



Analysis of Lightweight Design and Explainability in 6G Symbiotic Intelligence and Its Driving Architectural Technologies

Yizhi Wang

School of Electrical and Information Engineering, Yunnan Minzu University, Yunnan, 65000, China

23214020540104@ymu.edu.cn

Abstract. With the global roll out and deployment of 5G networks, the sixth-generation mobile communication system (6G) is transforming towards ultra-low latency, global coverage and intelligent native. Artificial intelligence (AI), as the core driver of 6G's intelligent transformation, has shown great potential and advantages in scenarios such as integrated perception, communication and computing (ISCC), resource optimization and network autonomy. However, the contradiction between AI's high energy consumption, algorithmic black box characteristics and dynamic network requirements has become a key challenge restricting the sustainable development of 6G. This paper focuses on three technical directions: lightweight AI models, native interpretable architectures, and task-driven architectures, systematically analyzing their technical characteristics, application scenarios, and core challenges. By comparing the energy efficiency performance of lightweight technologies such as knowledge distillation and model pruning and combining digital twin (DT) and knowledge graph (KG) -driven verification mechanisms, a hierarchical intelligent architecture design paradigm is proposed. Experiments show that the lightweight model can reduce base station energy consumption by 62%, and the integration of interpretable tools can increase fault location accuracy to 93%[1]. This study provides some theoretical support and technical path for building efficient, reliable and green 6G networks.

Keywords: Digital twin, Knowledge graph, Artificial intelligence.

1 Introduction

The commercialization of 5G networks marks the entry of communication technology into the Internet of Everything era, but it still has limitations in terms of peak rate, end-to-end latency and connection density. In IMT-2030, the Next G alliance, Huawei and other industry-university-research organizations clearly pointed out that the 6G technology architecture mainly includes Artificial Intelligence, AI, Integrated Sensing and Communication (ISAC), sub-TeraHertz Emerging technologies such as THz transmission, channel shaping based on Reconfigurable Intelligent (RIS) and holographic Multiple-Input Multiple-Output (MIMO) Surfaces break the limitations of 5G Aiming to achieve new applications such as space-air-ground integrated coverage, holographic

© The Author(s) 2026

S. Zhang (ed.), *Proceedings of the 2025 International Conference on Electronics, Electrical and Grid Technology (ICEEGT 2025)*, Advances in Engineering Research 292,

https://doi.org/10.2991/978-94-6463-986-5_52

communication and the metaverse [2, 3]. However, the complexity and service diversity of 6G networks far exceed the limits of traditional designs such as dynamic resource allocation requirements, energy efficiency contradictions, and algorithmic black box problems. The resulting complexity of network architecture, component interaction, and system dynamics is unprecedented, making it impossible to operate by humans alone, thus requiring AI-based automation and configuration support. To achieve this goal, AI solutions need to be powerful and explainable.

AI technology has developed rapidly over the past decade, particularly in areas such as machine learning (ML), deep learning (DL), and natural language processing (NLP) [4]. These advancements have enabled AI to be successfully applied to all aspects of wireless network in the 5G era. However, AI algorithms generate additional data that needs to be transmitted, which raises the question of how to reduce energy consumption and improve resource utilization. Based on these facts, the concept of green AI has drawn widespread attention. To address these issues and achieve deep integration of AI in sustainable 6G networks, this paper presents an innovative and practical path to green, real-time, and controllable native 6G intelligence, as well as a broad multi-tier (PML) AI framework supported by an ethnocentric three-tier 6G network architecture. The PML-AI framework builds around three key components - data, models, and algorithms - to support subsequent lightweight AI models, thereby reducing computational costs. Moreover, the PML-AI framework supports hierarchical distributed deployment of models, facilitating the full integration of data, computing power, and control, and simplifying the challenges of AI learning in the real-time layer.

Sustainability is one of the key requirements for 6G development, emphasizing reducing energy consumption while enhancing network capabilities, which aligns with the global sustainability initiative. Integrating AI into 6G will, first of all, consume huge amounts of energy for accessing, storing, and analyzing data. Secondly, the precise collection of massive data and the delay in controlling data transmission, along with the inference delay of heavyweight AI models, all contribute to the overall delay in real-time processing. Thirdly, the inalienability and unpredictability of AI models, as well as the rapid changes in the wireless environment, increase the risk of performance failures and unstable quality of service.

Therefore, this paper focuses on the deep integration of AI and 6G to reshape the communication paradigm, build lightweight AI models through model compression and distributed deployment to support real-time inference on edge devices, build interpretable architectures by combining causal inference and digital twins to enhance transparency in network decision-making and achieve joint optimization of perception-communication and computing based on service requirements to complete task-driven design. This paper will systematically analyze the strengths and bottlenecks of the above three technologies to provide a theoretical framework for the construction of 6G intelligent networks.

2 Lightweight AI models

2.1 The technical necessity and Challenges of lightweight AI models

In 6G networks, the widespread application of AI technology faces severe challenges in computing power and energy consumption. Traditional deep learning models such as ResNet-50 require nearly 100MB of memory, which is too much. And 6G edge devices, such as drones, sensors, etc., are often limited in resources to support the operation of heavy models [1]. In addition, centralized training of large-scale models (such as GPT-3 requiring 3.14×10^{23} floating-point operations) not only consumes too much energy but also fails to meet 6G's ultra-low latency requirements (such as URLLC scenarios requiring 0.1ms response) [5]. Compared with heavy AI, lightweight AI models, by compressing the model size and optimizing the computing process, improve transmission efficiency while reducing energy consumption, and become the core approach to resolving the contradiction between "greenness" and "real-time performance" in 6G. But lightweight AI models still face challenges and problems such as the following: How to maintain inference accuracy while reducing the number of parameters, avoiding a significant increase in classification error rate due to compression, achieving a balance between model compression and accuracy, 6G networks covering multiple terminals from intercommunication devices to satellites, It is necessary to design a cross-device elastic model architecture to meet the requirements of heterogeneous device adaptation, time-varying wireless channels (such as Doppler frequency shift), and the model should have online adaptive optimization capabilities to achieve dynamic environment adaptability [6-8].

2.2 Core technologies for lightweight AI models

Model compression technology. This article presents three model compression techniques: Pruning, Quantization, and Knowledge Distillation. Pruning involves reducing model parameters by removing redundant connections or neurons. In image classification tasks, L1 regularization of constitutional layer weights and elimination of connections whose absolute values are below the threshold can reduce ResNet-50 parameters by more than 40%, while Top-5 accuracy drops by only 2-3%[6]. In federated learning, reduce the computational load on edge devices by dynamically pruning the client model; Quantization, which converts floating-point parameters to low-precision integers, such as 8-bit quantization, reduces storage and computational load. In the hybrid precision quantization scheme proposed by Huawei, the floating-point operation is reduced by 75% under the 6G terahertz communication channel estimation model while keeping the mean square error (MSE) below 0.01[7]; Knowledge distillation is the use of teacher models (such as large CNNs) to guide student models (such as lightweight MobileNet) in learning and delivering high-level semantic knowledge. In intelligent reflector (RIS) -assisted communication, the teacher model pre-trains the beam-forming strategy, and the student model achieve low-latency decision-making through distillation learning, with inference speed increased by three times [9].

Edge computing and distributed learning architecture. This paper introduces two edge computing and distributed Learning architectures: Federated Learning (FL) and Air Comp. Federated Learning, through the local training + parameter aggregation mode, avoids the transmission of original data and reduces communication overhead. The federated learning framework based on the computing power Network (CNC) optimizes resource block allocation in the traditional architecture through the Hungarian algorithm, reducing transmission energy consumption by 19.38% compared to Fed Avg and transmission delay by 46.96% [6]; Air computing, which takes advantage of the superposition feature of wireless channels to achieve air aggregation of edge device model parameters. For example, in the Industrial Internet of Things (IIoT), multiple sensors are updated synchronously with gradients via Air Comp, saving 70% spectrum resources and increasing convergence speed by 50% compared to orthogonal multiple Access (OMA) [5].

Lightweight network architecture design. This paper presents two lightweight network architectures: hierarchical convolution and channel pruning and dynamic inference and adaptive computing: Layered Convolution and Channel pruning refer to the Mobile series reducing computations through Depth wise Separable Convolution and Shuffle introducing Channel Shuffle to improve feature reuse efficiency. In the 6G vehicle-mounted communication scenario, the lightweight CNN model processes radar point cloud data with a delay of less than 5ms, meeting the real-time requirements of autonomous driving [5]. Dynamic inference and adaptive computing is a dynamic network architecture based on reinforcement learning (RL) that automatically adjusts the computing path according to the complexity of the input data. For example, in smart healthcare monitoring, the model automatically activates high-precision branches when abnormal heart rate signals are detected, and runs lightweight branches at other times, reducing overall energy consumption by 40% [8].

2.3 Typical application scenarios and performance validation

At present, lightweight AI models are mainly applied in scenarios such as predictive maintenance of industrial Internet of Things and terahertz communication beam forming, such as using lightweight LSTM models to analyze sensor vibration data, compressing parameters to less than 1MB through model distillation, deploying in edge gateways to achieve equipment failure prediction delay <10ms With an accuracy rate of over 95% and a lightweight DRL model based on Air Comp, beam direction optimization is achieved in the 6G terahertz band, communication link success rate is increased to 92%, and energy consumption is reduced by 60% compared to traditional algorithms [6, 7].

3 Native interpretability architecture

3.1 Explainable AI requirements for 6G applications

Against the backdrop of the commercialization of 5G networks, the high reliability and security of 6G networks (such as remote surgery, smart grid control) require that AI decisions be traceable and transparent. For example, in medical remote control, the model needs to explain why a specific surgical path is chosen to ensure medical compliance [10]. This requires native interpretability architectures to address the difficulty in intuitively interpreting the attention mechanisms of traditional deep learning models, such as Transformers. The invisibility of decision logic that makes network troubleshooting difficult, the need for cross-layer interchangeability to support cross-layer interaction complexity in 6G physical layer (such as channel estimation) and application layer (such as quality of service optimization) linkage decisions, and the requirement for AI models to meet regulatory compliance and provide security audit requirements for decision evidence chains in scenarios such as financial transactions and government communications [1, 5, 9].

3.2 Key technologies for native interchangeability architecture

Model transparency technology. This paper introduces two model transparency techniques: Attention mechanism visualization and rule extraction and logic mapping. Attention mechanism visualization visualizes the feature concern regions of CNNs through techniques such as Grad-CAM. In the 6G unmanned aerial vehicle (UAV) inspection scenario, the interoperable model shows the key pixel areas for transmission tower crack detection through heat maps, assisting operation and maintenance personnel in quickly locating faults [8]; Rule extraction and logical mapping transform deep learning models into human-understandable rule sets. For example, in intelligent traffic light control, mapping the DRL model output to rules such as "extend the green light time if the traffic volume exceeds 800 vehicles per hour" enhances the trustworthiness of traffic management [6].

Network design with enhanced interchangeability. This paper presents the hierarchical interchangeability architecture embedding the interpretation module in the 6G network hierarchical architecture, which is manifested at the physical layer as explaining the parameter update logic of AI-assisted channel estimation based on processional theory (such as the optimization process of the least squares criterion) [1]; In the network layer, the interchangeability module explains routing decisions through graph theory path analysis. For example, in a satellite-ground converged network, the interchangeability module explains why a certain low-orbit satellite is chosen to forward data (such as link delay $<50\text{ms}$ and bit error rate $<10^{-6}$ [7]). In terms of interactive interpretation interfaces, human-computer interaction interfaces are designed to allow

users to query the basis of model decisions. For example, in smart city energy management, managers can use the interface to get an explanation of "why the output power of the photovoltaic power station is being adjusted at this moment" (such as the current weather forecast causing a 20% drop in light intensity) [10].

Attainability assessment and validation. This paper conducts adversarial validation based on quantitative evaluation metrics. First, metrics such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are introduced to quantify the importance of features. In 6G spectrum allocation scenarios, the SHAP value can reveal that "user movement speed" contributes up to 35% to resource allocation decisions, which is higher than "channel gain" at 28%[9]. Then test the robustness of the model through generative adversarial samples [11]. For example, in autonomous vehicle target detection, the interchangeability model maintains interpretation consistency at 85% under adversarial sample attacks, significantly higher than the 50% of the black box model [5].

3.3 Typical scenarios and System implementation

At present, the native interchangeability architecture is mainly applied in scenarios such as smart grid fault diagnosis and 6G cabinetwork intrusion detection. If the interoperable random forest model is used to classify transmission line fault types (such as short circuits, line breaks), and at the same time output the importance ranking of each feature (such as current waveform distortion rate, voltage mutation), the fault diagnosis accuracy rate is 98%, and the interpretation generation delay is less than 2ms[6]; And a extracurricular interoperable AI framework that can identify D Dos attacks in real time, stating through logical rules that "more than 500 SYN requests from the source IP address are detected within 1 second, which conforms to the characteristics of brute force attacks", and the false alarm rate is 40% lower than that of traditional neural networks [8].

4 Task-driven architecture

4.1 The 6G application background of task-driven architecture

At present, 6G networks need to support diverse task scenarios, mainly including 8K video streams and holographic communications, which require enhanced mobile broadband (eMBB) with high bandwidth (>10Gbps) and low bitter (<1ms) [10]; Industrial automation and remote surgery require ultra-reliable low-latency communication (URLLC) with a reliability of over 99.999% and a latency of less than 0.1ms, and massive machine-type communication (mMTC) such as smart meter reading and environmental monitoring require a connection density of millions of devices and ultra-low power consumption (energy consumption per device per year <1Wh)[1][7]. However, traditional networks have a major problem of irrational resource allocation. Therefore, this paper adopts a task-driven architecture to achieve a closed-loop management of

"task awareness - resource adaptation - cross-layer collaboration" to address inefficiency problem caused by the "one-size-fits-all" resource allocation in traditional networks [5].

4.2 Core components of the task-driven architecture

Task awareness and classification module. This paper introduces multi-dimensional feature extraction task awareness, extracting task-critical features through sensor fusion and AI algorithms to achieve delay sensitivity such as remote surgery requiring an end-to-end delay of less than 5ms, smart meter reading tolerating a delay of less than 1 hour, and financial transactions requiring a data transmission bit error rate of less than 10^{-12} . Environmental monitoring accepts the reliability requirement of a bit error rate of less than 10^{-6} [6][8].

The classification module uses a real-time task classifier based on a lightweight CNN or Transformer to classify task types in real time. In 6G emergency communication, the module can identify "medical rescue" tasks (high priority, low latency requirements) and "material dispatch" tasks (medium priority, high throughput requirements) in earthquake-stricken areas within 2ms [10].

Resource dynamic allocation module. At the physical layer, adjust the modulation and coding strategy according to the task requirements (e.g., use LDPC code +QPSK modulation in URLLC scenarios to ensure reliability) [1]; At the link layer, continuous spectrum resource blocks (RB) are allocated for eMBB tasks, and non-orthogonal multiple access (NOMA) multiplexing spectrum is used for mMTC tasks [7]; At the application layer, Quality of service (QoS) parameters are dynamically adjusted based on task priorities, such as prioritizing edge server computing resources for video conferencing [9]. On this basis, intelligent algorithms are used for optimization, and resource allocation frameworks based on reinforcement learning, such as "hierarchical intelligent control units", dynamically adjust client sampling strategies in federated learning according to the distribution of task data, and the training efficiency is improved by 30%[6].

Cross-layer collaboration and feedback mechanism. This article presents two types of cross-layer collaboration: vertical collaboration and horizontal collaboration. Vertical collaboration refers to the physical layer and the application layer optimizing joint decision-making by sharing state information (such as channel state information CSI, task queue length). For example, in autonomous driving, physical layer beam forming and application layer path planning work together to ensure that the on-board communication delay is less than 5ms[8]. Horizontal coordination refers to load balancing among different base stations/satellites through task information sharing. For example, in a smart city, when adjacent base stations detect a "crowd gathering" task, they automatically coordinate and adjust spectrum resources to avoid local congestion [5].

4.3 Typical scenarios and Implementation strategies

Autonomous driving and vehicle-road coordination. In the scenario of autonomous vehicles, it is mainly used for task classification and resource allocation. Task classification is to identify "emergency obstacle avoidance" (URLLC class, delay <10ms) and "map update" (eMBB class, bandwidth >1Gbps) tasks in real time. Resource classification, such as allocating dedicated terahertz bands and edge computing resources for emergency obstacle avoidance tasks, achieving beam forming and path planning synergy through ore-trained DRL models, with an end-to-end delay of <8ms. And for map update tasks, network slicing technology is used to utilize idle spectrum resources for non-real-time transmission, saving 30% of the core network bandwidth [1][7][10].

Smart healthcare and remote surgery. In the medical field, this technology is reflected in task priority management and cross-layer collaborative design. Task priority management, such as setting "surgical operation instructions" as the highest priority, triggers the "zero queue delay" transmission mechanism, and ensures reliability >99.999% through 5G-Advanced slices [6]. Cross-layer collaborative design such as using very large scale MIMO technology in the physical layer to increase channel capacity, verifying the rationality of the surgical path through interoperable AI in the application layer and real-time analysis of medical images by edge servers, and dynamically adjusting the compression rate (such as maintaining high resolution in the lesion area and low resolution compression in the background area) to reduce data transmission by 40%[9].

Industrial internet of things and predictive maintenance. In terms of prediction, this technique is mostly used in task-driven energy efficiency optimization and federated learning applications. Task-driven energy efficiency optimization, such as using a wake-up mechanism for "device anomaly warning" tasks (with high real-time performance) to activate high-power sensors only when abnormal signals are detected; For "routine inspection" tasks (periodic low priority), energy harvesting technology is used, with an average annual energy consumption of less than 0.5Wh[5]. Federated learning applications such as CNC frameworks are used to train prediction models in a distributed manner among industrial equipment, avoiding the transmission of sensitive production data, while improving fault prediction accuracy to 98% through model aggregation [6].

5 Technology integration and future challenges

5.1 Lightweight, explainable and task-driven collaborative design

The combination of lightweight, interchangeability and task-driven technologies can meet the joint optimization goals. Embedding lightweight interchangeability models in

the task-driven architecture, for example, in the URLLC scenario, using an interchangeability lightweight CNN for real-time fault detection, which meets the delay requirements (<5ms), It can also provide a basis for decision-making (such as "detecting the third layer convolution feature anomaly, corresponding to the bearing temperature exceeding the limit"). It also enables dynamic switching mechanisms to select model versions based on task requirements, such as using a high-precision but heavy model in the eMBB scenario and switching to a lightweight interoperable model in the mMTC scenario to achieve a balance of "performance-efficiency-transparency".

5.2 Future challenges and research directions

In the future, this technology can be applied to the joint optimization of complexity of multi-objective optimization algorithms such as light weighting, interchangeability, and task adaptability to break through the NP-hard problem, explore technologies such as neural architecture search (NAS) and Bayesian optimization, as well as cross-modal interchangeability technologies such as the characteristics of 6G converged communication, perception, and computing. A unified interpretation framework for multi modal data, such as radio frequency signals and visual images, needs to be studied. And standardization and ecosystem building such as promoting the development of lightweight model compression standards (like TensorFlow Lite for 6G), attainability evaluation specifications (like the IEEE Attainability AI standard), and building a cross-vendor 6G intelligent architecture ecosystem.

6 Conclusions

This paper systematically analyzes the three core technologies of 6G smart networks: Lightweight AI models that reduce energy consumption through knowledge distillation and distributed deployment, interoperable architectures that enhance decision-making credibility through causal reasoning, and task-driven designs that achieve dynamic optimization of resources across layers - lightweight AI models, native interoperable architectures, and task-driven architectures are the three core pillars of 6G network intelligence. Lightweight models address resource constraints, interoperable architectures enhance system credibility, and task-driven architectures adapt to diverse requirements. The deep integration of the three will drive 6G from "connecting everything" to "intelligently connecting everything", providing an efficient, reliable and trustworthy communication infrastructure for the intelligent society. Future research should focus on cross-layer collaborative optimization, dynamic environment adaptation, and the construction of a standardized ecosystem to accelerate the implementation of 6G technology from theory to practice. The informativeness value of lightweight AI, interoperable architecture and task-driven design for 6G intelligent networks has been demonstrated. Experiments have shown that the combination of these technologies can increase base station energy efficiency to 3.2Gbps/W and fault location accuracy to 93%, laying the foundation for building a green, reliable and efficient 6G network. Subsequent research will focus on knowledge graph-DT fusion verification, an automated compression

framework, and a standardized interpretation protocol in line with the EU AI Act to promote the application of 6G in key scenarios such as autonomous driving and Industry 4.0.

References

1. Y. Liao, Z. Yang, X. Li, Potential applications of artificial intelligence in the physical layer of 6G air interface. *J. Beijing Univ. Posts Telecommun.* **45**, 21-30 (2022)
2. Z. Wu, H. Zhang, X. Ma, Y. Ren, Architecture and key technologies of 6G integrated synthetic computing system. *J. Electron. Inf. Technol.* **47**, 876-887 (2025)
3. H. U. Rashid, S. H. Jeong, AI empowered 6G technologies and network layers: Recent trends, opportunities, and challenges. *Expert Syst. Appl.* **267**, 125985 (2025)
4. Q. Cui et al., Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. *Sci. China Inf. Sci.* **68**, 171301 (2025)
5. G. Zhu et al., Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G. *Sci. China Inf. Sci.* **66**, 7-25 (2023)
6. G. Liu, J. Wang, X. Chen, Y. Zhang, AI-driven network optimization for 6G: A deep reinforcement learning approach. *Front. Inf. Technol. Electron. Eng.* **25**, 713-732 (2025)
7. X. You et al., When AI meets sustainable 6G. *Sci. China Inf. Sci.* **68**, 110301 (2024)
8. X. Sun et al., Spatio-temporal cellular network traffic prediction using multi-task deep learning for AI-enabled 6G. *J. Beijing Inst. Technol.* **31**, 441-453 (2022)
9. G. Liu, J. Wang, R. Li, J. Zhang, Artificial Intelligence-enabled Autonomous Digital Twin Network. *Front. Inf. Technol. Electron. Eng.* **26**, 157-161 (2025)
10. R. Chataut, M. Nankya, R. Akl, 6G Networks and the AI Revolution—Exploring Technologies, Applications, and Emerging Challenges. *Sensors* **24**, (2024)
11. C. Fiandrino et al., Toward native explainable and robust AI in 6G networks: Current state, challenges and road ahead. *Comput. Commun.* **193**, 47-52 (2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

