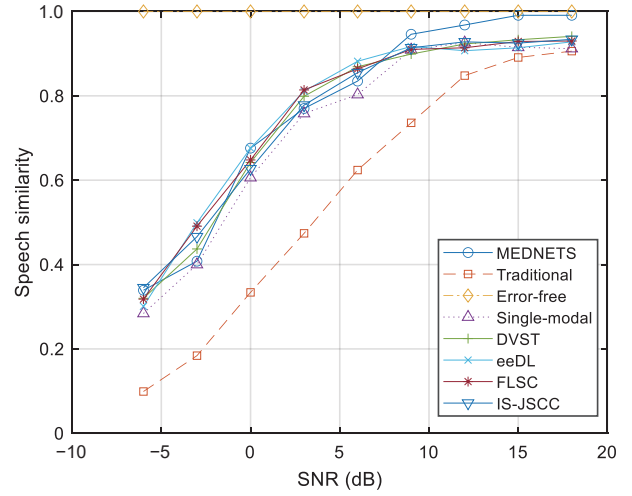


Fig. 6 Speech similarity.

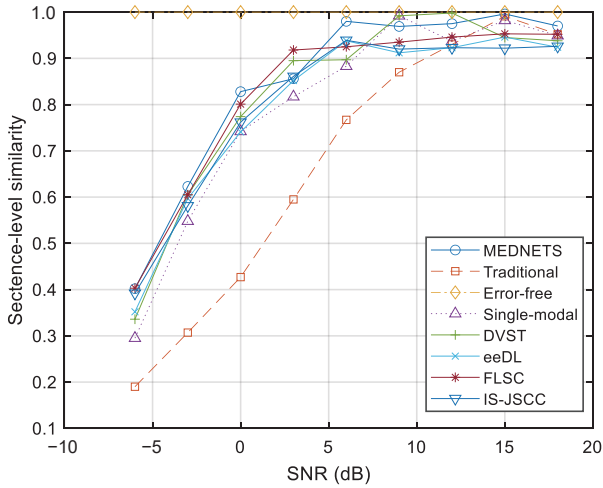


Fig. 5 Sentence-level text similarity.

$$S_{\text{speech}} = \frac{1}{N} \sum_{i=1}^N \frac{L_i - D_i}{L_i} \quad (32)$$

where L_i is the length of the i -th original speech signal, and D_i is the Levenshtein distance between the recognized text of the i -th original speech signal and the recognized text of the corresponding reconstructed speech signal.

The similarity in voice results from an acoustic model based speech recognition technique has been displayed in Fig. 6. The capacity of the MEDNETS system to retain semantic information in voice signals is demonstrated by the system achieving the greatest similarity scores across all SNR levels.

While the Perceptual Evaluation of Speech Quality (PESQ) score is calculated using Eq. (33),

$$\text{PESQ} = 4.5 - 0.1 \times \text{DMOS} \quad (33)$$

where DMOS is the degradation mean opinion score, which is estimated based on the perceptual difference between the original and reconstructed speech signals.

The PESQ scores for the various approaches are shown in Fig. 7. The capacity of the MEDNETS framework to maintain the perceptual quality of speech signals during transmission is demonstrated by its highest PESQ ratings.

We use a Mean Opinion Score (MOS) based subjective quality evaluation measure to assess the recovered speech's quality even more. A popular metric for evaluating the perceived quality and naturalness of synthesised speech is MOS. Using the same headphones, a group of native speakers of the same language listen to the produced audio samples and score them using a five-point numerical scale. A score of 5 represents the highest perceived quality,

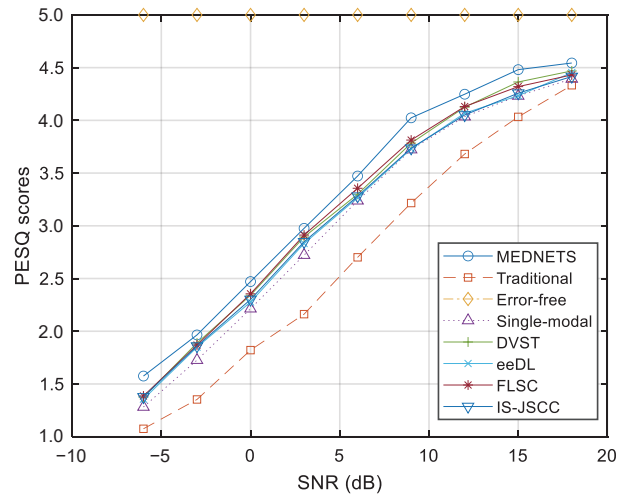


Fig. 7 PESQ score.

indicating very natural speech, while a score of 1 represents the lowest perceived quality, indicating highly unnatural speech. The results are presented in Fig. 8.

The results demonstrate that the proposed MEDNETS framework achieves the highest MOS scores among all the methods, except for the Error-free benchmark, which serves as an upper bound. At an SNR of 15 dB, the MEDNETS framework attains a MOS score of 4.4 ± 0.1 , indicating very natural and high-quality recovered speech. The MEDNETS framework consistently outperforms the traditional, single-modal, DVST, eeDL, FLSC, and IS-JSCC methods across all SNR levels.

The MEDNETS framework performs better than other frameworks because it can efficiently collect and maintain semantic information as it is being encoded and decoded. The cross-modal semantic fusion, similarity evaluation techniques, and deep neural network based architecture allow the MEDNETS framework to recover high-quality speech that is nearly natural-sounding and intelligible as the original speech. The MEDNETS framework offers a more realistic and robust solution for dependable and efficient speech transmission across wireless channels in real-world, practical medical communication settings.

The MEDNETS framework's high MOS ratings show that it can produce natural-sounding speech, improving user's experience and promoting efficient communication in medical settings.

5.2.3 Semantic image transmission

The image perceived similarity S_{image} is calculated by Eq. (34),

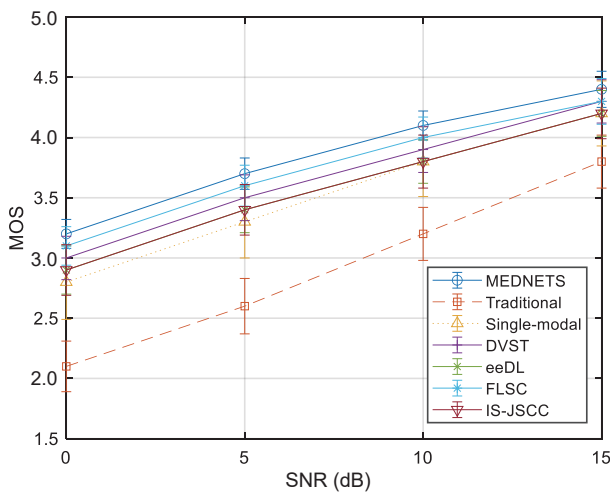


Fig. 8 MOS versus SNR.

$$S_{\text{image}} = \frac{1}{N} \sum_{i=1}^N \text{SSIM}(I_i, \hat{I}_i) \quad (34)$$

where I_i is the i -th original image, $\hat{I}_i$ is the corresponding reconstructed image, and $\text{SSIM}(\cdot, \cdot)$ denotes the structural similarity index measure.

The results shown in Fig. 9 shows that the MEDNETS framework is good in preserving both the semantic information and perceptual quality of images during transmission. It achieves the highest similarity scores compared to other methods. At an SNR of 18 dB, the MEDNETS framework achieves an image-perceived similarity score of 0.92.

Additional experiments were done to evaluate how well the proposed MEDNETS framework performs in terms of visual quality and efficiency in transmitting semantic images. The ratio of all image pixels to bits needed for transmission is known as the channel bandwidth ratio, or CBR for short. The MEDNETS architecture and benchmark techniques were employed to assess the visual quality of reconstructed images at different CBR levels. A portion of the MIMIC-CXR dataset, which contains medical images including CT scans and X-rays, is used in the testing. The average scores for each approach's Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) for the reconstructed images at various CBR levels are displayed in Fig. 10. Greater visual quality and structural resemblance to the original images are implied by higher PSNR and SSIM scores.

The MEDNETS framework has the highest PSNR and SSIM scores of all the techniques, according to the

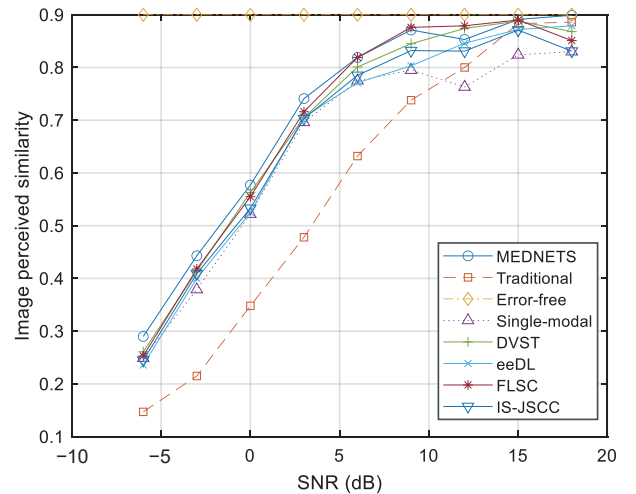


Fig. 9 Image perceived similarity.

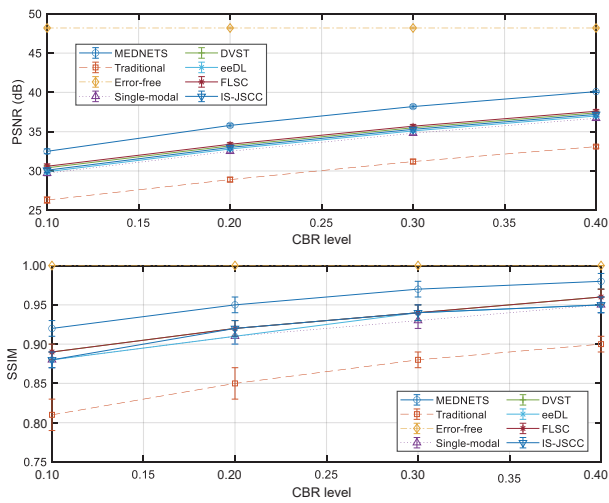


Fig. 10 Comparison for quality of visuals at several levels of CBR.

results shown in Fig. 10. With a CBR of 0.1, the MEDNETS framework yields an SSIM of 0.92 ± 0.01 and a PSNR of 32.5 ± 0.3 dB, demonstrating good visual quality and great structural similarity despite minimal bandwidth costs. The PSNR and SSIM scores of the MEDNETS framework increase with increasing CBR in a steady manner; at CBR being 0.4, they reach 40.1 ± 0.1 dB and 0.98 ± 0.01 dB, respectively.

The MEDNETS system's semantic process of cross-modal including algorithms related to encoding and decoding provides exceptional speed. Semantic decoders use these representations to reconstruct high-quality images after semantic encoders have extracted important semantic information from input images.

The cross-modal semantic fusion and similarity evaluation components of the framework enhance the semantic alignment and consistency between the original and reconstructed visuals. In contrast to benchmark methods, MEDNETS exhibits effective compression and transmission of medical images while lowering bandwidth requirements without compromising visual fidelity, offering better visual quality at reduced CBR costs. For situations involving real-time medical communication and effective network resource utilisation, this capability is essential. MEDNETS proves its potential to provide dependable and efficient medical picture transmission via 6G cross-modal semantic communication networks by consistently delivering high-fidelity reconstructions with higher visual quality and structural similarity, even under low bandwidth circumstances.

5.2.4 Cross-modal evaluation

The outcomes of similarity in the process of cross-modality between text and speech are shown in Fig. 11. The best similarity ratings go to the MEDNETS framework, indicating that it can capture and preserve the semantic relationship between various modalities. MEDNETS has a cross-modal similarity score of 0.95 at an SNR of 18 dB.

The process of similarity cross-modal results for text and images has been displayed in Fig. 12. The high similarity scores of the MEDNETS framework indicate that it does a good job of capturing and maintaining the semantic link between various modalities. MEDNETS has a similarity in cross-modal score of 0.90 at an SNR of 18 dB.

The outcomes of the similarity in cross-modal the analysis between voice and images are shown in

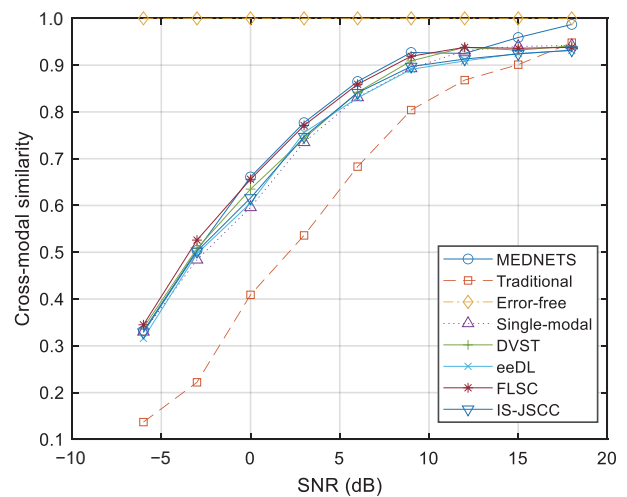


Fig. 11 Cross-modal similarity of speech-text.

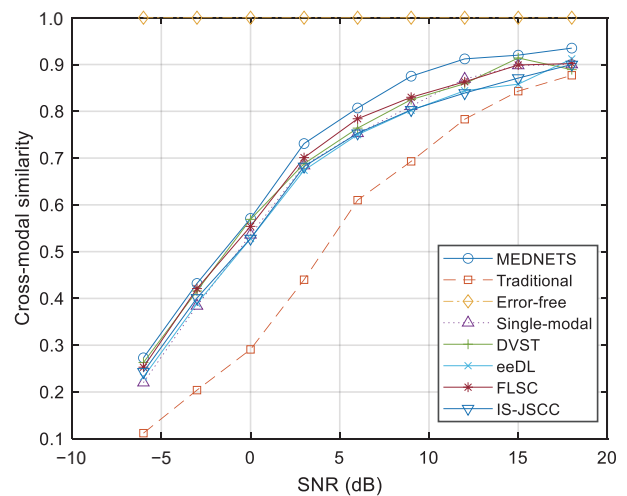


Fig. 12 Similarity cross-modal of image-text.

Fig. 13. MEDNETS achieves the highest similarity ratings, demonstrating its ability to preserve semantic links across many modalities. MEDNETS has a cross-modal similarity score of 0.88 at an SNR of 18 dB.

The similarity for the process of cross-modal results showing the effectiveness of the MEDNETS framework semantic connections has been made across several different modalities even under conditions where it would be even more challenging for the researchers where signal-to-noise ratios has been low. For real-time streaming in the field of 6G cross-modal semantic communication systems, MEDNETS performs better in a number of areas. It also guarantees high word and level of text similarity in a sentence, which is necessary for accurate transfer of vital medical data, including prescriptions and patient histories. At the next level, MEDNETS produces clear and intelligible transmission of audios for doctor and patient related consultations and medical diagnostics that has been done remotely thanks to its superior performance in speech similarity, PESQ scores, and MOS. At next stage, accurate medical image transmission (such as CT scans, MRIs, and X-rays) is ensured by MEDNETS, which maintains high levels of perceived picture similarity and visual quality across several bandwidth levels. This feature makes accurate diagnosis, disease monitoring, and treatment planning possible. It is helpful in emergency scenarios or resource-constrained situations where effective network resource utilisation is essential. Lastly, MEDNETS enhances cross-modal similarity between different data types (speech-text, image-text, and image-speech), enabling seamless integration and

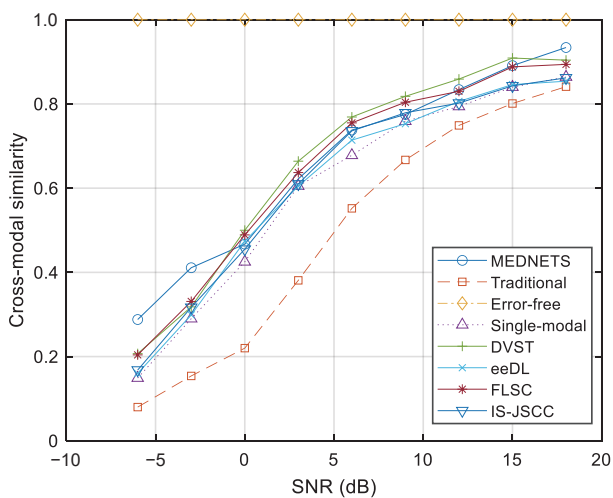


Fig. 13 Similarity cross-modal of image-speech.

understanding of multi-modal medical streaming data. This helps healthcare providers in making precise diagnoses and personalized treatment decisions based on comprehensive patient assessments.

Furthermore, we present experimental results on the inference time of the MEDNETS framework and compare it with the benchmarks. Table 1 shows the average inference time per sample for each method, and values are reported as mean $\pm$ standard deviation.

The results demonstrate that the MEDNETS framework achieves the lowest inference time among the benchmark methods, with an average of 60 ms per sample, highlighting the efficiency of MEDNETS and suitability for real-time medical data transmission. The optimized architecture and parallel processing capabilities of MEDNETS contribute to its fast inference speed, enabling timely and responsive cross-modal semantic communication in medical scenarios.

6 Conclusion

In our paper, we introduced MEDNETS, a deep neural network framework designed for 6G systems that enable real-time medical streaming through cross-modal semantic communication. Compared to other methods, MEDNETS is better performance wise in similarity scores, perceptual quality and understanding across different types of data and SNR levels. By combining various advanced multiple access techniques, like edge computing, AI integration and the capabilities of 6G networks, MEDNETS supported quick decision-making, intelligent processing of medical data and combined learning among healthcare providers. But along with this, there are two main limitations to consider. Firstly, the computational complexity of the deep neural networks used in MEDNETS, which can affect real-time processing speed; and secondly, the need for further validation

Table 1 Average inference time (mean $\pm$ standard deviation) comparison.

Method	Inference time (ms)
Traditional	120 $\pm$ 10
Error-free	5 $\pm$ 1
Single-modal	80 $\pm$ 5
DVST	95 $\pm$ 8
eeDL	90 $\pm$ 6
FLSC	85 $\pm$ 7
IS-JSCC	100 $\pm$ 9
MEDNETS	60 $\pm$ 3

across a broader range of medical datasets to confirm its effectiveness and reliability in diverse clinical scenarios. Future avenues for research include extending MEDNETS to accommodate real-time multimodal medical streaming data and interactive telemedicine applications, as well as incorporating privacy-preserving approaches and explainable AI.

Acknowledgment

This work was supported by the Local Service Project of the Department of Education of Liaoning Province (No. LJ212510160001) and the Key Research and Development Program Project of the Department of Science and Technology of Liaoning Province (No. 2024JH2/102600121).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

- [1] P. Meena, M. B. Pal, P. K. Jain, and R. Pamula, 6G communication networks: Introduction, vision, challenges, and future directions, *Wireless Pers. Commun.*, vol. 125, no. 2, pp. 1097–1123, 2022.
- [2] H. Zhu, L. Jia, J. Wang, and X. Xue, Distributed driver emotion prediction for intelligent 6G Internet of Vehicles, *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 10, pp. 17981–17992, 2025.
- [3] C. Yeh, G. D. Jo, Y. J. Ko, and H. K. Chung, Perspectives on 6G wireless communications, *ICT Express*, vol. 9, no. 1, pp. 82–91, 2023.
- [4] Y. Li, J. Huang, Q. Sun, T. Sun, and S. Wang, Cognitive service architecture for 6G core network, *IEEE Trans. Ind. Inf.*, vol. 17, no. 10, pp. 7193–7203, 2021.
- [5] T. B. Ahammed, R. Patgiri, and S. Nayak, A vision on the artificial intelligence for 6G communication, *ICT Express*, vol. 9, no. 2, pp. 197–210, 2023.
- [6] P. A. Babu and B. S. Babu, Deep learning-based Chatbox with attention bidirectional recurrent network-based puffer fish search for improved accessibility and cost reduction, *Cybern. Syst.*, vol. 57, no. 5, pp. 808–836, 2026.
- [7] X. Lv, S. Rani, S. Manimurugan, A. Slowik, and Y. Feng, Quantum-inspired sensitive data measurement and secure transmission in 5G-enabled healthcare systems, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 456–478, 2025.
- [8] H. Zhu, M. Xu, J. Wang, S. Zhu, and X. Xue, AIoT-enabled federated learning for green supply chain demand forecasting with privacy-preserving carbon reference effects, *IEEE Internet Things J.*, vol. 12, no. 19, pp. 39543–39555, 2025.
- [9] S. Praveena and S. Prasanna Devi, Optimizing retail operations through hybrid machine learning and deep learning techniques in demand forecasting, *Cybern. Syst.*, vol. 57, no. 1, pp. 1–43, 2026.
- [10] S. Wang, S. Gao, F. Qin, and W. Yang, A decomposition-based analysis of wireless multipath channel with full-wave simulation, *IEEE Antennas Wirel. Propag. Lett.*, vol. 21, no. 8, pp. 1615–1619, 2022.
- [11] X. Li, E. Leitinger, A. Venus, and F. Tufvesson, Sequential detection and estimation of multipath channel parameters using belief propagation, *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8385–8402, 2022.
- [12] T. Mao, and Z. Wang, Terahertz wireless communications with flexible index modulation aided pilot design, *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1651–1662, 2021.
- [13] D. Mahesh and T. S. Kumar, OptCosineDCNN: An advanced network attack detection and mitigation in SDN environments utilizing enhanced deep learning based approaches, *Cybern. Syst.*, doi: 10.1080/01969722.2025.2552941.
- [14] B. Wang, K. Xu, S. Zheng, H. Zhou, and Y. Liu, A deep learning-based intelligent receiver for improving the reliability of the MIMO wireless communication system, *IEEE Trans. Reliab.*, vol. 71, no. 2, pp. 1104–1115, 2022.
- [15] L. Zhao, H. Sun, F. Yang, Z. Wang, Y. Zhao, W. Tang, and L. Bu, A multimodal data driven rehabilitation strategy auxiliary feedback method: A case study, *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1181–1190, 2022.
- [16] H. Zhu, J. Wang, W. Zhu, Q. Huang, H. Xu, Z. Yu, W. Liu, and X. Xue, Distributed edge intelligence framework for secure and efficient data sharing in 6G-IoV, *IEEE Internet Things J.*, vol. 13, no. 5, pp. 8169–8181, 2026.
- [17] B. Cao, Z. Bi, Q. Hu, H. Zhang, N. Wang, X. Gao, and D. Shen, AutoEncoder-driven multimodal collaborative learning for medical image synthesis, *Int. J. Comput. Vis.*, vol. 131, no. 8, pp. 1995–2014, 2023.
- [18] M. Schmidhammer, C. Gentner, S. Sand, and U. C. Fiebig, Multipath-enhanced device-free localization in wideband wireless networks, *IEEE Antennas Wirel. Propag. Lett.*, vol. 20, no. 4, pp. 453–457, 2021.
- [19] H. Xie, Z. Qin, G. Li, and B. H. Juang, Deep learning enabled semantic communication systems, *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
- [20] S. Wang, J. Dai, Z. Liang, K. Niu, Z. Si, C. Dong, X. Qin, and P. Zhang, Wireless deep video semantic transmission, *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, 2023.
- [21] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, Semantic-preserved communication system for highly efficient speech transmission, *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 245–259, 2023.
- [22] B. Xie, Y. Wu, Y. Shi, D. W. K. Ng, and W. Zhang, Communication-efficient framework for distributed image semantic wireless transmission, *IEEE Internet Things J.*, vol. 10, no. 24, pp. 22555–22568, 2023.

- [23] S. Yao, K. Niu, S. Wang, and J. Dai, Semantic coding for text transmission: An iterative design, *IEEE Trans. Cognit. Commun. Networking*, vol. 8, no. 4, pp. 1594–1603, 2022.
- [24] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, Large AI model empowered multimodal semantic communications, *IEEE Commun. Mag.*, vol. 63, no. 1, pp. 76–82, 2025.
- [25] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, Large AI model-based semantic communications, *IEEE Wirel. Commun.*, vol. 31, no. 3, pp. 68–75, 2024.
- [26] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, and Y. Li, Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis, *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 186–201, 2023.
- [27] Y. Dai and G. Wang, A deep inference learning framework for healthcare, *Pattern Recognit. Lett.*, vol. 139, pp. 17–25, 2021.
- [28] X. Li, G. Wei, M. Jiang, and W. Zhou, Hierarchical Siamese network for real-time visual tracking, *Exp. Syst. Appl.*, vol. 238, p. 121651, 2024.
- [29] H. Huang, G. Liu, Y. Zhang, R. Xiong, and S. Zhang, Ensemble Siamese networks for object tracking, *Neural Comput. Appl.*, vol. 34, no. 10, pp. 8173–8191, 2022.
- [30] S. Tu, M. Waqas, A. Badshah, M. Yin, and G. Abbas, Network intrusion detection system (NIDS) based on pseudo-Siamese stacked autoencoders in fog computing, *IEEE Trans. Serv. Comput.*, vol. 16, no. 6, pp. 4317–4327, 2023.
- [31] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gündüz, and H. V. Poor, Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication, *IEEE Wireless Commun.*, vol. 30, no. 6, pp. 127–135, 2023.
- [32] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, Transformers in medical imaging: A survey, *Med. Image Anal.*, vol. 88, p. 102802, 2023.
- [33] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, Privacy-preserving object detection for medical images with faster R-CNN, *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 69–84, 2022.
- [34] S. Karthic and S. M. Kumar, Wireless intrusion detection based on optimized LSTM with stacked auto encoder network, *Intell. Autom. Soft Comput.*, vol. 34, no. 1, pp. 439–453, 2022.
- [35] F. Fang, P. Zhang, B. Zhou, K. Qian, and Y. Gan, Atten-GAN: Pedestrian trajectory prediction with GAN based on attention mechanism, *Cognit. Computat.*, vol. 14, no. 6, pp. 2296–2305, 2022.
- [36] H. Chauhan, D. Gupta, S. Gupta, M. J. Haque, A smart cradle system to monitor infants for healthcare baby wards based on IoT and blockchain, in *Proc. 3rd Int. Conf. Advances in Computing, Communication Control and*

Networking (ICAC3N), Greater Noida, India, 2021, pp. 606–609.

- [37] L. Kakkar, D. Gupta, S. Tanwar, S. Saxena, K. Alsubhi, D. Anand, I. D. Noya, and N. Goyal, A Secure and efficient signature scheme for IoT in healthcare, *Comput. Mater. Continua*, vol. 73, no. 3, pp. 6151–6168, 2022.



Xiaohong Lv received the master of medicine degree from Dalian Medical University, China, in 2008, and the PhD degree from China Medical University, China, in 2022. She is currently a professor at Department of Imaging, The First Affiliated Hospital of Jinzhou Medical University, China. Her research interests include intelligent processing, transmission, reconstruction of medical data, and intelligent analysis of medical images.



Jing Wang received the BS degree in computer science from Suzhou University of Science and Technology, China, in 2013, the MS degree in computer science from Northeastern University, China, in 2015, and the PhD degree in software engineering from Southeast University, China, in 2021. He is currently an assistant researcher at School of Computer Science and Engineering, Southeast University, China. His research interests include pattern recognition and machine learning.



Shalli Rani completed her postdoctoral fellowship from Manchester Metropolitan University, UK, in 2023. She is currently a professor at Institute of Engineering and Technology, Chitkara University, India. She received the master of computer applications degree from Maharishi Dyanand University, India, in 2004, the master of engineering degree in computer science from Janardan RaiNagar Vidyapeeth University, India, in 2007, and the PhD degree in computer applications from Punjab Technical University, India, in 2017. Her main area of interest and research are wireless sensor networks, underwater sensor networks, machine learning, and Internet of things. She has published/accepted/presented more than 100+ papers in international journals/conferences.



Ayush Dogra is currently working as an assistant professor at Institute of Engineering and Technology, Chitkara University, India. He received the master degree in business management from Indira Gandhi National Open University, India, in 2015, and the PhD degree in electronics and communications engineering from Panjab University, India, in 2019. He has published more than 100 papers in highly reputed journals and conferences. His main research interests include medical imaging, digital imaging, and infrared imaging.



Adam Slowik received the BEng and MEng degrees in computer engineering from Koszalin University of Technology, Poland, one PhD degree in electronics from the same university, in 2007, and another PhD degree in computer science from Czestochowa University of Technology, Poland. Since October 2013,

he has been an associate professor at Department of Electronics and Computer Science, Koszalin University of Technology, Poland. He is the author or co-author of over seventy articles and two books. His research interests include soft computing, computational intelligence, machine learning, and bio-inspired global optimization algorithms, and their engineering applications. He is also an associate editor of *IEEE Transactions on Industrial Informatics* and a reviewer for many international scientific journals.



Daohua Pan received the PhD degree in computer science and technology from Harbin Institute of Technology, China, in 2021. She is currently an associated professor at College of Computer Science, Harbin Institute of Technology, China. Her research interests include artificial intelligence, pattern recognition, and

pervasive computing.



Huamao Jiang received the master degree of medicine from Jinzhou Medical College, China, in 2000, and the PhD degree from China Medical University, China, in 2006. He is currently working at Multi-modal Data Fusion and Precision Medicine Laboratory, The First Affiliated Hospital of Jinzhou Medical University,

China. His research interests include the application and analysis of medical big data, intelligence of surgical operations, and robot assisted minimally invasive surgery.



Yanhong Feng received the master degree of medicine from Jinzhou Medical College, China, in 2003, and the PhD degree from China Medical University, China, in 2006. She is currently a professor and director at Department of Imaging, The First Affiliated Hospital of Jinzhou Medical University, China. Her research

interests include the development and transformation of cutting-edge technologies for intelligent analysis of medical images.