Evolution of AI-Driven Mobile Networks: Trends, Toolkits, and Deployment Frameworks

Ian Joseph Chandra*, Fransiscus Asisi Bimo*, Cheng-Chang Chen†, and Ray-Guang Cheng*
* Dept. of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taiwan
† Institute for Information Industry (III), Taiwan
Email: crg@mail.ntust.edu.tw

***Abstract*—Artificial Intelligence (AI) is rapidly transforming mobile networks by enabling autonomous, adaptive, and data-driven functionalities across the radio access network (RANs) and core networks (CNs). This paper aims to analyze the emerging trends in the adoption of AI technologies within mobile networks, evaluate relevant AI toolkits and their application potential, and investigate the deployment conditions and computational requirements of AI models across various mobile network nodes. Based on these insights, we propose system integration strategies and future application directions to support scalable, efficient, and standardized AI deployment in next-generation mobile networks.**

I. Introduction

Artificial Intelligence (AI) is becoming a foundational technology in mobile network evolution from 5G to 6G [1]. The network planning and design, self-organizing networks (SON), shared infrastructure, protocol code generation, and capacity forecasting are identified by Linux Foundation Networking (LFN) as key enablers of intelligent networking [1]. Network operators may adopt generative AI to support precise small cell placement, MIMO antenna configuration, beamforming, and optimized backhaul connections. AI-based algorithms can be employed for autonomous optimization and network resource management. Additionally, leveraging 5G RAN infrastructure for model training and inference can enhance both AI capabilities and overall network efficiency. Generative AI may also enable co-pilot functionality for network protocol software development and be used to optimize network capacity—avoiding unnecessary upgrades or degraded performance caused by overloaded circuits [1].

Artificial Intelligence (AI) is rapidly becoming a foundational technology driving the evolution of mobile networks from 5G to 6G. As mobile networks grow in complexity and scale, AI offers powerful tools for enhancing efficiency, automation, and adaptability. The Linux Foundation Networking (LFN) identifies six key enablers of intelligent networking, including network planning and design, self-organizing networks (SON), shared infrastructure, protocol code generation, and capacity forecasting [1]. The network operators may adopt generative AI for precise small cell placement, MIMO antenna configuration, beamforming optimization, and backhaul planning; utilize AI-based algorithms to enable autonomous network optimization and dynamic resource management; leverage existing mobile infrastructure for model training and inference to enhance both AI performance and overall network efficiency; provide co-pilot functionality for network protocol software development; and prevent unnecessary upgrades and mitigate performance issues caused by overloaded circuits for capacity optimization [1].

6G envisions distributed AI across radio access networks (RANs) and core networks (CNs) []; digital twins and real-time adaptation []; and intent-driven network behavior [].

The integration of AI presents both technical opportunities and deployment challenges that span across architecture [], performance [], and standardization [].

The 5G to 6G mobile networks shift from automation to autonomy [].

AI is increasingly applied in both the RAN and the CN to support traffic prediction, mobility management, dynamic resource allocation, anomaly detection, and network optimization [2], [4], [6]. AI enhances network efficiency, resilience, and service quality by enabling predictive and autonomous behaviors.

However, several key challenges hinder widespread deployment. These include:

- **Data Collection and Quality:** AI models require large volumes of high-quality, real-time network data. Yet data collection is often limited by privacy constraints, inconsistent formats, and insufficient labeling [6].
- **Interoperability:** Integrating AI solutions across multi-vendor RAN and CN environments remains complex due to varying data models and proprietary interfaces [5].
- **Real-Time Processing:** Many AI applications (e.g., beamforming, resource scheduling) must meet strict latency and timing constraints, challenging the capabilities of edge inference platforms [3].
- **Security and Trustworthiness:** AI introduces new attack surfaces, such as model poisoning and adversarial inputs, and raises concerns about explainability, accountability, and compliance [].

Ongoing efforts by 3GPP, the O-RAN ALLIANCE, and ITU-T aim to address these issues by defining standardized architectures, data handling frameworks, and security guidelines for AI-native networks.

This study focuses on the following key topics:

- Analysis of AI technology development trends in mobile networks
- Evaluation of AI toolkits and potential application for mobile networks

- AI model deployment and computational requirements across mobile network nodes
- Recommendations for system integration and future development

II. Trends and Standardization in AI-Driven Mobile Networks

In this section, we will analyze the trends and research directions of AI technologies in mobile networks; consolidate their application models and evolution paths; and summarize the standardization efforts in 3GPP and O-RAN ALLIANCE and potential applications addressed by the AI-RAN Alliance.

AI-native networks represent a new paradigm in which AI is deeply embedded within the core architecture, enabling unprecedented levels of automation, optimization, and intelligence [10]. The ITU-T Focus Group on AI-Native Networks (FG AI-Native) was established in July 2024 to identify the requirements, challenges, and opportunities associated with AI-native networks.

A. Trends and research directions of AI technologies in mobile networks

B. Application models and evolution paths

C. Standardization efforts

D. Potential applications

Global momentum is growing toward embedding AI in mobile networks to enhance automation and efficiency [13].

AI is increasingly applied in prediction, optimization, and control across RAN and core networks [24].

Standards bodies such as 3GPP and O-RAN Alliance are formalizing AI interfaces and functions within the network stack [15].

A convergence between academic research and industrial standardization is accelerating practical AI-RAN frameworks [16].

Emerging trends highlight the need for unified data-driven architectures to enable scalable and interoperable AI deployments [17].

III. AI Toolkits and Potential Applications for Mobile Networks

In this section, we provide a brief overview of potential AI toolkits and assess their technical feasibility, potential applications, and deployment constraints in mobile networks. These insights help us identify suitable use cases and their corresponding integration strategies.

A. AI toolkits

Toolkits like NVIDIA Aerial SDK provide GPU-accelerated inference capabilities tailored to RAN workloads [18].

B. Technical feasibility

C. Potential applications in mobile networks

These toolkits facilitate use cases such as beam management, interference mitigation, and traffic prediction [19].

D. Deployment constraints in mobile networks

Evaluation of these tools reveals deployment challenges including hardware dependency, data locality, and latency sensitivity [16].

By mapping toolkit capabilities to mobile use cases, we can guide optimal integration paths [24].

These findings support a readiness matrix for toolset selection aligned with AI-RAN goals [20].

IV. Deployment and Resource Requirements for AI Models in Mobile Network Nodes

This section analyzes the deployment constraints and computational resources for AI models in mobile network nodes. We evaluate the deployment conditions and resource allocation strategies for AI models based on the varying requirements of different network nodes, including latency tolerance, bandwidth constraints, types of data collection, and inference cycles to enhance the feasibility assessment.

A. Deployment constraints for AI models in mobile network nodes

B. Computing resource requirements for AI models in mobile network nodes

AI deployment in RAN, MEC, and core nodes differs in latency constraints, data types, and model sizes [21].

Real-time inference at the edge demands lightweight models with localized training and decision-making [22].

Centralized training and distributed inference architectures face trade-offs in accuracy, efficiency, and cost [23].

Different scenarios require tailored resource allocation policies for computing, memory, and bandwidth [24].

A systematic framework is needed to model AI workload requirements and match them to node capabilities [20].

V. System Integration and Future Deployment

This section presents comprehensive system integration recommendations and outlines future development directions. It includes integration strategies for incorporating AI technologies into mobile networks, proposed demonstrative architectures, potential implementation challenges, and corresponding mitigation strategies to support the evolution of AI-driven mobile networks.

A. Integration strategiesn

B. Demonstration architectures

C. Potential implementation challenges

D. Corresponding mitigation strategies

Integrating AI across control, user, and management planes requires harmonized APIs and orchestration platforms [25].

Open testbeds are critical for validating cross-layer AI strategies and standard-compliant designs [15].

Key challenges include lifecycle management, real-time orchestration, and data governance [26].

Techniques like federated learning offer a scalable and privacy-preserving alternative to centralized AI [21].

Future success depends on a co-designed ecosystem involving operators, vendors, and research bodies [20].

REFERENCES

[1] Linux Foundation Networking, "Intelligent Networking, AI and Machine Learning for Telecommunications Operators."

[2] 3GPP, "Study of enablers for Network Automation for 5G," 3GPP TR 23.791, Release 16, June 2019.

[3] 3GPP, "Management and orchestration; 5G end to end Key Performance Indicators (KPIs)," 3GPP TS 28.554, Release 19, March 2025.

[4] O-RAN ALLIANCE, "O-RAN Use Cases and Deployment Scenarios," White Paper, February 2020.

[5] O-RAN ALLIANCE, "O-RAN Architecture Description," v. 13.0, Feb. 2025.

[6] ETSI, "Experiential Networked Intelligence (ENI); System Architecture," ETSI GS ENI 005 V1.1.1, Sep. 2019.

[7] ITU-T, "Framework of artificial intelligence enhanced telecom operation and management (AITOM)," Recommendation ITU-T M.3080, Feb. 2021.

[8] Andrey Shorov, "AI/ML security in mobile telecommunication networks," Ericsson Blog, April 2024.

[9] J. Malik, R. Muthalagu and P. M. Pawar, "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies," in *IEEE Access*, vol. 12, pp. 99382-99421, 2024

[10] ITU-T, Focus Group on Artificial Intelligence Native for Telecommunication Networks (FG AINN),

[11] R. Ali and A. Imran, "AI for 5G: A Survey on Promising Machine Learning Models and Future Research Directions," IEEE Access, vol. 9, pp. 127523–127563, 2021.

[12] F. Tang et al., "On Removing Routing Protocol from Future Wireless Networks: A Real-time Deep Learning Approach for Intelligent Traffic Control," IEEE Wireless Communications, vol. 27, no. 3, pp. 118–125, 2020.

[13] M. Chen et al., "Artificial Intelligence for Wireless Networks with Applications to 5G and Beyond," IEEE Communications Surveys Tutorials, vol. 24, no. 1, pp. 452–492, 2021.

[14] N. Zhang et al., "AI-Driven Next-Generation Base Station Architecture for 6G Mobile Network," IEEE Wireless Communications, vol. 28, no. 3, pp. 102–109, 2021.

[15] Y. Sun et al., "Artificial Intelligence-Based Communication-Aware Task Offloading for Autonomous Driving in MEC," IEEE Journal on Selected Areas in Communications, vol. 40, no. 1, pp. 216–231, 2022.

[16] S. Zanzi et al., "O-RAN: Towards an Open and Smart RAN," IEEE Wireless Communications, vol. 29, no. 1, pp. 64–71, 2022.

[17] W. Wu et al., "AI-Native Network Slicing for 6G Networks," arXiv preprint arXiv:2105.08576, 2021.

[18] M. A. Habibi et al., "Machine Learning for 5G and Beyond," IEEE Access, vol. 8, pp. 133995–134017, 2020.

[19] M. Giordani et al., "Toward 6G Networks: Use Cases and Technologies," IEEE Communications Magazine, vol. 58, no. 3, pp. 55–61, 2020.

[20] X. Lin et al., "Autonomous Networks: Architecture, Enabling Technologies, and Open Issues," IEEE Communications Surveys Tutorials, vol. 23, no. 2, pp. 1027–1051, 2021.

[21] S. Samarakoon et al., "Federated Learning for Ultra-Reliable Low-Latency V2V Communications," IEEE Transactions on Communications, vol. 68, no. 6, pp. 3241–3255, 2020.

[22] Z. Zhou et al., "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," Proceedings of the IEEE, vol. 107, no. 8, pp. 1738–1762, 2019.

[23] P. Wang et al., "Slicing: A New Approach to Accelerate the Deployment of 5G," IEEE Communications Magazine, vol. 57, no. 3, pp. 56–61, 2019.

[24] N. Zhang et al., "Edge Intelligence in 6G: Vision, Enabling Technologies, and Applications," IEEE Wireless Communications, vol. 28, no. 2, pp. 12–18, 2021.

[25] M. Polese et al., "Edge Computing for 5G: A Survey," IEEE Access, vol. 8, pp. 85854–85877, 2020.

[26] A. Hassen et al., "End-to-End Network Slicing for 5G and Beyond," IEEE Communications Magazine, vol. 59, no. 3, pp. 96–102, 2021.

[27] Y. Wang, J. Wang, W. Zhang, Y. Zhan, S. Guo, Q. Zheng, X. Wang, "A survey on deploying mobile deep learning applications: A systemic and technical perspective," Digital Communications and Networks, Volume 8, Issue 1, pp. 1-17, Feb. 2022.