Intelligence Driven Wireless Networks in B5G and 6G Era: A Survey

GAO Yin[1,2], CHEN Jiajun[1,2], LI Dapeng[1,2]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. Wireless Product R&D Institute, ZTE Corporation, Shanghai 201203, China)

DOI: 10.12142/ZTECOM.202403012

https://kns.cnki.net/kcms/detail/34.1294.TN.20240903.1032.004.html, published online September 3, 2024

Manuscript received: 2024-03-11

Abstract: As the wireless communication network undergoes continuous expansion, the challenges associated with network management and optimization are becoming increasingly complex. To address these challenges, the emerging artificial intelligence (AI) and machine learning (ML) technologies have been introduced as a powerful solution. They empower wireless networks to operate autonomously, predictively, on-demand, and with smart functionality, offering a promising resolution to intricate optimization problems. This paper aims to delve into the prevalent applications of AI/ML technologies in the optimization of wireless networks. The paper not only provides insights into the current landscape but also outlines our vision for the future and considerations regarding the development of an intelligent 6G network.

Keywords: intelligent network; native AI; load prediction; trajectory prediction

Citation (Format 1): GAO Y, CHEN J J, LI D P. Intelligence driven wireless networks in B5G and 6G era: a survey [J]. ZTE Communications, 2024, 22(3): 99 – 105. DOI: 10.12142/ZTECOM.202403012

Citation (Format 2): Y. Gao, J. J. Chen, and D. P. Li, "Intelligence driven wireless networks in B5G and 6G era: a survey," *ZTE Communications*, vol. 22, no. 3, pp. 99 – 105, Sept. 2024. doi: 10.12142/ZTECOM.202403012.

1 Introduction

The rapid growth of mobile subscribers and the introduction of numerous new services have led to the continuous expansion of wireless communication networks. In addition, the diversity of network deployment and the increase of network parameters in the 5G era also make network management quite complicated. The large scale and high complexity make it infeasible to achieve the best network optimization solution by human engineers. Fortunately, the advancement of artificial intelligence (AI) and machine learning (ML) technologies has provided a powerful solution to addressing these challenges. AI and ML technologies offer efficient ways to tackle complex problems in wireless communication network management. By leveraging these technologies, the wireless communication network can be autonomous, predicted, on-demand and smart operated, and realize accurate parameter prediction, intelligent resource allocation, and green energy savings, thus greatly enhancing the network performance with less human intervention.

To apply AI/ML to the existing 5G network, the 3rd Generation Partnership Project (3GPP) has also begun the study on AI/ML topics[1–3]. In terms of the radio access network (RAN), the enhancement of data collection for intelligence has been studied, including the high-level principles, the functional framework, and scenarios (network energy saving, load balancing, and mobility optimization) for AI-assisted network optimization. The RAN3 work group further discusses the corresponding normative work in Release-18 to enhance the collection of measurement through signaling based on the existing next-generation (NG)-RAN interfaces and architecture. In addition, the research on the intelligence of air interface is also carried out by the RAN1 work group, which studies the life-cycle management, scenarios such as channel state information (CSI) feedback enhancements, beam management and positioning improvement, evaluations for each use case and potential impact to the current specification.

This paper presents the popular application of AI/ML techniques in wireless network optimization and provides our future vision and consideration on a 6G intelligent network. The subsequent sections of this paper are outlined as follows. In Section 2, we review some achievements in AI/ML assisted wireless communication network optimization. Section 3 describes the potential implementation of AI/ML based use cases over existing network architecture. Finally, future vision and consideration are provided in Section 4, followed by the conclusion of this paper in Section 5.

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62201266, and in part by the Natural Science Foundation of Jiangsu Province under Grant No. BK20210335.

2 Applications of AI/ML Techniques in Wireless Network Optimization

In wireless communication systems, AI/ML algorithms have been extensively used for various usage such as traffic load prediction, mobility prediction, radio link failure prediction[4], positioning[5], and network slicing resource management[6]. Various experiments or practices have been carried out to show the potential benefits of AI/ML-assisted wireless network optimization.

2.1 Network Energy Saving

Energy conservation has always been a global and eternal topic in various walks of life. Especially in the mobile communication industry, the relentless expansion of mobile communication networks to meet the demands of an unprecedented surge in mobile subscribers has resulted in a rapid increase in energy consumption. To reduce the huge energy consumption and achieve a greener mobile communication network, numerous research projects have been started with different contribution areas like services, architecture, and intelligence during the past years[7].

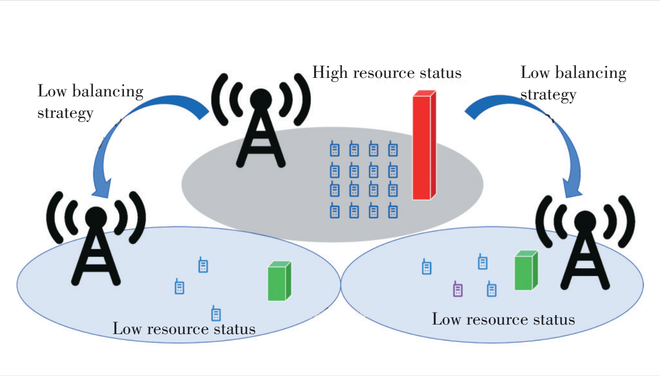
Numerous energy-saving strategies, including symbol shutdown, carrier shutdown, channel shutdown, deep sleep, and symbol aggregation, have been proposed[8]. However, existing energy conservation is usually vulnerable due to some potential issues such as imprecise traffic prediction, imbalance between performance and efficiency, inflexible parameter adjustment, and localized energy efficiency improvement leading to an overall deterioration[1]. AI/ML based energy-saving strategies can realize accurate load prediction, flexible parameter optimization, service forecasting, and scene identification, to select the most suitable shut-down schemes for a certain scenario without significantly deteriorating the system performance.

In Ref. [9], the energy-saving performance of strategies including symbol switch-off, channel switch-off, and carrier switch-off with the assistance of Auto Regressive Integrated Moving Average (ARIMA) is evaluated in real scenarios. The results show that the machine learning technique can bring the percentage of the switch-off duration per cell up to 176%, increase the switch-off duration and 1.24 kWh power saving per cell per day without affecting basic key performance indicators (KPIs), and the total electrical saving per week is increased by CNY 2 223 compared with the conventional energy saving strategies[9]. AI/ML based channel shutdown, symbol shutdown and deep sleep are used in the real test involving 54 active antenna units (AAU) for energy saving, resulting in a reduction of 23.87% in power consumption and an improvement of 23.4% in energy efficiency[10]. Ref. [11] illustrates an energy-efficiency optimization algorithm through deep reinforcement learning (DRL) by simulation, and shows that the DRL-assisted energy-saving algorithm can bring about 50 W or 40% power savings compared with the initial system. AI-based service awareness, capable of discerning variations in the energy efficiency of different service types, is integrated with AI-based traffic forecasting for the optimization of energy-saving strategies. This approach results in a daily energy saving of 13.7 kWh, with energy saving increasing by nearly 10%[12].

2.2 Traffic Load Prediction

Traffic prediction plays a pivotal role in network optimization and serves various use cases, including energy conservation and mobile load balancing. Traditional models such as linear regression and support vector machines (SVM) have reached maturity in traffic load prediction. With the quick evolution of deep learning, increasingly sophisticated algorithms are being employed for more accurate predictions. Load prediction plays a significant role in assisting load balancing as illustrated in Fig. 1.

Several supervised and unsupervised learning algorithms for predicting the resource status on sites are compared in terms of accuracy, time consumption, and memory usage by using the real-world datasets of the wireless Long-Term Evolution (LTE) network, and the results show that the Automated Neural Net (ANN) has the highest prediction accuracy about 80%, and the SVM and Self-Organizing Maps (SOM) can also provide above 70% accuracy[13]. Ref. [14] proposes a new type of federated learning (FL) mechanism to solve data security and privacy issues in the commonly-used-centralized training models, and the prediction accuracy can reach 86.02%, close to the state-of-the-art FL models while significantly reducing the communication cost. Ref. [15] introduces two models, Ensemble and ResNet, for traffic load prediction, and compares their prediction performance in the same scenario with ARIMA and Prophet as the baseline. The result states that the prediction accuracy of the ensemble model which takes time, space, and historical information into consideration is much higher than ARIMA and Prophet. The calculation complexity of ResNet is significantly lower than that of baseline models, as it can generate results for all cells in a single training session. This characteristic makes it particularly suitable for traffic load prediction tasks involving large datasets.



▲Figure 1. Benefits to load balancing with assistance of artificial intelligence (AI)/machine learning (ML)

2.3 UE Trajectory Prediction

The ultra-dense network deployment in the 5G mobile communication system consists of numerous small cells to satisfy the requirements of ultra-high reliability, low latency, and high data rate. This can lead to more frequent handovers for high-mobility UE, thus resulting in problems such as high latency, throughput reduction, radio link failure (RLF), and ping-pong effect. Therefore, mobility optimization is of vital importance, and one of the key parts of mobility optimization is UE trajectory prediction. Since human mobility is predictable to some extent, it is feasible to analyze user's mobility patterns through their history trajectory information[16], and can be enhanced by using machine learning techniques.

Ref. [16] learns the mobility pattern of user equipment (UE) from historical trajectories and predicts its future movement trends using the Long Short Term Memory (LSTM) structure, and the prediction result is used in the proposed intelligent dual connectivity mechanism for handover optimization. The simulation results demonstrate that, even as cell density increases, the average handover prediction accuracy for low-speed users remains high[16]. Ref. [17] proposes multiple features that combine UE history trajectory with the reference signal receiving power (RSRP) measurement reports from serving and neighbor base stations as input of the sequence-to-sequence model for next-time location prediction, and introduces orientation loss function to analyze the direction of movement. Simulation results show that with RSRP and orientation loss function, the average distance error decreases from 48.634 m to 38.457 m, and the accuracy of predicted connected nodes the next time can be up to 98.26%[17]. Ref. [18] compares the performance of various models for AI-based mobility prediction, and Bidirectional Long Short Term Memory (Bi-LSTM)-attention shows the highest accuracy up to 91.78%, while ANN consumes the shortest training time due to its low complexity.

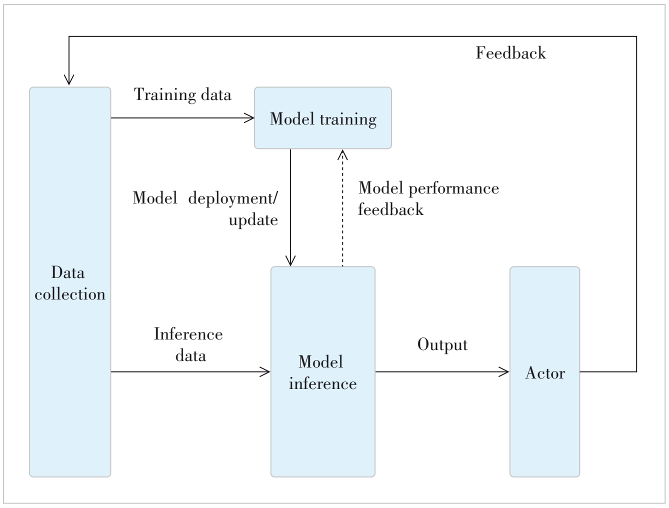
3 Implementation of AI/ML Based Use Cases over Network Architecture

The key to integration between AI/ML and wireless networks is to resolve the issue of how to implement the AI/ML based use cases over existing network architecture. Existing network architecture includes next-generation radio access network (NG-RAN) nodes, UE, and core networks, and each network entity can be in charge of different functions to support AI/ML, e. g., training and inference. Different scenarios may call for various deployment methods for AI/ML models.

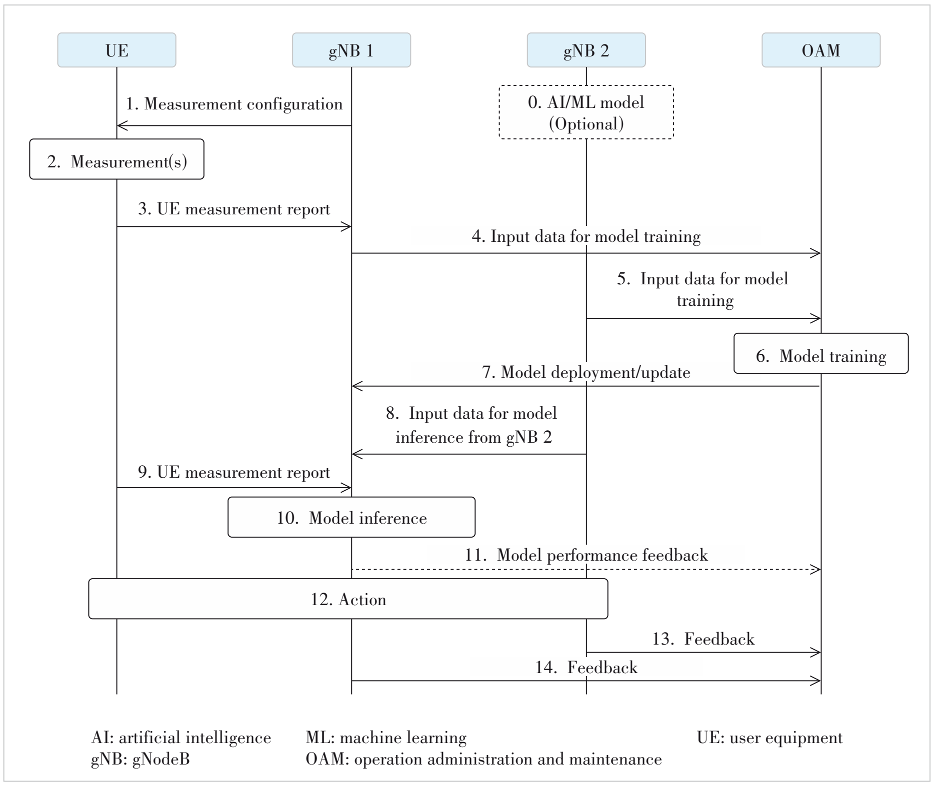
A framework for RAN intelligence (Fig. 2) has been described in 3GPP TR 37.817[1]. The data collection function collects different kinds of data for the AI/ML model, such as measurements from UE or gNBs, predictions or decisions output from AI/ML models, and feedback from the actor, and provides the input data for training and inference function. Input data can be used after the data preparation procedure, such as data pre-processing, is performed by the model training function and model inference function. The model training function trains, validates and tests the AI/ML model, which will be deployed to the model inference function after these procedures. The model inference function generates predictions or decisions and a trained AI/ML model from the model training function. Additionally, it offers model performance feedback to monitor and optimize model performance. The actor function executes actions based on decisions made by the model inference function and provides feedback to the data collection.

Some enhancements are needed in the current 5G RAN to integrate AI/ML functions. From the perspective of specification, three use cases, i.e., network energy saving, load balancing, and mobility optimization, are first considered to be standardized for supporting AI/ML functions. For these three AI/ML based use cases, model training can be located in either operation administration and maintenance (OAM) or gNB, while the model inference is located in gNB. In the case of centralized unit/distributed unit (CU-DU) split RAN architecture, model training can be located in either OAM or gNB-CU, while the model inference is located in gNB-CU.

Fig. 3 shows the general flow chart of an AI-based use case with model training at OAM and model inference at gNB. UE is currently served by gNB 1, while gNB 2 can be the neighbouring gNB optionally with an AI/ML model. The OAM collects the input data needed for model training, including the measurement report of UE, input data from serving gNB and neighbouring gNB, and performs model training. The trained AI/ML model is then deployed/updated into the gNB 1 (this step is out of the RAN3 Rel-17 scope) for further training or model inference. Based on the local input data from gNB 1 and other indicated input data from UE and gNB 2, gNB 1 performs model inference to make decisions or predictions, and this output can also be the model performance feedback sent to the OAM. The gNB 1 executes the action based on the model inference output and provides feedback to OAM for



▲Figure 2. Functional framework for radio access network (RAN) intelligence



▲**Figure 3. Deployment of AI/ML functionality: model training at OAM and model inference at next-generation radio access network (NG-RAN)**

requesting node requests the reporting of predictions by sending the Data Collection Request message and configures the requested prediction time (a specific point of time in the reasonable future for which the prediction information is requested) as t_p. The requested node makes predictions based on its own AI/ML model and reports the successfully initiated predictions in the Data Collection Update message to the requesting node for one time at t_{Report}, which is a time point ahead of t_p. For periodic reporting, the reporting periodicity is also configured in the Data Collection Request message. The requested predictions will be reported every T_R at time points of $t_{\text{Report}} + N \times T_R$ (N = 0, 1, 2,⋯), corresponding to the requested prediction time of $t_p + N \times T_R$, until the requesting node sends the Data Collection Request message to stop the report.

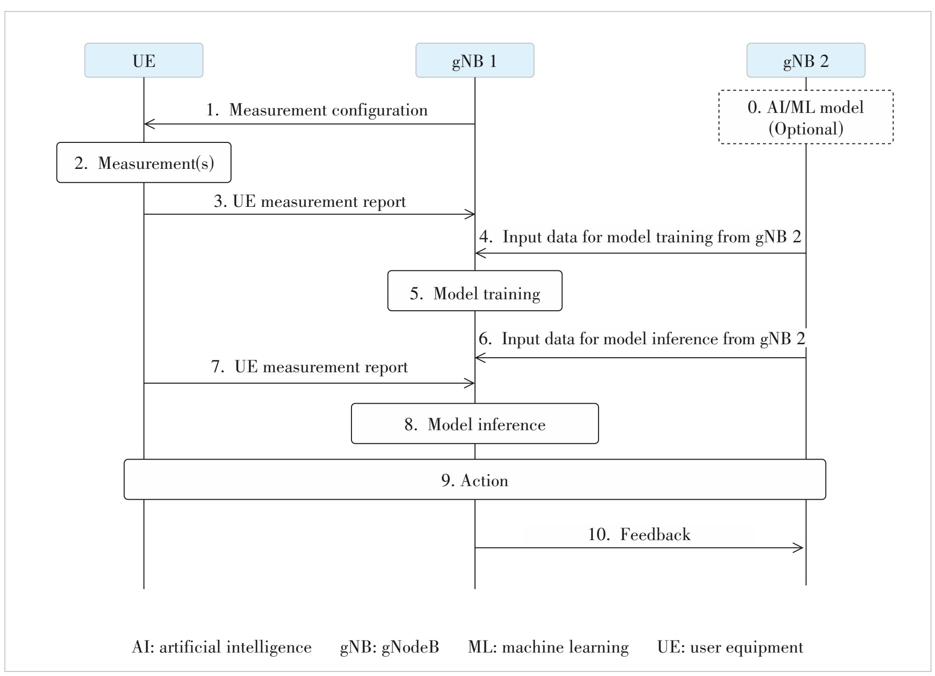
Meanwhile, UE performance feedback (including average UE throughput DL/UL, average packet delay, and average packet loss), measured UE trajectory, and energy cost are considered measurements to support AI/ML functions, such as model performance evaluation.

model performance monitoring and training optimization.

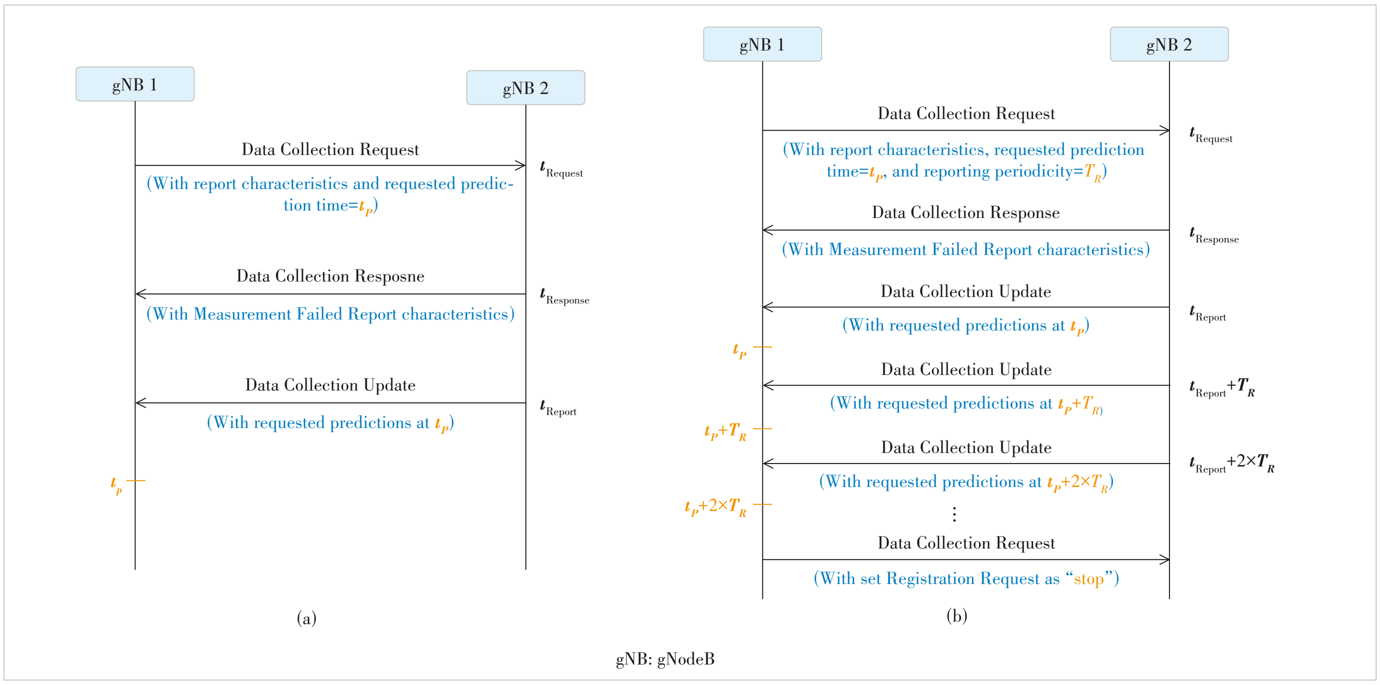
Fig. 4 shows the general flow chart of an AI-based use case with model training and model inference at gNB. It can be seen that the overall procedure is similar to the previously introduced flow chart. The difference is that since the model training and model inference are both performed in gNB 1, the input data for model training and model inference and the feedback after the action are directly sent to gNB 1.

Across the three AI/ML-based use cases, the predicted resource status, the predicted number of active UEs, predicted radio resource control (RRC) numbers and predicted UE trajectory are considered as types of predicted assistance information to be reported between NG-RAN nodes.

Take predicted resource status information as an example. It can be reported at one time or periodically between NG-RAN nodes. The flow charts of one-time reporting and periodic reporting are shown in Figs. 5(a) and 5(b). In one-time reporting, the



▲ **Figure 4. Deployment of AI/ML functionality: model training and model inference at next-generation radio access network (NG-RAN)**



▲**Figure 5. Flowcharts of transferring predicted information: (a) one-time reporting; (b) periodic reporting**

Take the UE performance feedback as an example. It can also be reported at one time or periodically, as illustrated in Figs. 6(a) and Fig. 6(b). Since the UE performance is the average information over a period of time measured at the traffic offloaded neighbouring gNB, a new data collection ID, IE, is included in the Handover Request message as a trigger indication to request the measurement of UE performance at the target gNB after the successful handover, while the configuration of measurement and reporting is still indicated in the Data Collection Request message. In one-time reporting, the target gNB starts the UE performance measurement collection after the successful handover until the measurement collection duration expires, and reports the measured UE performance for one time to the source gNB. In periodic reporting, for one pair of measurement IDs, the reporting periodicity is calculated from the egress of the Data Collection Response message, namely, the UE performance feedback is reported through the Data Collection Update message every T_R at time points of $t_{\text{Response}} + N \times T_R$, which can effectively avoid the signaling storm caused by the UE handed over at different times.

With the introduced solutions to supporting AI/ML functions over the Xn interface between NG-RAN nodes, the RAN node can infer future information, aiding operators in optimizing their network and enhancing the user experience.

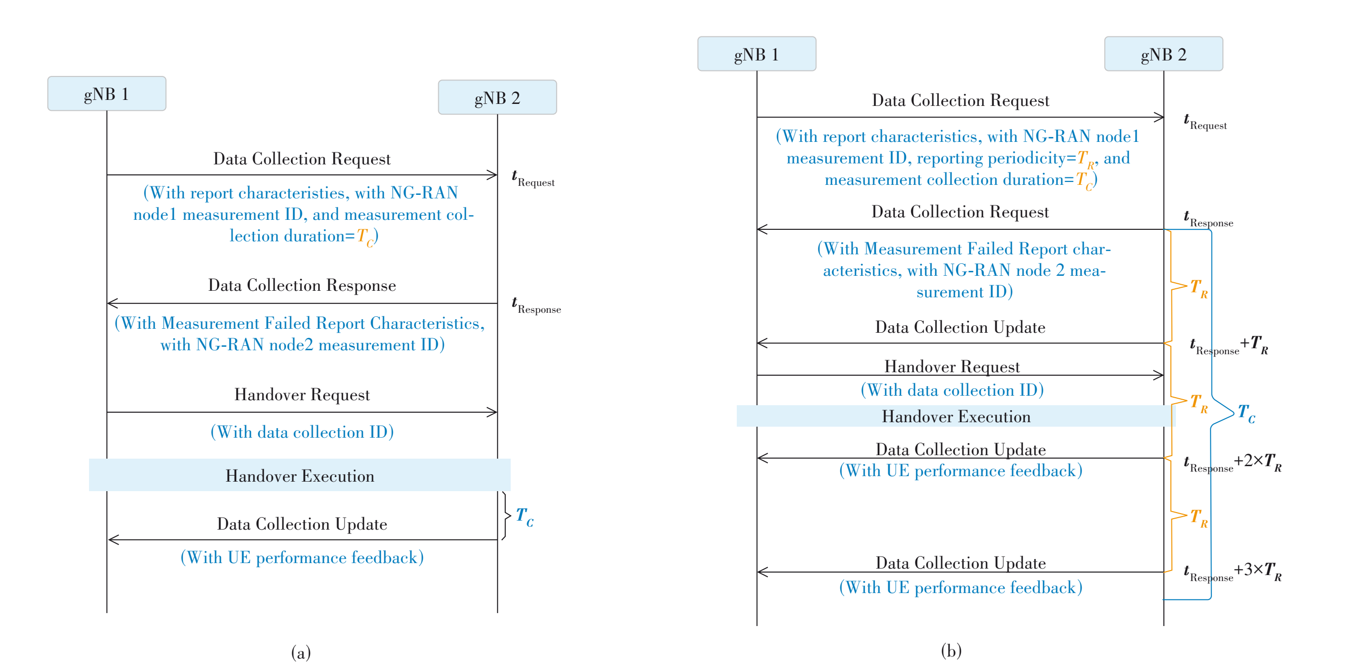
4 Future Vision on AI/ML Assisted Wireless Network

The development of the 6G network is currently a vigorously researched topic in both the telecommunications industry and the academic community. For 6G, there are higher and more requirements compared to 5G mobile networks in terms of key objectives, such as coverage, speed, latency, capacity, AI, integrated sensing and communication (ISAC), and computing. From our perspective, 6G networks need to achieve seamless human-machine and machine-machine interactive communications, while humans are at the center of control and judgment. Inter-working between humans and machines will become more frequent and broader in the future, not only for the devices bought by people, such as wearable devices and sensing devices, but also for those variant devices in society and industry, e. g., cameras, vehicles, robots, and unmanned aerial vehicles (UAVs). Such collaborative intelligent interaction can be achieved by AI/ML tools based on the amount of data perception, while machine cognition must be handled carefully. During human-machine interaction, three key points need to be considered: intelligence, energy efficiency, and security, as illustrated in Fig. 7.

1) Intelligence

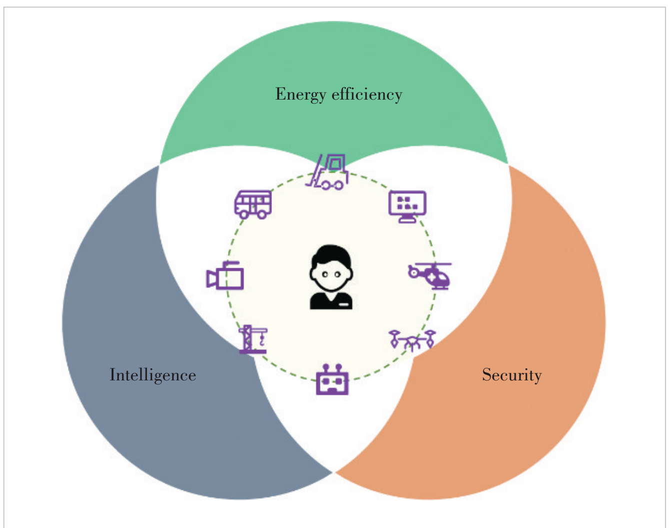
In the realm of artificial intelligence, the current focus of 5G is to treat AI as a tool to assist networks in operations such as load balancing, energy efficiency, and mobility optimization. However, with the advent of 6G, the consideration goes beyond treating AI/ML merely as a tool. Instead, the focus shifts towards deeply integrating AI/ML functionalities into the network layer, aiming to achieve native AI.

From now on, artificial intelligence can be divided into data intelligence, perception intelligence, cognitive intelligence, and autonomous intelligence. Data intelligence refers to the ability of computing hardware to analyze and categorize data stored, which can be considered the most fundamental level of



NG-RAN: next-generation radio access network UE: user equipment

▲**Figure 6. Flow charts of transferring UE performance feedback: (a) one-time reporting; (b) periodic reporting**



▲ **Figure 7. 6G network: inter-working between human and machine for sustainability**

intelligence. Perception intelligence means that computing units have perceptual capabilities to recognize diverse information, such as videos, images, and sounds. From an implementation and security perspective, perception intelligence will become an efficient level of intelligence in 6G networks. By perceiving and analyzing data from various network elements and layers (including communication quality, user experience, use case requirements, etc.), it will make a large number of human-machine interactions more efficient.

2) Energy efficiency

Energy efficiency is one of the key concerns for operators, whether in the current 5G or future 6G networks. For 6G, as there will be an introduction of a larger number of terminal devices and the need to support computing capabilities for AI/ML functions, the substantial increase in data transmission can lead to a significant rise in energy consumption. This increased energy consumption can adversely affect the sustainability of 6G networks and result in a considerable number of carbon emissions. Therefore, energy-saving efficiency strategies need to be further developed. This may involve using artificial intelligence to predict traffic volume and minimize energy consumption or treating energy services as a specific service criterion.

3) Security

6G security aims to ensure that systems are protected against unintended and unauthorized access, safeguarding personal data and sensitive network information. Enhanced encryption algorithms can be used to protect the privacy of data during transmission and storage. In addition, federated learning is leveraged to use the local training and global training mechanisms to protect the privacy data from UE. Blockchain processes data through decentralization and uses distributed data management to protect user privacy.

5 Conclusions

AI/ML-enabled RAN intelligence has the potential to significantly enhance network performance and user experience. This paper aims to delve into a comprehensive overview of achievements in optimizing wireless communication networks through the application of AI/ML techniques. Additionally, the paper provides an overview of the implementation of AI/ML based use cases over existing network architecture. The inclusion of this aspect underscores the importance of aligning AI/ML advancements with industry standards, ensuring seamless integration and widespread adoption. As computational capabilities continue to strengthen, diverse application scenarios emerge, and standardization progresses, the paper anticipates an escalating role for AI/ML techniques in shaping the landscape of 5G and the imminent 6G era. The convergence of these factors positions AI/ML as a pivotal force, poised to drive innovation and efficiency in next-generation wireless communication networks.

References

[1] 3rd Generation Partnership Project (3GPP). Study on enhancement for data collection for NR and EN-DC [R]. 3GPP R 37.817, 2022

[2] CHEN W S, LIN X Q, LEE J, et al. 5G-advanced toward 6G: past, present, and future [J]. IEEE journal on selected areas in communications, 2023, 41 (6): 1592 – 1619. DOI: 10.1109/JSAC.2023.3274037

[3] 3rd Generation Partnership Project (3GPP). Study on artificial intelligence (AI)/machine learning (ML) for NR air interface [R]. 3GPP TR 37.843, 2023

[4] BOUTIBA K, BAGAA M, KSENTINI A. Radio link failure prediction in 5G networks [C]//Proc. IEEE Global Communications Conference (GLOBECOM). IEEE, 2021: 1 – 6. DOI: 10.1109/GLOBECOM46510.2021.9685587

[5] BUTT M M, PANTELIDOU A, KOVÁCS I Z. ML-assisted UE positioning: performance analysis and 5G architecture enhancements [J]. IEEE open journal of vehicular technology, 2021, 2: 377 – 388. DOI: 10.1109/OJVT.2021.3110134

[6] DEBBABI F, JMAL R, CHAARI L, et al. Overview of AI-based algorithms for network slicing resource management in B5G and 6G [C]//Proc. International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2022: 330 – 335. DOI: 10.1109/IWCMC55113.2022.9824988

[7] LARSEN L M P, CHRISTIANSEN H L, RUEPP S, et al. Toward greener 5G and beyond radio access networks: a survey [J]. IEEE open journal of the communications society, 2023, 4: 768 – 797. DOI: 10.1109/OJCOMS.2023.3257889

[8] TAN R M, WU T, SHI Y, et al. Intelligent energy saving solution of 5G base station based on artificial intelligence technologies [C]//Proc. IEEE International Joint EMC/SI/PI and EMC Europe Symposium. IEEE, 2021: 739 – 742. DOI: 10.1109/EMC/SI/PI/EMCEurope52599.2021.9559261

[9] GAO Y, CHEN J J, LIU Z, et al. Machine learning based energy saving scheme in wireless access networks [C]//Proc. International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2020: 1573 – 1578. DOI: 10.1109/IWCMC48107.2020.9148536

[10] CHEN J J, GAO Y, LIU Z, et al. Future vision on artificial intelligence assisted green energy efficiency network [J]. ZTE communications, 2023, 21 (2): 34 – 39. DOI: 10.12142/ZTECOM.202302006

[11] GIANNOPOULOS A, SPANTIDEAS S, KAPSALIS N, et al. Supporting intelligence in disaggregated open radio access networks: architectural principles, AI/ML workflow, and use cases [J]. IEEE access, 2022, 10: 39580 – 39595. DOI: 10.1109/ACCESS.2022.3166160

[12] TAN R M, SHI Y, FAN Y Y, et al. Energy saving technologies and best practices for 5G radio access network [J]. IEEE access, 2022, 10: 51747 – 51756. DOI: 10.1109/ACCESS.2022.3174089

[13] SHRIVASTAVA P, PATEL S. Selection of efficient and accurate prediction algorithm for employing real time 5G data load prediction [C]//Proc. IEEE 6th International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2021: 572 – 580. DOI: 10.1109/ICCCA52192.2021.9666235

[14] NAN J H, AI M, LIU A J, et al. Regional-union based federated learning for wireless traffic prediction in 5G-Advanced/6G network [C]//Proc. IEEE/CIC International Conference on Communications in China (ICCC Workshops). IEEE, 2022: 423 – 427. DOI: 10.1109/ICCCWorkshops55477.2022.9896702

[15] GAO Y, ZHANG M, CHEN J J, et al. Accurate load prediction algorithms assisted with machine learning for network traffic [C]//Proceedings of International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2021: 1683 – 1688. DOI: 10.1109/IWCMC51323.2021.9498910

[16] WANG C, ZHAO Z, SUN Q, et al. Deep learning-based intelligent dual connectivity for mobility management in dense network [C]//IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018: 1 – 5. DOI: 10.1109/VTCFall.2018.8690554

[17] GAO Y, CHEN J J, LIU Z, et al. Deep learning based location prediction with multiple features in communication network [C]//Proc. IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2021: 1 – 5. DOI: 10.1109/WCNC49053.2021.9417278

[18] WANG W H, DUAN X Y, SUN W F, et al. Research on mobility prediction in 5G and beyond for vertical industries [C]//Proc. IEEE/CIC International Conference on Communications in China (ICCC Workshops). IEEE, 2021: 379 – 383. DOI: 10.1109/ICCCWorkshops52231.2021.9538889

Biographies

GAO Yin (gao.yin1@zte.com.cn) received her master's degree in circuit and system from Xidian University, China in 2005. Since 2005, she has been engaged in the study of 4G/5G technology and is currenly a wireless expert and project manager at the research center of ZTE Corporation and the State Key Laboratory of Mobile Network and Mobile Multimedia Technology, China. She has authored or co-authored hundreds of proposals for 3GPP meetings and journal papers in wireless communications and has filed more than 200 patents. She has also worked as 3GPP RAN3 Chair since 2017 and was elected as RAN3 Chair in May 2021.

CHEN Jiajun received his master's degree in communication engineer from Shanghai University, China in 2019. He is currently a 5G research engineer at the R&D center of ZTE Corporation and the State Key Laboratory of Mobile Network and Mobile Multimedia Technology, China. His research interests include 5G wireless communications and artificial intelligence.

LI Dapeng received his MS degree in computer science from University of Electronic Science and Technology of China in 2003. He is currently a senior researcher at ZTE corporation and mainly focuses on research and implementation of wireless access network systems.