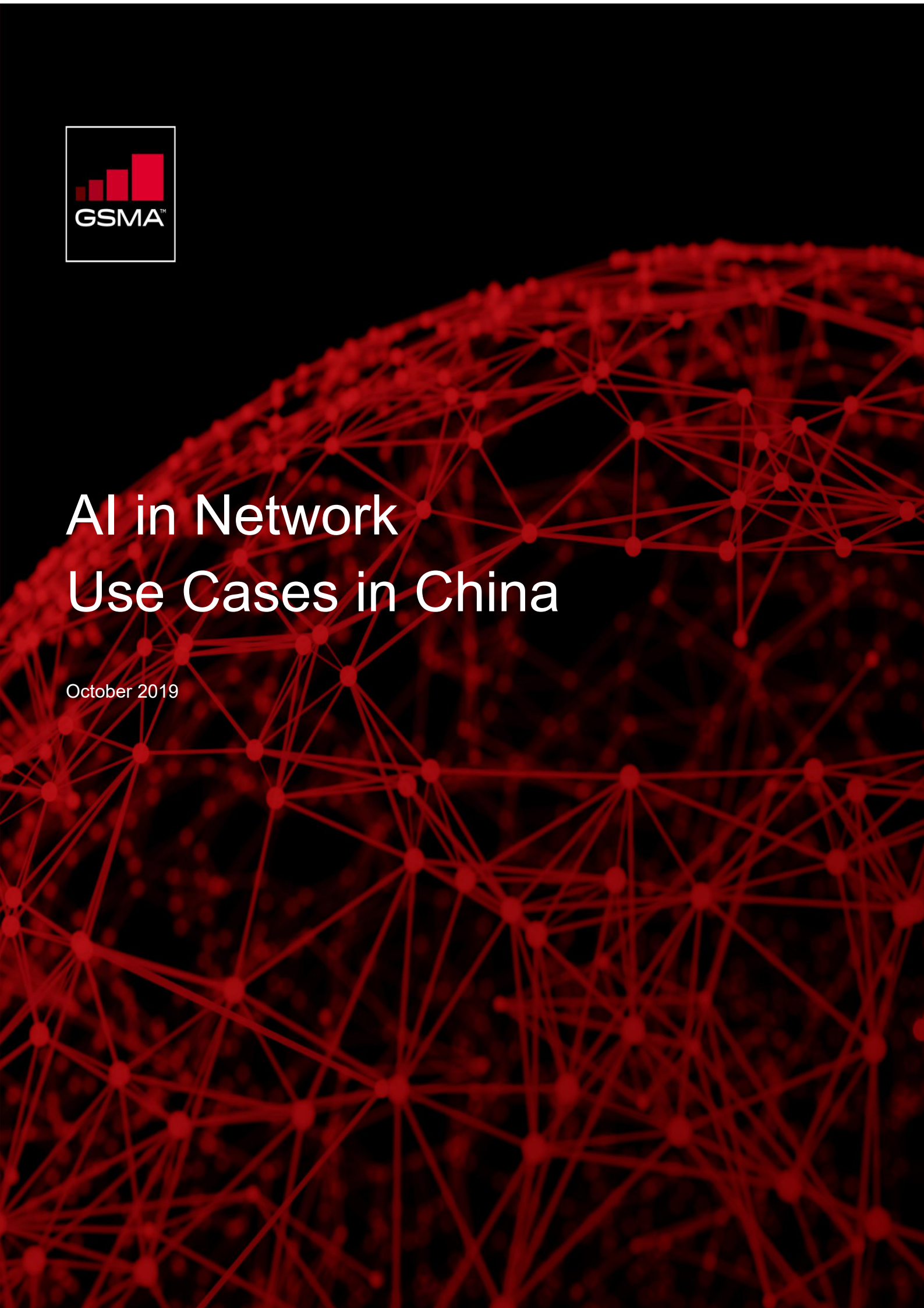



# AI in Network Use Cases in China

October 2019

# Table of Contents

# 1 Intelligent Autonomous Network: an Important Component of 5G

As mobile communication enters the 5G era, new technologies, features, services, and applications emerge one after another. The traditional telecom network operation and management mode do not meet the increasing requirements for network evolution, service development, user experience, and operation analysis. Also, the traditional mode is not able to effectively improve network operation efficiency and control the operating costs. The industry has realised that the 5G era requires a highly intelligent, automated network, followed by an intelligent autonomous network. Intelligent autonomy is an essential enabler for innovative business models of mobile communications and will become the essential element of mobile communications networks in the post-5G era. The introduction of AI into mobile networks will be an inevitable requirement for network design, deployment, operation, assurance, and optimisation in the 5G and post-5G eras.

To implement intelligent autonomous networks, the entire industry needs to:

- Have a unified understanding of intelligent autonomous networks.
- Clearly define the concepts of intelligent autonomous networks.
- Specify each development phase and objective of intelligent autonomous networks.
- Jointly incubate intelligent autonomous network use cases.

The convergence of AI and communications networks will inject new technological vitality into communications networks, and open unprecedented possibilities. AI in Networks is the key to success, truly promoting the GSMAs Intelligent Connectivity vision, connecting everyone and everything to a better future.

## 1.1 New Challenges and Opportunities in the 5G Era

5G networks are currently being deployed around the world. Compared with 4G networks, 5G networks have a qualitative leap in key performance indicators (KPIs) such as the transmission rate, transmission delay, and connection scale. Therefore, 5G networks can support more diverse service scenarios and applications. However, 5G networks also bring the challenges of increasing CAPEX and OPEX for mobile operators.

**O&M Mode Innovation Required by Increasingly Complexity**

To support the eMBB, mMTC, and URLLC typical service scenarios and ensure excellent network performance, new technologies with high complexity, such as massive MIMO and flexible air interface, are introduced to 5G networks. These technologies enable 5G networks to meet more stringent technical specifications, such as the peak rate, spectrum efficiency, low delay, high reliability, and connection density. The 5G core network, constructed based on the virtualisation and cloud concepts, requires flexible resource scheduling and diversified network entities and interfaces. The 5G core network also requires unified network scheduling and management. As a result, the problems of the traditional network O&M mode become more apparent. Besides, 5G networks will coexist with existing 2G, 3G, and 4G networks for a long time, which brings unprecedented challenges to 5G network O&M as well.

**Accelerated Service Innovation Raising Higher Requirements for Network Intelligence**

In the 4G era, mobile traffic usage increases exponentially, but the unit price of traffic per bit decreases continuously. In the 5G era, traditional data services alone can hardly bring new revenue growth. Major

service innovation will mainly relate to the the digital transformation of other industries. Mobile operators will be required to transform their business models and greatly enhance network flexibility to meet user requirements and service operation requirements.

In this 5G era, the single-mode communication between people will gradually evolve into full-scenario communication that includes Person-to-Person, Person-to-Machine, and Machine-to-Machine. Therefore, service scenarios will become more complex. Many new service scenarios raise differentiated requirements for SLA (Service Level Agreements), complicating network operation.

Based on 5G network capabilities and rich service development, service experience will be diversified and personalised. Such service experience examples include immersive experience, real-time interaction, accurate perception of emotion and intent, and 'you get what you want'. The traditional mode of network support for user experience will need to be reshaped.

**Introducing AI Technologies to Effectively Cope with Related Challenges and Bring New Opportunities**

5G networks have a large amount of available data, including transport layer data (channel, spectrum, and customer link), network layer data (signalling and management data), and various types of application layer data. Based on such data, operators can leverage AI technologies to cope with the challenges of 5G networks.

AI technologies can be introduced to:

- Implement big data analysis and adaptive policy decision-making.
- Optimise the automation solution.
- Understand and predict user and network requirements.
- Implement better resource orchestration and scheduling.
- Realise the goal of a complete intelligent autonomous network.

Intelligent autonomous networks have the following advantages:

- Reduce network construction and operation costs.
- Respond to user and service requirements more precisely.
- Innovate business models.
- Bring huge opportunities for operators.

5G network planning and design should quickly and concisely reflect operators intention. Service provisioning requires fewer manual configuration errors and quick service rollout. In terms of network O&M, breakthroughs need to be made to remove the limitations of the traditional expert-experience-based O&M mode. The automatic network operation capability will become the indispensable 4th dimension of the 5G era together with eMBB, mMTC, and URLLC, and become one of the most important driving forces for 5G service innovation and development.

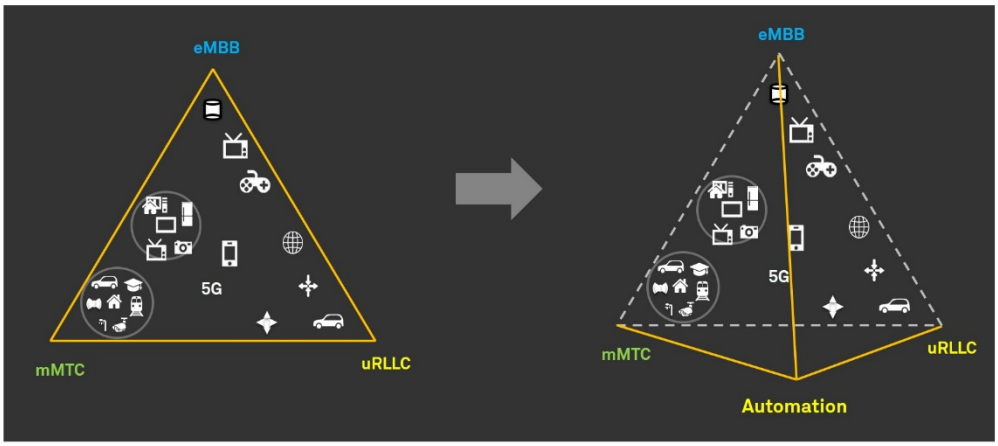


Figure 1: Network automation is the 4th dimension of 5G networks

## 1.2 AI Technologies is in its Prosperous Development Phase

AI is a technical science that studies and develops theories, methods, technologies, and applications for simulating and extending human intelligence. Since the birth of AI in 1956, AI theories and technologies have become increasingly mature, and AI application fields have been expanding.

The purpose of AI research is to enable intelligent machines to:

- Listen: speech recognition and machine translation.
- Watch: image recognition and text recognition.
- Speak: speech synthesis and human-machine dialogue.
- Think: man-machine chess playing and theorem proving.
- Learn: machine learning and knowledge representation.
- Act: robot and self-driving car.

The AI development process is similar to the exploration of the unknown; so far, the AI development process is divided into the following six phases:

1. Initial development phase: from 1956 to the early 1960s

   After the concept of AI was proposed, several remarkable research results were achieved, such as machine theorem proving and checker-playing program, which set off the first climax of AI development.

2. Reflection and development phase: from the 1960s to the early 1970s

   The breakthroughs at the early stage of AI development raised peoples expectations for AI. People started to try more challenging tasks and set unrealistic R&D goals. However, successive failures (including goal achievement failures) caused AI development to enter the trough.

3. Application development phase: from the early 1970s to the mid-1980s

   The expert system in the 1970s stimulated the knowledge and experience of human experts to solve problems of specific fields. This system made a significant breakthrough in the transformation of AI from theoretical research to practical applications and from the exploration of general reasoning strategies to some specific applications. The expert system achieved successes in the fields of medical treatment, chemistry, and geology, and thus pushed AI into a new high tide of application development.

4. Slow development phase: from the mid-1980s to the mid-1990s

The following problems with the expansion of AI application scale were gradually exposed:

- Narrow application fields
- Lack of common knowledge
- Difficulties in obtaining knowledge
- Single reasoning method
- Incompatibility with existing databases

5. Steady development phase: from the mid-1990s to 2010

   With the development of network technologies, especially Internet technologies, AI innovation and research were accelerated, making AI technologies more practical. In 1997, IBMs Deep Blue supercomputer won the chess world champion, Kasparov. In 2008, IBM proposed the concept of Smart Planet. These are the landmark events of this phase.

6. Prosperous development phase: from 2011 till now

   With the development of information technologies such as big data, cloud computing, Internet, and IoT, computing platforms such as graphics processors have driven the rapid development of AI technologies represented by deep neural networks. The technical gap between science and applications is substantially narrowed. From hard-to-use technologies to usable and human-surpassing technologies, AI technologies such as image classification, voice recognition, knowledge Q&A, man-machine chess playing, and autonomous driving usher in a new boom. When AlphaGo Zero started from a piece of white paper and beat the AlphaGo that beat top Go players like Li Shishi and Ke Jie, with a score of 100 to 0 through self-learning in a few days, self-evolving, self-perfecting, and universal AI seems to be no longer a dream.

In all fields where a large amount of training data is available, the application of AI technologies has been fruitful. Massive data generated by mobile networks every day is the basis for applying AI technologies to mobile networks. AI has natural advantages in analysing large-scale data, mining cross-field features, and generating effective policies. AI technologies will gradually bring the capabilities of listening, watching, speaking, thinking, learning, and acting into mobile networks, and implement intelligent autonomous networks gradually.

## 1.3 Operators' Requirements for Intelligent Autonomous Networks

Currently, leading global operators regard AI as one of their key strategies. For operators in China, China Mobile released a unified AI R&D cloud platform, dedicated to large-scale AI applications in the network, market, service, security, and management fields. China Telecom released *China Telecom AI Development White Paper 1.0*, clarifying the intelligent transformation from the network on-demand to the intent-based network. China Unicom is building smart, agile, intensive, and open networks and unified network AI capability platforms and is developing network planning, design, construction, maintenance, and optimisation models to support network operation services. International operators such as AT&T and Vodafone are also actively deploying intelligent autonomous networks.

**AI Planning of China Mobile**

China mobile has released its self-developed JiuTian AI Platform, which is an integral part of China mobiles 5G+AICDE plan . The platform integrates the excellent AI capabilities of China Mobiles internal and external and supports AI R&D in various fields. The platform connects to the centralised big data platform which provides rich, high-quality, and tagged AI sharing training databases for the entire company and supports centralised management of multi-domain massive data. In this way, computing power, data, and capabilities, shared across the entire China Mobile Group. I In the future, China Mobile

will actively build an AI ecosystem which will serve the network, market, service, security, management and other fields, and empower all significant vertical industries to promote the development of the AI industry. China Mobile is committed to becoming an AI application pioneer and AI industry enabler.

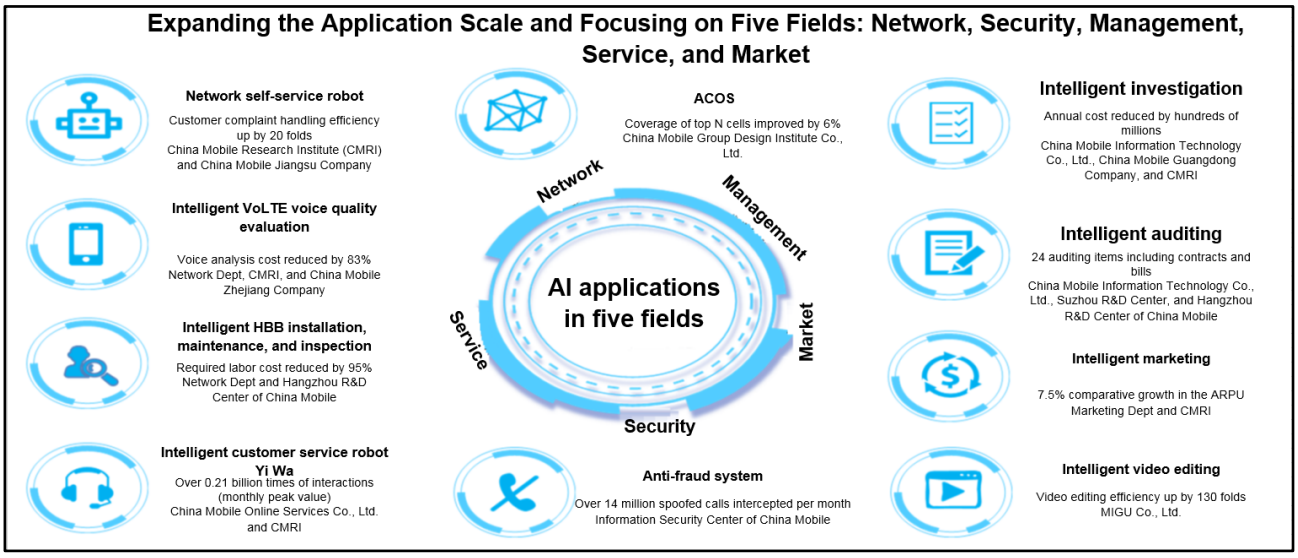


Figure 2: AI application planning of China Mobile

**Network AI Development Strategy of China Telecom and its AI development Layout**

In 2016, China Telecom officially released its transformation strategy (Transformation 3.0) and network architecture reconstruction plan (CTNet2025) to adapt to the global development trend of the intelligent era. By leveraging network intelligence, service ecology, and smart operation, China Telecom has expanded comprehensive intelligent information services to help boost national cyber development and serve the peoples livelihood.

During the steady progress of network reconstruction and the 5G pilots, the traditional manual O&M mode cannot meet the requirements of future network operation any more. In early 2017, China Telecom, together with Huawei and other partners, set up the network AI industry-standard workgroup (ENI) in ETSI. This workgroup aims to jointly promote the formulation of research and standards on AI application scenarios, requirements, and system architecture in operators networks.

With the development of network AI research and practice, China Telecom released *China Telecom AI Development White Paper 1.0* during MWC Shanghai 2019. This white paper clearly states that China Telecom will focus on strategic transformation 3.0, and deeply embed AI technology capabilities and provide common AI capability platforms, applications, and solutions based on the CTNet2025 architecture. In the initial phase, cloud-network convergence is used as the breakthrough point to provide public, government, and enterprise users with quickly provisioned, customised, automated, and multi-layer intelligent services. In the future, the smart, E2E DICT solutions and services will be gradually provided. The following figure shows the overall layout of China Telecoms AI development.

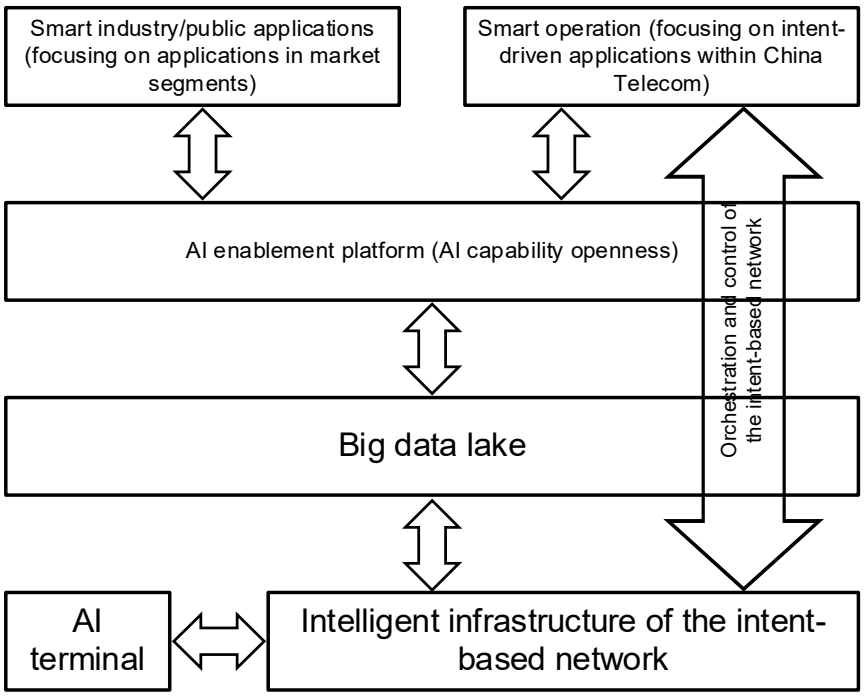



Figure 3: Overall layout of China Telecoms AI development

According to the network practice of China Telecom, the intent-based network is regarded as an advanced network on demand. That is, the on-demand, self-service, and elastic network services evolve to an automated, closed-loop, and intent-driven network organisation — a basic features of the evolution from CTNet2025 1.0 to 2.0.

The construction of China Telecoms intent -based network follows the overall layout of China Telecoms AI development. The network consists of four parts: smart brain, orchestration and control layer, and intelligent infrastructure of the intent-based network, and AI terminals. The smart brain of the intent-based network includes the big data lake and AI enablement platform in the overall layout as well as various AI capabilities developed based on them. The orchestration and control layer of the intent-based network is the key channel hub for implementing smart operation based on the network infrastructure. The following figure shows the overall architecture.

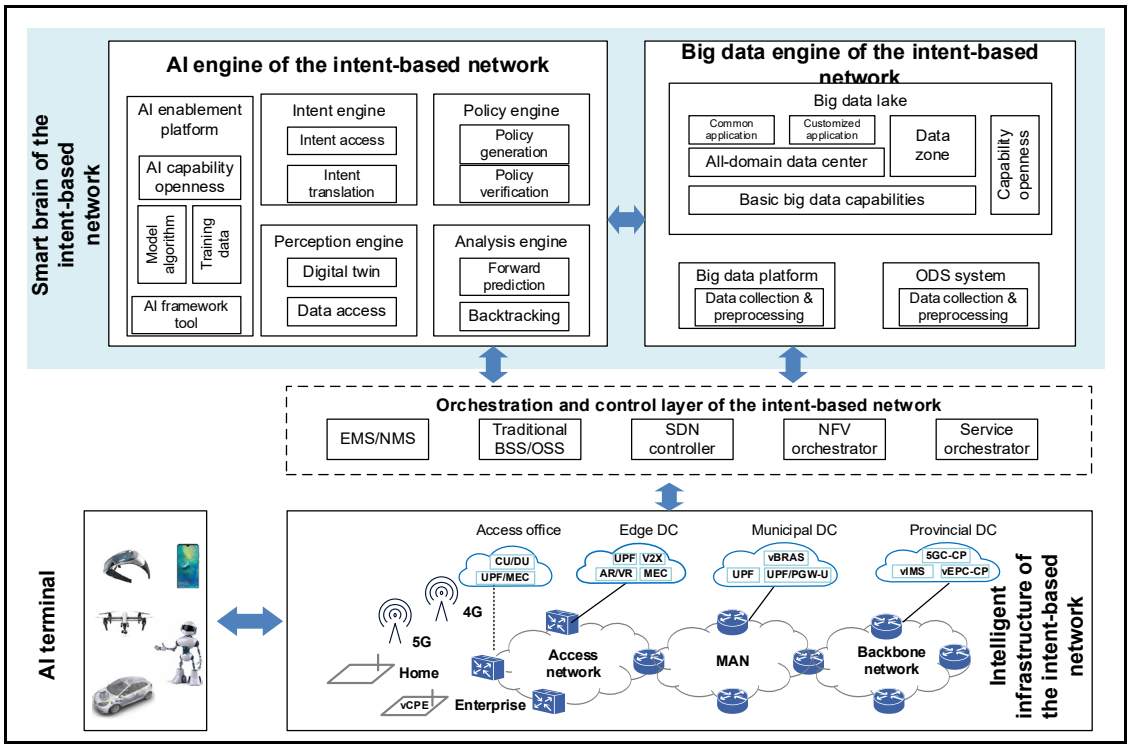



Figure 4: Target architecture of China Telecoms intent -based network

China Telecom positions itself as an AI network builder, an AI industry driver, an AI technology adopter, and an AI service provider. The comprehensive introduction and development of AI technologies enable China Telecom to:

- Accelerate the upgrade of intelligent networks.
- Form an intelligent industry ecosystem.
- Improve the level of intelligent operations.
- Build a comprehensive intelligent intent-based network by using the practical requirements and specific scenarios of telecom networks and services as the breakthrough point.
- Provide get-as-you-wish services oriented to customer requirements.

**Intelligent Network Development Strategies and Plans of China Unicom**

In 2018, China Unicom released the upgraded version of CUBE-Net2.0: CUBE-Net2.0+. China Unicom introduced AI-based on software-based and cloud-based networks to build its intelligent, agile, intensive, and open next-generation intelligent network CUBE-NET2.0+.

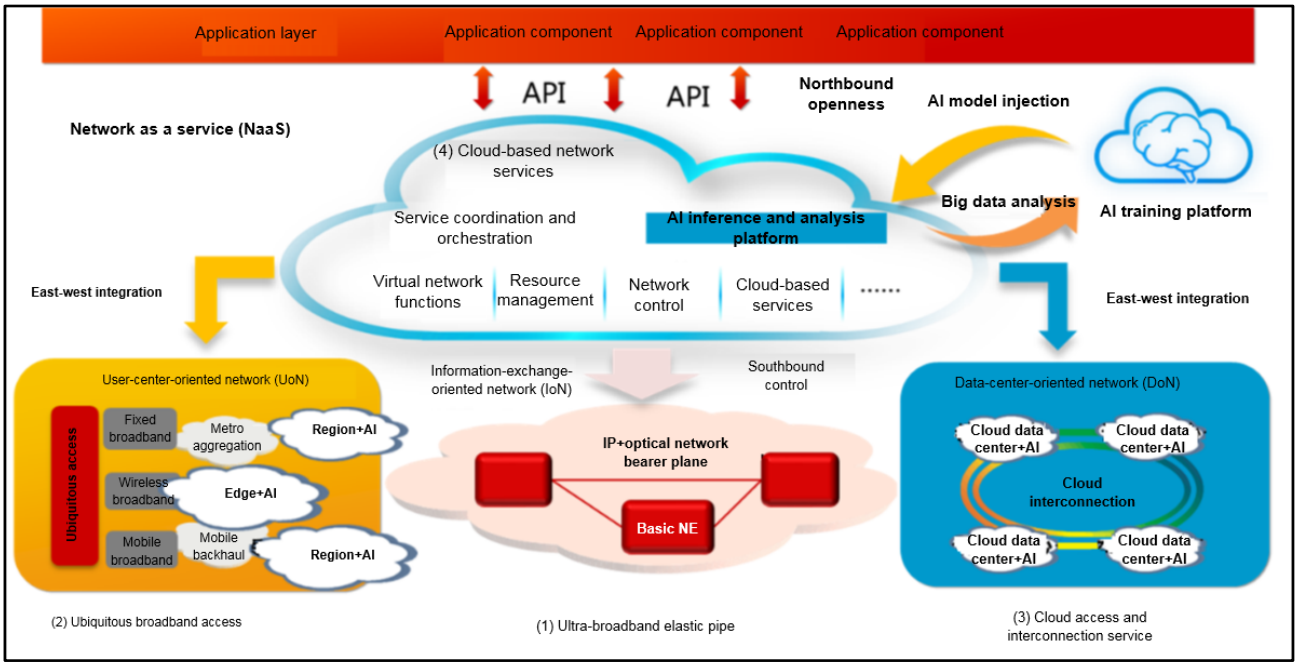



Figure 5: Architecture of China Unicoms CUBE -Net2.0+ network

China Unicoms network AI is implemented from the following aspects:

- Follow the application-driven, platform enablement, and ecosystem win-win development strategy.
- Focus on network AI application innovation such as 5G+AI and industry innovation.
- Build the CubeAI platform for the network AI development engine.
- Build a win-win network AI ecosystem and an open cooperation system.

At MWC Shanghai 2019, China Unicom officially released the CubeAI platform. CubeAI is a platform for network AI technical services, industry cooperation, and communication and sharing. Technical services refer to the provisioning of network AI algorithms, models, services, and applications to support network production, operation, and service innovation. Industry cooperation aims to build a network industry cooperation ecosystem and give full play to advantages of all parties in the industry chain so that they can jointly promote network AI innovation and application. Communication and sharing mean carrying out internal and external technical communication, open-source cooperation, standards formulation, test and verification, application demonstration, and experience sharing.

As an essential component of CubeAI, the AI model sharing and capability openness testbed was also released by China Unicom. China Unicom independently developed it based on the design concept of the Linux Foundation AI open-source project Acumos. The testbed is a full-stack open-source network AI platform supporting model management, model sharing, and model microservice deployment. The source code has been released to the GitHub open source community. In the future, China Unicom will invite partners to participate in platform building and sharing, jointly promote the application demonstrations of typical network AI and actively carry out technical cooperation with related open source communities.

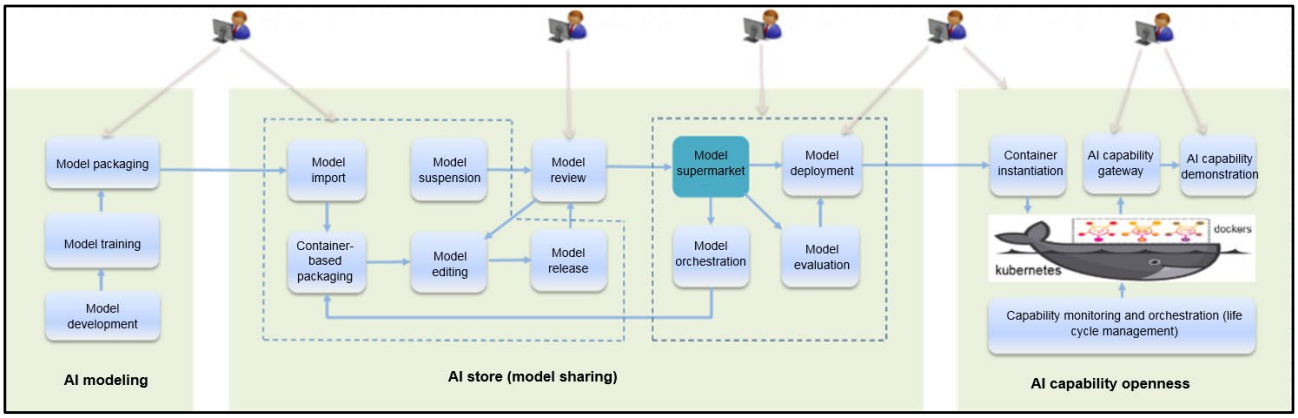



Figure 6: CubeAI testbed project of China Unicom

China Unicom has initiated the CubeAI partner program and won wide support from the industry. Nearly 30 companies in the industry, including Huawei, ZTE, and Baidu, have jointly participated in the initiation. The partners include leading equipment vendors, AI technology providers, and chip vendors.

**The practice of International Operators in Intelligent Autonomous Networks**

AT&T, a North American operator, believes that we can create a better world more quickly than ever by integrating AI with human expertise and evolving it into the AI that enables people. AT&T believes that the core of future 5G technologies includes mobile technology, fixed broadband technology, and edge computing and that thousands of small cells to be deployed in large quantities. Therefore, operators must collect and analyse massive terminal data in real-time. The introduction of AI technologies can ensure the continuous, secure, and efficient operation of 5G networks. AT&T also believes that AI technologies can help optimise network and enhance network security in addition to assisting network deployment. For example, AI technologies can optimize the traffic and network rate, enable each user to obtain the optimal bandwidth anytime anywhere, and ensure data security when users exchange data with each other.

Vodafone, a European operator, believes that AI is at the core of the digital society strategy and provides a new perspective for solving practical problems. AI has helped Vodafone improve its product and service levels as well as business operation efficiency. In just four years, Vodafone has built a large big data team. This team uses AI algorithms to process a large number of anonymised datasets to provide varied services for different customers from a new business perspective. Vodafone has deployed a large number of wireless access devices in European and Asian markets. These devices adopt AI algorithms to determine the locations where customers need larger capacity, thereby improving network coverage. Also, Vodafone applies AI technologies to optimise handovers between cells. These successful applications can help Vodafone save network energy consumption by about 15% in these areas.

**Driving Forces of Intelligent Autonomous Networks**

In a survey conducted by Analysis Mason in 2018, 76 surveyed operators listed the main driving factors for AI-assisted network automation. As shown in the following figure, the main reason for using AI to assist automation is to **reduce OPEX**. Nearly 80% of operators put this reason in the top three driving factors. Other factors are as follows:

- Improve customer QoE.
- Densify the network efficiently.
- Support increased E2E automation.

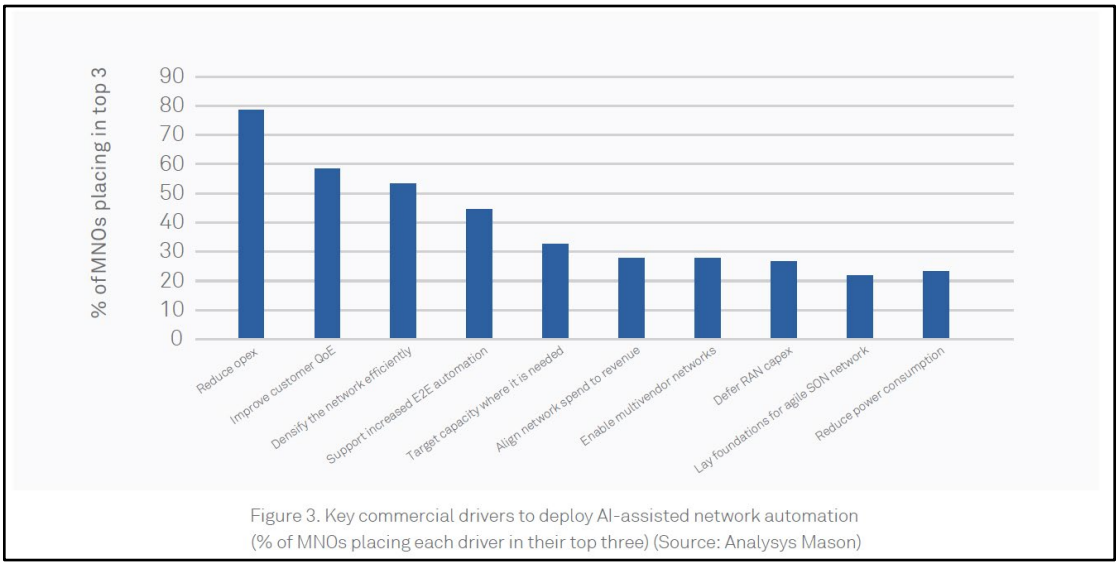



Figure 7: Driving forces for operators to use AI to assist network automation

**Trend Prediction of Intelligent Autonomous Networks**

As mobile operators begin to evaluate their commercial 5G strategies, some operators have introduced automation capabilities in their network processes, mostly in aspects of O&M, planning, and optimisation. According to the survey and prediction of Analysis Mason, although 56% of mobile operators around the world did not have automation in their networks during the survey, nearly 80% of operators expected more than 40% automation by 2025, and one-third of operators expected the automation rate to exceed 80%.

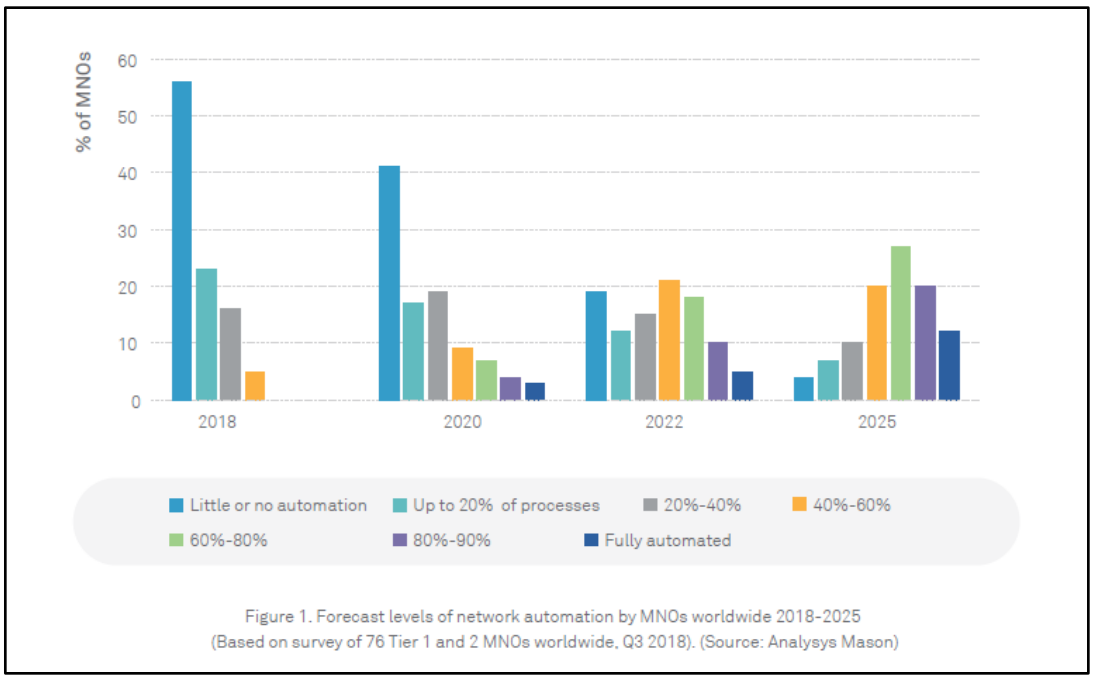



Figure 8: Automation trend prediction for operators networks

## 1.4 Progress in Standards Related to Intelligent Autonomous Networks

A hot topic in the industry, AI in Network have achieved significant growth in the standards and industry organisations of the telecom field. International standards or industry organisations such as 3GPP, ITU-T, ETSI, and CCSA have initiated their researches on AI in Network as follows.

**3GPP**

In August 2018, 3GPP SA WG5 started a study item on Intent driven management service for the mobile network. This project aims to study mobile network communication intent-driven management scenarios that improve O&M efficiency and define intent-driven management service interfaces to implement automatic closed-loop control. The purpose is to implement simplified control of complex networks or services. In August 2018, 3GPP SA WG5 also started a work item on Self-Organising Networks (SON) for 5G networks. This project aims to support the SON of 5G networks through scenario awareness and data analysis provided by the management data analysis function. The SON of 5G networks features self-configuration (for example, base station self-establishment), self-optimisation (for example, coverage and capacity optimisation), and self-healing. 3GPP RAN WG3 started RAN-centric Data Collection and Utilization SI in June 2018. The objective is to study the wireless big data collection and application-oriented to network automation and intelligence, including the process and information interaction required for studying different use cases. At present, the research phase of this project has concluded, and the project has entered into the standardisation phase. 3GPP SA WG2 started a study item Study of Enablers for Network Automation for 5G SI (FS_eNA) in January 2018 and launched the standardisation work in January 2019. This project aims to provide big data analysis by using the network data analytics function (NWDAF) of 5G core networks. This project defines the input and output of network data analysis and implements intelligent network optimisation based on big data analysis, and the optimisation includes:

- UE-level customised mobility management, such as paging enhancement and mobility template, and connection management based on UE service behaviour
- 5G QoS enhancement, such as non-standardized QoS profiling based on user QoE
- Network load control, such as UPF selection and network performance prediction

**ITU-T**

In November 2017, ITU-T SG13 initiated the Focus Group on Machine Learning for Future Networks including 5G (FG-ML5G).

This group aims to:

- Identify the relevant standardisation gaps of machine learning for 5G and future networks.
- Improve the interoperability, reliability, and modularisation capabilities of 5G-oriented machine learning.
- Deliver technical reports and specifications for ML for future networks, including interfaces, network architectures, and protocols, algorithms, and data formats.
- Analyse the impact of the adaption of machine learning for future networks, such as network autonomous control and management.

In July 2019, SG13 approved the *Architectural framework for machine learning in future networks, including IMT-2020* submitted by the focus group in March 2019 as a technical specification. This specification contains the following information:

- Terminology for ML in the context of future networks
- Architecture framework for machine learning in future networks and 5G
- Guidance for applying the architecture framework to specific technologies (such as 3GPP, MEC, edge network, or transport network).

Now, ITU-T FG-ML5G has started the second phase (which lasts until July 2020), the group will further study and expand:

- Use cases and basic requirements of machine learning in 5G and future networks
- Data handling framework to enable Machine Learning in 5G and future Networks
- Methods for evaluating the intelligence level of mobile networks

**ETSI**

ETSI Zero-Touch Network & Service Management (ZSM) was established in January 2018 to automate the network or service O&M workflow (including deployment, configuration, maintenance, and optimisation) and realise the vision of zero-touch E2E workflow. In June 2019, three research subjects were initiated: key problems, basic capabilities, and solutions and the objectives are as follows:

- Study and explore the issues and challenges of network automation.
- Study and analyse technologies related to automatic closed-loop operations.
- Define automation-related policies and intent interfaces to implement closed-loop, E2E domain management and interaction and coordination between domain.
- Define the E2E automation process based on typical cases.

ETSI Experiential Networked Intelligence (ENI), founded in February 2017, aims to define an experience-based perception network management architecture using a perception-adaptation-decision making-execution control model and to improve operators network deployment and operation experience using AI technologies. The core concepts are network perception analysis, data-driven decision-making, and AI-based closed-loop control. The release 1 phase of the related use cases, requirements, and architecture was completed August 2019. The release 2 phase focuses on closed-loop control of real-time networks.

**CCSA**

In July 2017, CCSA TC1 WG1 launched the research project named Application Research of Artificial Intelligence in Telecom Network Evolution. This project studies the application of AI technologies from the following aspects:

- Telecom network maintenance, for example, fault analysis, locating, and prediction
- Network optimisation, for example, intelligent collection, analysis, and forecasting of network performance data
- SDN/NFV network self-management, self-control, self-adaptation, and decision-making control, for example, intelligent network optimisation

At the first meeting of the Technology Management Committee held July 2019, CCSA agreed to change the name of TC1 WG1 to Internet Application and AI Working Group" and explicitly encouraged TCs to carry out AI standardisation within the TC research scope. In January 2019, CCSA TC3 WG1 started the research project named AI-based Network Traffic Prediction and Application Scenarios. This project aims to study various potential application scenarios of network traffic prediction in future networks and analyse

the application of AI algorithms in network traffic prediction. In July 2017, CCSA TC5 WG6 started a research project named Application of AI and Big Data in Wireless Communications Networks.

By August 2019, after WG6s 52nd meeting , the research output includes:

- Wireless channel modelling method based on AI and big data
- Application of AI and big data in the following fields:
  - Wireless signal detection and estimation
  - Resource management
  - Intelligent evolution of the wireless network architecture
  - Wireless network planning, optimisation, and O&M
  - Data service push
- Study on grading method for intelligent capability of mobile networks

# 2 Overview of Intelligent Autonomous Networks

Intelligent autonomous networks need to:

- Be based on cloud infrastructure.
- Build AI and big data engines.
- Consider the features of different network layers, and decouple design, micro-module implementations.
- Focus on high-value scenarios.
- Introduce AI capabilities on-demand and gradually.

## 2.1 Architecture Considerations

To achieve the goal of a fully intelligent autonomous network, and realise AI in Network, without further increasing the network complexity, hierarchical architecture needs to be ensured. A more upper-layer and centralised deployment location indicates a larger data volume, raises higher computing power requirements and is more suitable to perform cross-domain global policy training and reasoning which does not have real-time requirements. For example, cross-domain scheduling and E2E orchestration have high demands on computing capabilities, require massive cross-domain data, and have low requirements on real-time performance. While at the lower-layer and closer to the end, the domain-specific or entity-specific analysis capability may be possible and real-time performance requirements may be met. The following figure shows a three-layer architecture consisting of the cross-domain orchestration layer, single-domain autonomous layer, and network entity (NE) layer. Information needs to be coordinated and exchanged through open interfaces (such as open APIs and SDKs) within different levels of closed-loops, for example, within cross-domain closed-loop or single-domain closed-loop.

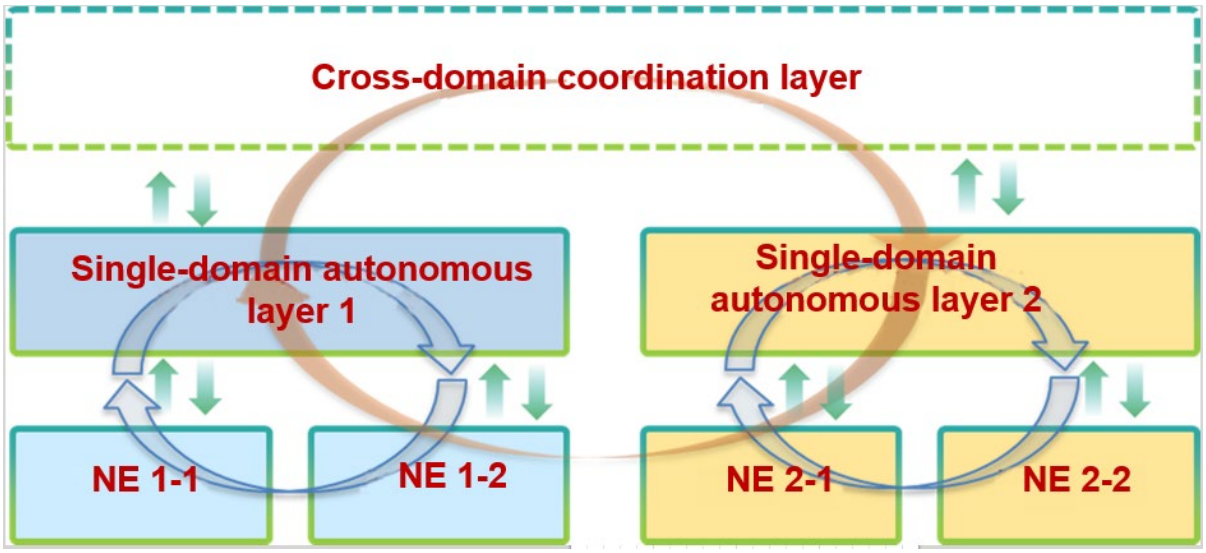



Figure 9: The three-layer architecture of an intelligent autonomous network

- Cross-domain orchestration layer: Operators use expert experience and global data to complete AI model training and enable cross-domain and entire-network closed-loop control. The goal is to convert expert knowledge into models and provide more intelligent services for customers. Cloud-based intelligence can mark and classify data and scenarios based on actual network scenarios to train accurate AI models and offer new intelligent service: AI as a Service (AIaaS).
- Single-domain autonomous layer: A single network domain consists of a group of NEs that can jointly complete the same work. According to different scenarios, a single network domain may be a core network, an access network, or a private enterprise network containing its core network and access network. At the single-domain autonomous layer, an intelligent engine that integrates management and control needs to be built to implement autonomous and closed-loop management of each single network domain. For example, for the radio access network (RAN), transport network, and core network, the capabilities of this layer are required to implement automation in a single domain. Network intelligence is necessary to analyse and infer data in the local domain, identify various network scenarios, predict and prevent possible events, analyse root causes of occurred events, and provide decisions.
- NE layer: Based on the embedded system, the framework and algorithm platform of machine learning and deep learning are built inside the NE devices to provide scenario-based AI model libraries and structured data. Local intelligence provides two key capabilities: data extraction and model inference. Massive data generated by sites is extracted as useful sample data. Real-time AI model inference is performed on the CPU, DSP, or AI chip through the embedded AI framework. In this way, adaptive scenario matching is implemented locally, and real-time parameters and resources are automatically optimised.

To ensure optimal performance and minimise the complexity of integration between layers, the layered framework requires simplified open interfaces between layers. The information exchanged through open interfaces will gradually be simplified from massive data and parameters to intents. Conversely, the simplification of open interfaces depends on the network autonomy capability of each domain and layer. Therefore, autonomy by layer and coordination with openness is the key to success.

## 2.2 Phases of Intelligent Autonomous Network

To implement a fully intelligent autonomous network is a long-term goal, which needs to be implemented step by step as follows:

- Provide an alternative solution for repeated operations.
- Sense and monitor the status of the network environment and network devices, then make decisions based on multiple factors and policies.
- Since the intent of operators and users, and perform self-optimisation and evolution.

The complexity of communication networks determines that intelligent autonomous networks cannot be achieved overnight and maybe gradually deployed in phases. The GSMA supports having an aligned industry view on the phasing, and the below serves as a comparative framework which may be implemented in six possible phases:

- Network with manual operation (L0)
- Network with a computer-assisted operation (L1)
- Preliminary intelligent autonomous network (L2)
- Intermediate intelligent autonomous network (L3)
- Advanced intelligent autonomous network (L4)
- Fully intelligent autonomous network (L5)

| **Phase** | | **Key Feature** | **Evaluation Dimension** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **Execution** | **Perception** | **Analysis** | **Decision-making** | **Intent-driven** | **Scenario** |
| **L0** | Network with manual operation | All manual operations | Manual | Manual | Manual | Manual | Manual | None |
| **L1** | Network with computer-assisted operation | Computer-assisted data collection, manual analysis and decision-making | Mainly system | Mainly manual | Manual | Manual | Manual | Few scenarios |
| **L2** | Preliminary intelligent autonomous network | Automatic analysis and manual decision-making based on static policies in some scenarios | System | Mainly system | Mainly manual | Manual | Manual | Some scenarios |

| Phase | | Key Feature | Evaluation Dimension | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **Execution** | **Perception** | **Analysis** | **Decision-making** | **Intent-driven** | **Scenario** |
| **L3** | Intermediate intelligent autonomous network | Automatic analysis of dynamic policies in specific scenarios and system-assisted manual decision-making in pre-designed scenarios | System | System | Mainly system | Mainly manual | Manual | Most scenarios |
| **L4** | Advanced intelligent autonomous network | The system implements the complete closed-loop operation of dynamic policies, and automatically performs intent perception and implementation in pre-designed scenarios. | System | System | System | Mainly system | Mainly manual | Overwhelmingly most scenarios |
| **L5** | Fully intelligent autonomous network | The system implements the closed-loop operation of all scenarios, and automatically performs intent perception and implementation. | System | System | System | System | System | All scenarios |
| Remarks | | | All levels of decision-making and execution support manual intervention. Manual review conclusions and execution instructions have the highest priority. | | | | | |

Table 1: Phase division of an intelligent autonomous network

**Level 0 network with the manual operation:** The system provides only the auxiliary monitoring function. All dynamic tasks must be manually executed.

**Level 1 network with the computer-assisted operation:** The system executes a subtask based on existing rules to improve the execution efficiency. For example, in the network optimisation field, the

network performance data is automatically sensed in specific scenarios. That is, the system automatically collects the network coverage performance data (such as the network coverage and capacity KPIs) based on the predefined rules.

**Level 2 preliminary intelligent autonomous network:** The system enables closed-loop operation for some units in some external environments, reducing requirements on personnel experience and skills. For example, in the network optimisation field, experience models can be used to assist in the analysis of network problems (such as coverage, access, and handover). Alternatively, in the network maintenance field, automatic fault association and compression are implemented. That is, the system automatically associates fault data in a specific environment according to the preset policies. In the network deployment field, automatic detection and configuration of base station hardware are implemented. That is, the system automatically detects and configures hardware devices according to the preset hardware configuration policies.

**Level 3 intermediate intelligent autonomous network:** The system can sense real-time environment changes, and optimise and adjust itself in some fields to adapt to the external environment, implementing intent-based closed-loop management. For example, in the network optimisation field, closed-loop, automated network coverage optimisation is performed. That is, the system automatically detects network coverage problems in a specific environment, and automatically optimises and adjusts the network based on the issues identified. Alternatively, in the network maintenance field, automatic fault root cause analysis is implemented. That is, the system automatically analyses the fault root causes in a specific environment based on the preset policies.

**Level 4 advanced intelligent autonomous network:** The system can predict the service-driven and customer-experience-driven network or implement proactive closed-loop management in a more complex cross-domain environment. In this way, operators can solve network faults before customers complain, reduce service interruption and customer complaints, and finally improve customer satisfaction. For example, in the network optimisation field, coverage parameters are dynamically adjusted based on scenario perception and prediction. That is, the system dynamically adjusts network coverage parameters based on scenario perception and prediction results to achieve optimal coverage. Alternatively, automatic fault prediction is implemented in the network maintenance field.

**Level 5 fully intelligent autonomous network:** This level is the ultimate goal of telecom network development. The system provides closed-loop self-design, self-implementation, self-optimisation, and self-evolution across multiple services, domains, and the entire life cycle. Intelligent autonomous networks are available to support perception and implementation of operators and users intent.

In the early stage, the technical solution of a lower phase can be used to gain advantages in the cost and agility. In the later stage, the solution can be evolved to a higher phase to obtain extra benefits and support more scenarios.

## 2.3 Overview of Use Cases

The working process of an intelligent autonomous network is directly related to the service value of operators. Therefore, operators need to participate in defining the related working process based on the digitalisation degree, enterprise organisation structure, and personnel quality:

- Balance the total cost of ownership (TCO), including CAPEX and OPEX.
- Evaluate the strategic relevance and potential value.
- Determine the essential intelligent autonomous network scenarios.

Operators, infrastructure vendors, and third-party vendors have started to explore intelligent autonomous networks. Cases in network traffic prediction, automatic base station deployment, automatic fault locating, and on-demand experience optimisation emerge one after another. The next section describes typical cases of AI technologies in terms of network planning and construction, maintenance and monitoring, configuration optimisation, service quality assurance, energy-saving, security protection, and operational service. These cases are introduced from the aspects of scenario description, technical solution overview, application effect, and work plan suggestions.

# 3 Typical Use Cases of Intelligent Autonomous Networks

## 3.1 AI for Network Planning and Construction

### 3.1.1 Intelligent Planning Robot

The all-things-connected concept of "4G changes lives, and 5G changes societies" is gradually accepted by all sectors of society, and will bring revolutionary impact on the society, economy, and peoples life.

AI algorithms are developing rapidly. The traditional static data learning is gradually changed to dynamic-data-based continuous learning, making the burst, unpredictable, and unrepeatable wireless network data more traceable.

Although AI has a promising application prospect in intelligent planning, there are still no benchmark cases and large-scale applications in the industry. The plan, construction, maintenance, and optimisation processes of AI and communications networks are gradually combined. The AI application in wireless network planning faces the following challenges:

- Wireless network scenarios are complex and diverse, data dimensions are diversified, time variability is substantial, and channel changes are random. As a result, network parameters change considerably. Traditional AI algorithms may fail to converge or have poor effects, making it difficult to perform accurate modelling. Therefore, AI algorithms need to be adjusted continuously.
- Problem-solving is complex. Optimal site planning involves multiple dimensions, such as MR weak coverage, user level, other-network competition, and clustering complaint. In many cases, it is difficult to obtain an optimal solution.
- Classification is inaccurate. Wireless networks differ significantly from each other, and it is difficult to find common features accurately describing them. For example, the Internet scenario-based problems of intelligent planning are usually diversified and concurrent. Even network optimisation personnel cannot distinguish between them. When AI algorithms are used, a large amount of blacklist and whitelist marking work may be required.

  Efficient network operation is the key to success. The following factors need to be considered during network planning:

  - How to determine the priorities of network construction and resource allocation?
  - Where are the traffic hotspots? Where are high-value users?
  - How to determine high-value areas on the network?
  - Should indoor base stations, new indoor distributed base stations, or small cells be used for penetration coverage in this scenario?
  - How to ensure the best experience of VIP users?

- How to identify areas with suppressed traffic and unleash network potential?
- How to evaluate the effect/effectiveness of planning and construction?
- What is the real experience of users in buildings?

As shown in the following figure, for the intelligent planning robot, AI can be used in the entire integrated intelligent management process of wireless network planning, construction, maintenance, and optimisation.

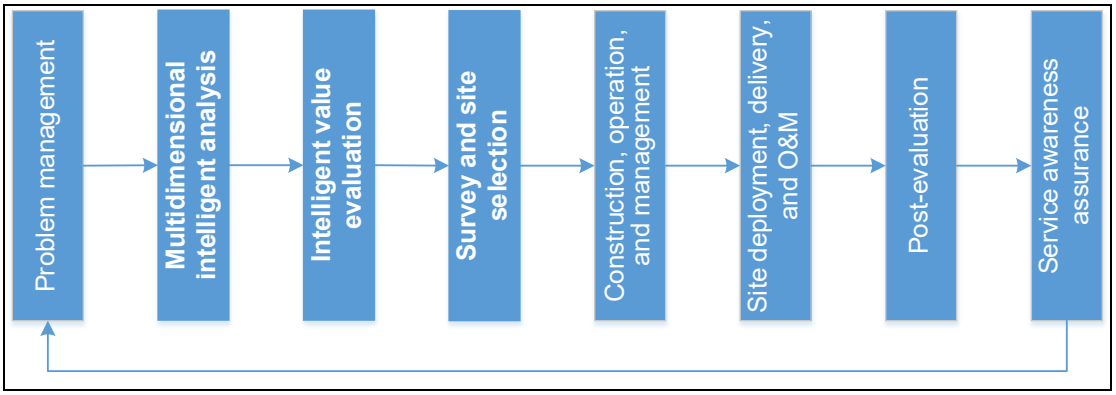



Figure 10: The AI-based entire process of the intelligent planning robot

The basic data to be collected by the intelligent planning robot includes OMS, VMOS, DT/CQT test data, and Internet app data. This solution uses the instruction adaptation, scene recognition, and KQI-KPI model matching algorithms to implement four core capabilities (quick evaluation, sensitive monitoring, automatic optimisation, and iterative planning) for wireless network planning, construction, maintenance, and optimisation.

Currently, AI algorithms are mainly used in the multidimensional intelligent analysis and intelligent value evaluation processes. The algorithms are used for neural-network-based radio frequency pattern matching (RFPM), neural-network-based indoor and outdoor UE differentiation, and intelligent management of Internet-based scenarios.

**Neural-network-based RFPM**

Neural-network-based RFPM aims to extract hidden association characteristics and rules from massive network optimisation data (MR weak coverage+users, competition, and complaint data) by introducing the self-learning and deep learning capabilities of AI. The future network evolution is predicted based on the extraction and summary of common characteristics.

Specifically, data is collected, including the DPI user-plane data, MR data, high-precision building map, CAD building files, KPI data, complaint data, word-of-mouth data, and package data. Based on the collected data, machine learning, neural networks, and algorithms such as density clustering and regression analysis are used to display segmented scenarios, network value, traffic suppression analysis, high-value area positioning, and traffic potential mining in the form of a fingerprint database.

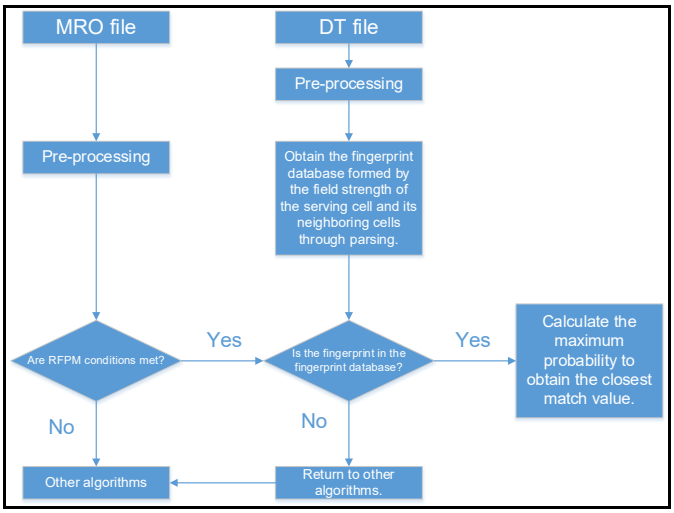



Figure 11:Fingerprint database algorithm selection process based on neural networks

**Neural-network-based Indoor and Outdoor UE Differentiation**

With the development of the MR positioning technology and continuous improvement of the positioning accuracy, the wireless network planning and optimisation is evolved to be more refined and intelligent.

Most UEs and services are located indoors. Traditional DTs and MRs that do not distinguish between indoor and outdoor scenarios cannot reflect indoor network quality. The indoor test workload is heavy, and the test scope and coverage evaluation accuracy cannot be ensured. The indoor and outdoor differentiation technology is a crucial technology for solving deep coverage problems of the LTE network. This technology can quickly and accurately evaluate the deep coverage problem points of outdoor and indoor buildings, providing more refined data for network planning and optimisation.

The principles of the algorithm for differentiating indoor and outdoor UEs based on neural networks are as follows:

- If the serving cell is an indoor distributed cell, indoor distributed signal leakage is not considered, and all UEs are regarded as indoor UEs.
- If the serving cell is an outdoor cell, the following radio signal propagation principles are used: For indoor and outdoor UEs at the same or similar positions, the main difference between them is the penetration loss of buildings. The RSRP of the outdoor UE is 10–20 dB higher than that of the indoor UE, and the mid-value is about 15 dB.
- After deep machine learning, the penetration loss at the same or similar location is compared to obtain a large amount of DT RSRP data of indoor and outdoor UEs in the same cell and TA. Also, the accuracy of indoor and outdoor UE locating is improved with the assistance of Wi-Fi.
- The time and space dimensions are considered to improve locating accuracy.

**Intelligent Management of Internet-based Scenarios**

Intelligent management of Internet-based scenarios is to use Internet-based scenario data (Point of Interest information) and coverage cell data for automatic association analysis to implement scenario data management and scenario-based automated network quality management.

Internet-based scenarios are clustered by coverage capacity, awareness comparison, community distribution, fixed and mobile convergence, and revenue index.

- Level 1: coverage capacity. The coverage capacity is determined and classified by a grid (50 m x 50 m) based on the wireless network coverage strength, 3G/4G network coordination value, and network capacity value.
- Level 2: awareness comparison. Awareness comparison aggregates and classifies scenarios based on user complaints, service awareness DPI data, DT data, and VoLTE awareness data.
- Level 3: community distribution. Community distribution aggregates and classifies scenarios based on scenario definitions (such as business areas and residential areas), population density, and network structure data.
- Level 4: fixed and mobile convergence. Fixed and mobile convergence mainly considers the MR coverage rate and broadband port provisioning rate in the scenario.
- Level 5: revenue index. The revenue index is used to classify the revenue based on the number of content users, several provisioned broadband ports, and Internet service scale in a specific scenario.

The intelligent planning robot has been applied in China Unicom Jiangsu.

To improve the accuracy of traditional network planning, high time and labour costs are required. Take China Unicom Jiangsu as an example. For about 1 million buildings and 200,000 kilometres of roads, it is challenging to perform the ergodicity test, and a large number of personnel, devices, and vehicles are required for the test. Also, the positioning accuracy of the best cell positioning method based on OMC statistics and propagation is only 300–500 m.

The disadvantages of traditional planning are as follows:

- Manual network data maintenance: Basic data is stored in tables. The data maintenance period is extended. Data cannot be automatically updated and needs to be updated manually and periodically.
- Lagging-behind network evaluation method: Traditional DT/CQT evaluation requires 1920 person-days and handles tens of millions of data records each time, resulting in high costs and low efficiency. Also, the evaluation is greatly affected by human and environment factors. The evaluation period is long, and the assessment of integrity is inadequate.
- Lack of MR data collection methods: There is no unified MR data management platform, resulting in high dependence on vendors.
- Inconsistent evaluation counters: Different vendors have different data standards. Manual statistics collected at each level cannot be displayed as unified counters on the platform.

After the intelligent planning robot is introduced, the evaluation criteria and construction criteria are unified, and the whole process is adequately controlled (closed-loop automatic management and control from requirement initiation to post-evaluation).

**Efficiency Improvement**

Take three municipal pilots of China Unicom Jiangsu as an example. The planning accuracy is improved from 500 m x 500 m to 50 m x 50 m. The site planning efficiency is doubled, and the accuracy is higher than 80%. Besides, 40% of problems that cannot be detected by manual means are detected.

| | Number of Weak Coverage Clusters Automatically Output by the Intelligent Planning Robot | Number of Weak Coverage Clusters Manually Checked and Confirmed | Accuracy | Number of Known Weak Coverage Clusters (Traditional Method) | Number of Weak Coverage Clusters That Can Only Be Detected by the Intelligent Planning Robot | Percentage of Newly Identified Problems |
|---|---|---|---|---|---|---|
| City A | 120 | 96 | 80% | 64 | 32 | 33% |
| City B | 79 | 65 | 82% | 44 | 21 | 32% |
| City C | 86 | 71 | 83% | 28 | 38 | 57% |

Table 2: Comparison between the intelligent planning robot and traditional method

For example, in the production of a month, the intelligent planning robot solved 11121 problems, shortened the analysis duration by 2224 hours, and improved the efficiency by 74% compared with the original optimisation working mode.

**Cost-Effectiveness**

Take China Unicom Jiangsu as an example. If the traditional method is used, one administrator needs to be assigned for each vendors 2G/3G/4G network and all third -party services. About ten administrators are required in total. After the intelligent planning robot is used, the Jiangsu province is divided into five areas. One administrator is assigned for each area, and only five administrators required. The management cost is reduced by about 50%.

Also, the network optimisation investment per carrier is reduced from about RMB350 to RMB250. The investment is saved by approximately 28.57%.

**Standardisation**

The standard of the MR data format plays a vital role in intelligent analysis and evaluation. However, the data standard varies depending on vendors and much depends on vendors. In the future, the standardisation of the wireless MR data format needs to be further enhanced.

**Technology Development**

In practice, the positioning technology, especially the positioning technology for differentiating upper and lower layers, is the key to accurate network planning in the future. In the future, the user location information in the application layer of the third-party app can be parsed and associated with the MR on the RAN side to form OTT fingerprints. Based on the Wi-Fi positioning data, indoor MRs can be used to differentiate upper and lower layers further.

**New Challenges and Requirements of 5G**

The space loss and penetration loss of 5G significantly increase due to 5G high-frequency bands. The increase of site density brings new challenges to site construction and requires more refined site location selection. In the future, 5G network planning, site location selection, and antenna location selection need

to be performed based on information such as 3D scenario reconstruction, object identification in coverage scenarios, and planning knowledge graphs.

### 3.1.2 Intelligent Traffic Forecast in the Bearer Network

With the introduction of unlimited quota package, the existing network will inevitably be affected. To cope with this situation, carriers want to predict the peak traffic in advance to guide the adjustment and capacity expansion of the bearer network.

The service guarantee of festivals has always been an essential O&M task for operators. According to the prediction algorithm, the prediction traffic of weekends, weekdays and holidays can be provided to guide the O&M in the existing network.

Currently, bearer networks are used to serve wireless base stations. Carriers hope that predictions of busy and idle hours of various services in each area can be provided as a basis for proper network architecture and reasonable allocation of network resources.

The bearer network traffic prediction solution learns the inherent laws of traffic in terms of location and time through the machine learning algorithm and forms the future growth trend and daily traffic baseline, which can be used in bandwidth early warning, capacity expansion prediction and differential service guarantee.

The solution consists of two parts: algorithm exploration and traffic prediction application. Through algorithm exploration, an algorithm candidates list is formed, and when these algorithms are used to train and predict the current network traffic data, including growth prediction and busy/idle prediction. Then an evaluation result is formed for each algorithm, which facilitates the comparison and optimisation of algorithms. The detailed steps are as follows:

1. Collects traffic in the existing network.
2. Processes the historical traffic data of the existing network and trains models.
3. Use models for traffic prediction. For each model, evaluate a result.
4. Select an appropriate algorithm model by the target.
5. New traffic data is collected, and the forecast is made.
6. Add factors such as holidays to get the final traffic prediction result.

Involved algorithms: In the traffic growth prediction part, the linear regression/sarima/fbprophet algorithm can be used for traffic growth or holiday traffic forecast. Traffic prediction for busy/idle hours can be based on algorithms such as K-Means/DBSCAN for clustering first, then the algorithm such as SVM/Bayesian network is used for 24-hour busy/idle prediction.

Interfaces involved: interfaces between NMs and NEs (e.g., SNMP, Netconf, Telemetry)

Data sets included: Historical traffic data of resource (ports, links, ring networks, and services, etc) for which traffic needs to be predicted.

According to the test, the prediction accuracy is higher than 90%. Compared with rough manual prediction, the prediction can also focus on specific resources and specific time and date.

The result of the traffic bandwidth usage prediction in the convergence ring of the PTN transmission network is shown as follows:

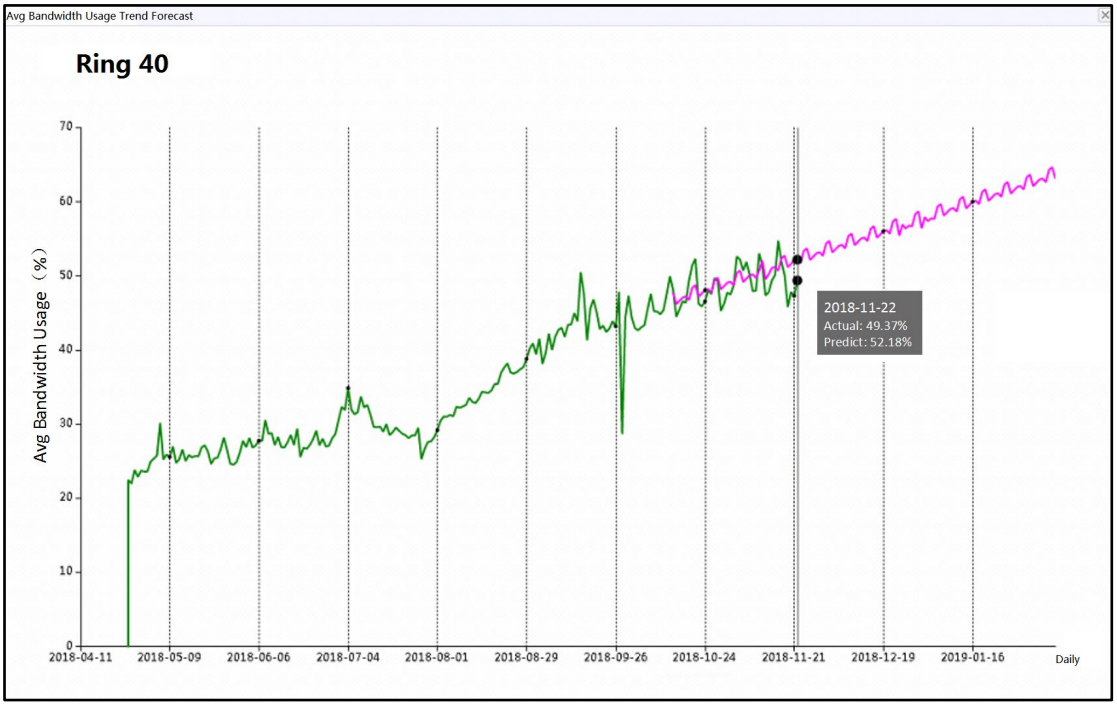



Figure 12: Prediction Result of Bandwidth Usage of Convergence Ring in PTN Transmission Network

Engineering application suggestion: The prediction has high requirements for data, and the quality and quantity of data must be ensured.

Standardisation suggestion: It is recommended to standardise the evaluation model of prediction and application effects.

### 3.1.3 Site Deployment Automation

The base station deployment scenario refers to the entire workflow of deploying the base station after the site survey. The workflow includes network planning and design, site design, configuration data preparation, site installation, onsite commissioning, and onsite acceptance. The following figure shows the complete site deployment process.

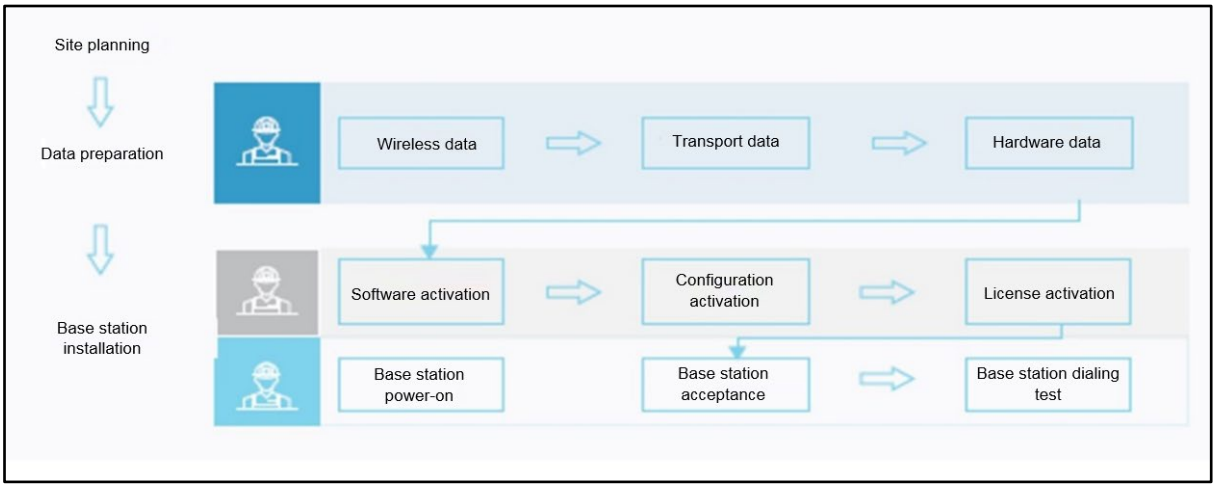



Figure 13: Base station deployment workflow

Traditional base station deployment faces challenges in the following aspects:

- A large number of parameters (usually thousands of parameters) need to be configured, including necessary transmission, device, and radio parameters.
- In the site design and planning phase, all the base station design parameters and changes need to be understood and mastered to ensure correct configuration.
- Site installation may differ from site planning, and manual dialling tests result in long site access time and various site access problems.

Currently, most site deployment solutions are between using tools for auxiliary management and partially autonomous networks. Some leading platforms support automatic conditional deployment. It is foreseeable that the E2E automatic site deployment process will be realised shortly.

The development and introduction of AI technologies will bring revolutionary changes to the implementation of E2E automatic deployment. For example, if big data analysis and deep learning algorithms are introduced when new base stations are deployed on inventory networks, parameter planning can be simplified, deployment policy development can be significantly reduced, deployment accuracy can be greatly improved, and inventory networks can be intelligently followed. On inventory networks, parameters are classified by scenario, and many settings are fixed. A large amount of data exists on inventory networks of operators. Based on the characteristics data of the live network (wireless, transmission, and hardware), deep learning algorithms can be used for online learning to generate deployment policies and templates for different scenarios (such as heat absorption and coverage hole filling).

Therefore, new base stations in the same scenario do not need to be planned one by one. Instead, configuration matching can be performed based on the parameters of inventory base stations to automatically generate parameter configuration planning for new base stations. In this way, the correct simplified input and simplified parameter planning can be realised.

Even for the same project of the same operator, the parameters are not entirely the same because of different characteristics of site coverage areas. Too many differences increase the O&M complexity. Also, network is optimised continuously, and wireless network configuration parameters keep changing dynamically. Therefore, site deployment cannot be performed based on fixed information for a long time. Otherwise, subsequent optimisation work will be more substantial. Based on the above reasons, some initial configuration parameters of the base stations newly deployed on the live network under normal O&M can be automatically generated based on the rules of the live network.

Before deployment, the deployment characteristics are generated based on the planning data and the geographical location of the base station. Then, the system automatically matches the optimal parameter settings and deployment policies on the live network based on the characteristics of the base station. After deployment policies are applied, essential information about neighbouring base stations can be detected for real-time learning. In this way, existing policies can be further optimised to generate supplementary information such as neighbouring cells and power. Scenario-based deployment policies are automatically analysed and obtained on the live network, greatly reducing policy development for similar scenarios. After the deployment, online learning capability can optimise some planned parameters in real-time. This reduces planning deviation caused by problems in obtaining information such as engineering parameters and greatly improves deployment accuracy.

The following figure shows the overall solution process.

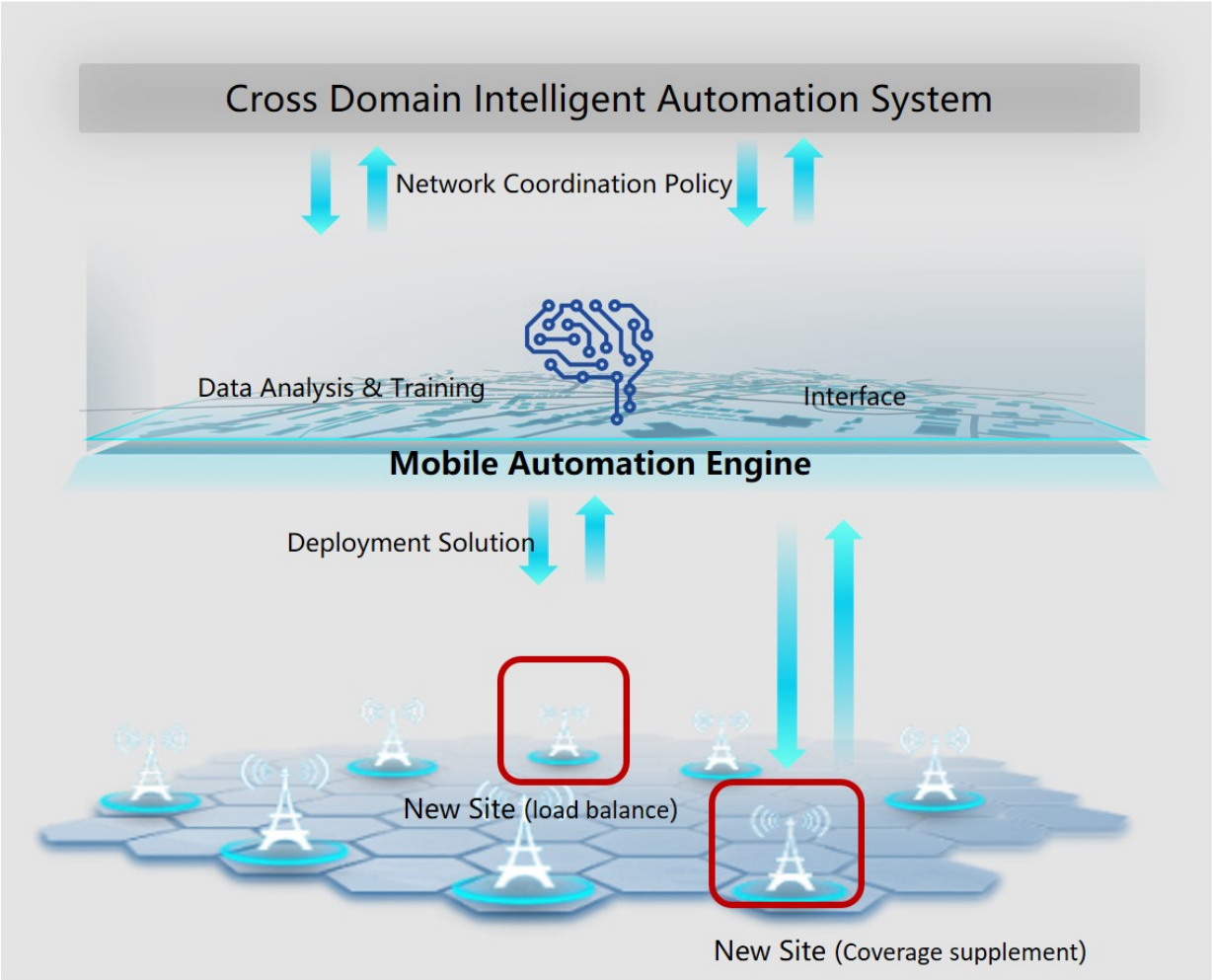



Figure 14: Process of AI-based Site Deployment Automation for Wireless Networks

The process is as follows:

1. Generate a configuration experience library based on expert experience, use the association algorithm, and set the confidence threshold.

2. Based on the inventory wireless network data and associate item sets, use the data mining and analysis technologies to generate possible configuration recommendations for specific scenarios. Obtain frequently associated parameters.
3. Calculate the confidence of configuration prediction based on the live network data and the confidence threshold entered by the user, and verify the obtained configuration recommendations.
4. If the confidence exceeds the confidence threshold, the system automatically generates the optimal configuration rules based on the association items.
5. If a recommendation rule is manually verified to be valid, add the rule to the unified initial configuration rule library.
6. When deploying a new site, invoke the existing configuration rules to generate configuration data of the new site without external input data.

Based on typical frequently associated parameters, the relevant primary vital parameters of RF modules are used at 23 5G sites. A total of 21 valid recommendation rules are generated after verification. The overall validity exceeds 90%, and no error recommendation is made. Logical rules can reduce the number of initial planning parameters by about 5%.

Automatic site deployment plays a vital role in quick site deployment in the 5G era because it greatly improves the deployment efficiency and simplifies parameter configuration. Although AI for automatic deployment is still being tested in the lab, site automation that streamlines operators workflows has been applied to 5G sites of multiple operators around the world. For example, after an operator in Korea adopted automatic site deployment, the time for deploying a site is reduced from 2 hours to about half an hour. For an operator in China, the efficiency of using 5G new sites in Beijing is improved by more than two times, and the effectiveness of deploying 3D MIMO sites is enhanced by more than three times.

In the next step, this solution needs to be better interconnected with operators workflows and systems, and the standardisation of northbound intent-based interfaces needs to be promoted so that E2E automatic deployment can be better embedded into operators workflows.

## 3.1.4 Broadband Installation Quality Monitoring

In China, home broadband users mainly use fibre to the home (FTTH) and fibre to the building (FTTB) to access networks. To guarantee broadband installation quality, the following problems need to be resolved. First, fibre access construction has high requirements for installation and maintenance personnel. Currently, photos are taken during installation, and spot checks are carried out after installation to ensure that installation and maintenance personnel follow the technical process during installation. However, this method requires considerable workforce, and only part of installation points can be checked, leaving potential risks to construction quality. Second, passive optical networks (PONs), the dumb resources, are difficult to manage due to their passive features. As installation personnel focus on construction rather than maintenance during the fast development of network access using optical fibres, the actual usage of optical splitter ports is often inconsistent with that recorded in the resource system. Installation and maintenance personnel often randomly occupy ports for quick construction, affecting port usage. As a result, construction fails to be implemented according to work orders. Also, to correct port usage and effectively use ports, operators have to periodically invest heavily to check resource usage manually. However, manual check is inefficient and cannot ensure quality, as there is no tool for quick resource recognition and automatic comparison.

Currently, operators ask installation and maintenance personnel to take and upload photos of key steps during installation. Then, operators perform manual spot checks for photo-based quality inspection. AI image recognition technology can be applied to the following aspects to improve standards compliance, manage dumb resources, reduce labour costs, and improve broadband installation quality:

- Facial recognition for installation attendance recording
- Installation technique quality inspection
- Optical splitter port identification for construction based on work orders
- Quick resource identification

Before the photo-based quality inspection, invalid, fake, and duplicate photos uploaded by installation and maintenance, personnel can be identified, recorded, and filtered through preliminary preprocessing.


Figure 15: Image preprocessing

Installation attendance recording adopts facial recognition and object detection technologies to check whether installation personnel are appropriately dressed (for example, wearing safety helmets and union suits). Attendance recording by facial recognition is also supported. Construction positions can also be located to check whether installation personnel are on site.


Figure 16: Installation attendance recording

Basic installation technique quality inspection based on deep neural network AI algorithm detects target objects in images. Diversified techniques can be checked:

- Whether cables are neatly bound.
- Whether the optical splitter chassis is contained in the photo.
- Whether a QR code is labelled on the optical splitter.
- Whether a service card is attached to the optical modem.


Figure 17: Quality inspection of basic installation techniques

In-depth installation technique quality inspection detects complicated check items:

- Whether dustproof caps are lost.
- Whether the optical splitter chassis is intact.
- Whether the QR code is correct.
- Whether drop cables are secured.
- Whether optical fibres of the optical modem are correctly wound.


Figure 18: In-depth installation technique quality inspection

During manual spot checks, the most significant step is to check whether installation personnel allocate ports (dumb resources) according to work orders. As the EMS can not manage PON devices, such resource allocation can be assured only through spot checks and standards compliance of installation personnel. The AI image recognition and optical character recognition (OCR) technologies need to be applied to port number recognition and comparison with work orders to ensure construction based on work orders.

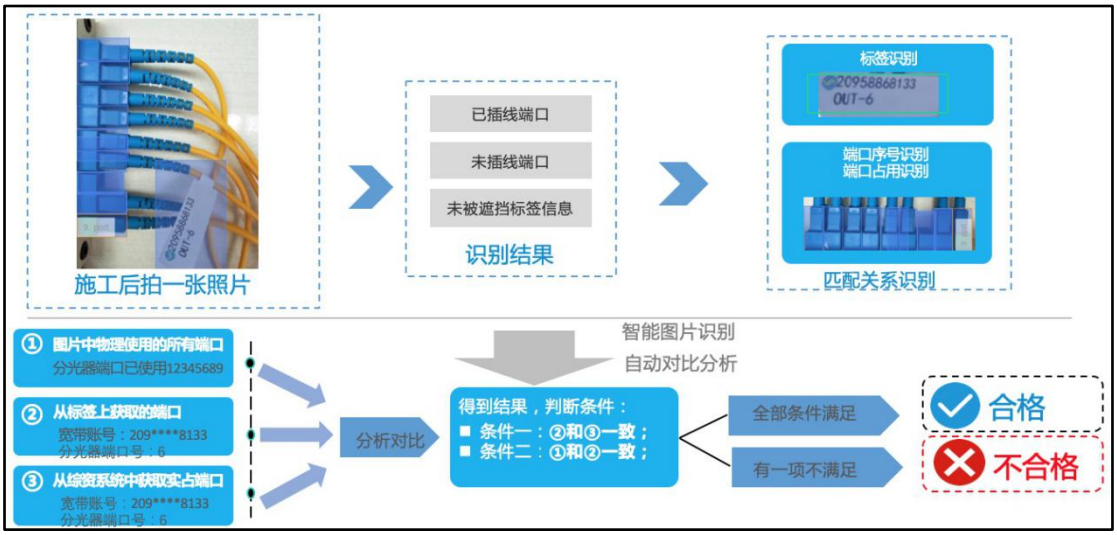



Figure 19: Guarantee of work order–based construction

Whether resources can be identified and checked quickly is another key during the fibre distribution terminal (FDT) and optical distribution frame (ODF) construction. Installation and maintenance personnel need to confirm plate information about optical cable sections and ODF port usage, compare the identification result with records in the resource management system, and update the resource usage statuses. Operators have to set up special teams to work on this each year, causing high costs. If photography and image recognition technologies can be used during construction to manage resource information continuously, much operator investment can be saved. Such resource information management is more precise, gradually improving construction quality.

OCR technology is used to identify plate information of optical cable sections, including optical cable section name, specifications, length, completion time, and construction unit. ODF port usage is determined by recognising the usage of port *X* on disk *X* (**1** indicates this port is used, and **0** suggests this port is not used).

The application algorithm requirements mainly include facial recognition and match, device type recognition, sticker/tag recognition, and port usage recognition. Generally, the technologies required for algorithm implementation can be divided into the following aspects: region of interest extraction (ROI), OCR, and target detection.

- ROI: Detects and locates the sticker or tag area in images using YOLOv3+CTPN algorithms.
- OCR: Recognizes and converts the text area in images into text characters using CRNN algorithm.
- Target detection: Detects targets to be located (herein referred to like used and unused ports, cables, and devices) in images using YOLOv3-spp algorithm.

The data obtained using these algorithms need to be compared with the customer, work order, and resource information gathered over the BOSS system interface. The match result is stored in a relational database for manual statistics collection or further mining.

Image-based quality inspection has been applied to broadband installation quality monitoring in some regions. AI is used to recognise photos taken by installation and maintenance personnel after construction, classify devices and cables, review port usage, and analyse information on tags or stickers. This helps

check whether the construction meets the technical requirements and whether specified downstream ports on the optical splitter are used, and quickly recognise ODF port usage through maintenance inspections.

The following takes an operator in a medium-sized province in China as an example. This solution greatly improves installation and maintenance quality, achieving fruitful results after half a year:

- Saving 30 persons/month: During usage of AI image-based quality inspection for half a year, the construction quality of 150, 000 work orders are checked each month with over 1.3 million images being checked. This saves 30 persons/month compared with manual checks.
- 10% -> 60%: AI enables all construction photos to be checked. Field workers have to implement construction and take pictures in a standard manner. The construction standards compliance rate increases to 60% from 10%.
- 10% -> 4%: During usage of AI image-based quality inspection for half a year, construction quality has been significantly improved, reducing the second visit rate from 10% to 4%.

AI has been initially applied to broadband installation and maintenance, greatly enhancing installation and maintenance quality. This improves quality inspection coverage and contributes to resource management optimisation. The next step will focus on the following aspects:

- Algorithm improvement and introduction of auxiliary tools: When the usage of complex devices (for example, the ODF) is high, cables are crossed in a disordered manner. In this case, even human eyes cannot identify or check cables. Object segmentation algorithms can be improved for such scenarios, for example, using Mask R-CNN or MS R-CNNN algorithm. Auxiliary tools can also be used to reduce algorithm difficulty. For example, a reflector, photo frame, or auxiliary sticker can be used to facilitate image recognition.
- Dataset standardisation: The application datasets come from photos taken during construction. Currently, operators and even branches of the same operator use different installation specifications and stickers. Therefore, datasets obtained from these photos are not standardised, which is not conducive to promoting image-based installation and maintenance quality inspection. The next step is to standardise construction materials further, summarise conventional construction techniques, and improve unified photographing specifications.

## 3.2 AI for Network Maintenance and Monitoring

### 3.2.1 Intelligent Operation Analysis Platform for Wireless Networks

The rapid development of the Internet has offered real benefits to every corner of life. When telecom operators provide network services for individuals, societies and countries, massive data is continuously generated. Some data generated by a server or other network devices reflect server or network features that are compliant with specific rules. Exploring such information is key to monitoring network security. Currently, wireless networks face the following challenges:

- Network data quality differs among operators and cells. Network performance data of some cells may be missing or incorrect, which is difficult to be manually checked. This poses threats to future network optimisation.
- Changes in the UE quantity and surrounding network environment of each cell affect network data in real-time. Personnel cannot check these changes on-site or take measures accordingly. Therefore, early warnings of future network traffic changes cannot be provided.

- Network coverage and signal quality of indoor and outdoor cells differ a lot.

As the radio environment, network structure, user behaviour, and user distribution continuously change, the network needs to be continually optimized and monitored. In this way, network faults can be detected, and factors that affect network quality can be identified. Optimal network status can be achieved by adjusting algorithm parameters and adopting technical methods. Also, the network growth trend can be better understood to lay a foundation for future capacity expansion, thereby improving network service quality.

Models are built based on related studies in terms of four directions, namely historical wireless network performance data distribution and comparison, exception diagnosis, trend prediction, and capacity expansion prediction. Characteristics of 5H1L cells in these aspects are observed. These models focus on operation strategies for intelligent autonomous networks, truly improving intelligent wireless network operations for operators.

AI can be applied to diversified network operation analysis to quickly and precisely locate exceptions, predict network trends, and provide rectification suggestions. This makes network management and operations more efficient, improves intelligent network management, and contributes to centralised, intelligent, and automatic wireless network optimisation.

Key technologies of big data include data storage and data mining, which rely on distributed databases and big data platforms, respectively. Therefore, distributed data mining is the key to wireless network optimisation. The automatic data processing function is deployed on multiple nodes to update data in real-time, ensuring data stability and continuity.

An intelligent wireless network operation analysis solution is provided for operators based on analysis of their requirements and requirements of the wireless network system. This solution targets the following aspects:

- Data problems at each layer during centralised collection and aggregation of NMS data over the entire network are analysed based on existing data and functions of the NMS for 4G wireless networks and big data analysis approach. The data sources include Indicator distribution data: Historical performance indicator data of the same indicator on the same network management equipment before and after a certain time point is compared. Historical performance indicator data of the same indicator during the same period on different network management equipment is analysed. Exception diagnosis data: NMS data of nationwide cells during the last week or month is analysed based on KPIs in the wireless performance data. Trend prediction data: Similarly, NMS data of national cells for the next week or month is predicted based on KPIs in the wireless performance data. Capacity expansion prediction data: Multi-dimension NMS data, such as wireless performance, user experience, and cell configuration data, is analysed to predict capacity expansion.
- A diagnosis model is built for data integrity, rationality, and stability. The most practical algorithm for exception diagnosis of wireless network data, optimised LOF, is selected by sorting up and comparing multiple advanced algorithms (quartile deviation method, cluster analysis, and LOT). This algorithm enables the model to reflect the data fluctuation diagnosis result within the next slot, enhancing basic data quality.
- The optimal prediction model, LSTM+DNN/GRU, is selected based on diagnosis results of historical wireless performance data and by comparing practical prediction models (ARIMA, wavelet analysis, and LTSM) for time sequence analysis. This model helps predict changes in wireless KPI data to learn the KPI trends of cells to some extent.

- Key tags are established for cells based on external information, such as NMS capacity expansion and complaints. Then, models are built for cell load warning and possible risk detection through data training. The application modes of machine learning in problem analysis are discussed, and an intelligent warning mechanism is created for the integrated NMS.

The intelligent operation analysis platform for wireless networks based on AI and big data was piloted in China Telecom Fujian in 2018. The application effects of its functions are as follows:

**Indicator Distribution Comparison**

This solution can display indicator distribution (on different network management equipment in the same period and on the same network management equipment in different periods), visualise abnormal data changes of vendors or China Telecom Fujian, and compare distribution trends in different periods or on various network management equipment. This ensures the proper allocation of network resources and improves network usage.

**Exception Diagnosis**

Exceptions of cells and performance indicators in cities and provinces, changes in abnormal signs, and distribution of abnormal cells are displayed based on deep learning algorithms and wireless network performance data. These cells, cities and provinces where exceptions potentially occur are marked on the grid-based map to develop related rectification solutions.

For poor-quality cells, this solution can analyse key indicators that cause poor service quality, find out root causes, and take measures accordingly. This prevents the expansion of network incidents and ensures wireless network operation quality.

**Trend Prediction**

Trends of specific performance indicators or performance indicators of specific cells for the next week or month can be predicted based on historical KPI data trends and time sequence deep learning algorithms. The prediction accuracy is higher than 80%. In this way, future traffic distribution can be accurately predicted, and KPIs can be monitored to diagnose possible incidents and take preventive measures in time, lowering the network incident rate and enhancing network connection quality.

**Capacity Expansion Prediction**

This solution predicts cell capacity expansion for the next half-year based on wireless network performance data, perception data, and configuration data together with supervised deep learning algorithms with a prediction accuracy rate of 99%. This solution can update the number of cells to be expanded, distribution of these cells, and capacity expansion of 5H1L cells in cities and provinces in real-time each day. In this way, management personnel can adjust the cell capacity expansion solutions and perform capacity expansion promptly to improve network service quality.

Currently, data services account for a more significant proportion of services provided by mobile operators, and mobile networks are transformed into hybrid multi-layer networks. This poses a great challenge to operators, for such transformation requires faster and more flexible management and control mechanisms to improve the operation efficiency and meet the ever-changing market requirements innovatively. The key to this challenge is efficiently sorting up various data resources on mobile networks and using big data technology to implement in-depth association analysis.

As relying more on mobile Internet, people are more sensitive to network service quality, which is the focus of operators. Network quality has become a key factor for the development of operators,

determining their operations. Big data and AI technologies have been widely used in mobile Internet, and are initially applied by operators in more fields. These technologies enable an analysis of multiple factors (such as network traffic, terminals, and users), improving the network optimisation efficiency and saving investments. If further analysis can be performed, these technologies can provide data support for refined marketing and user experience improvement.

The operation analysis platform for wireless networks can properly support decision makings on the live network during the pilot phase. However, as live-network adjustment requires high preciseness and automation capabilities of existing equipment are insufficient, policies cannot be automatically delivered and executed based on the analysis results provided by the platform. In the future, algorithms for the operation analysis platform can be continuously optimised to make predictions and detections more accurate. Equipment can also be reconstructed or upgraded to prepare for automatic operations. In this way, resource usage can be further improved, and the incident rate and costs can be lowered. During the pilot phase, if vendors use different calculation methods or data formats for the same KPI, it is recommended that these calculation methods and data compositions for related parameters be standardised.

5G, which requires better O&M due to its network complexity, is developing rapidly worldwide. In China, 5G commercial licenses have been issued to operators. Therefore, demands for automatic network operations and fewer manual operations are more than ever. Challenges in massive data collection, storage, standard applications, formulating AI core capabilities targeting networks, and cultivating talents understanding both networks and AI technology, need to be resolved for 5G/future networks.

## 3.2.2 IP RAN Alarm Compression

IP RANs mainly carry 3G and 4G mobile services and VIP leased line services by dynamically using the IP/MPLS protocol. Compared with traditional networks, IP RANs use more complex protocols and logical connections. Compared with the traditional NMS, the NMS of IP RANs receives a large number of device alarms, many of which are caused by root alarms.

Currently, the massive alarm data is processed based on experts experience. That is, experts knowledge is summarised into rules, and non-key alarm information is filtered out based on these rules. To avoid filtering out critical alarms, the filtering control policy is loose, which leads to a limited filtering capability.

During maintenance of IP RANs, operators hope to apply the AI technology to alarm compression, enhancing alarm handling efficiency. Typical service scenarios are as follows:

**Scenario 1: Transient Alarms**

A transient alarm refers to an alarm whose interval between the generation time and clearance time is less than a specified threshold. This type of alarms has a short lifecycle and offers little value. Besides, the increase of these alarms also distracts O&M personnel from critical alarms that need to be focused on, making alarm recognition more difficult.

**Scenario 2: Frequently-Generated Alarms**

If the number of the same alarms or events generated within a period reaches a specific value, these alarms/events are considered correlated. Rules are set for the event/alarm frequency analysis. If the number of specified alarms/events generated within a period exceeds the preset threshold, these alarms/events are considered correlated. For example, if the temperature of the same board on the same NE is too high or low for *Y* times in *X* minutes, a new alarm is generated, reporting that the board temperature is abnormal.

**Scenario 3: Intra-NE Fault Impact Analysis**

If an alarm is generated for a physical object (such as a board or topology) of a NE, correlative alarms are generated for other physical objects and logical objects of this NE.

For LTE, boards in a base station and boards and logical objects are correlated. A faulty board often leads to abnormal cells. For example, when an alarm indicating unavailable optical modules is generated for the BBU, an alarm indicating RRU link disconnection is produced for the RRUs. Meanwhile, an alarm indicating LTE cell out of service is reported for cells served by these RRUs. The alarm indicating unavailable optical modules is the root one.

**Scenario 4: Fault Impact Analysis of Intra-Network Services at Different Layers**

In this scenario, a fault leads to a large number of alarms. The root cause needs to be quickly located. Service-layer alarms cause client-layer alarms. For example, if a branch optical fibre is broken, a loss of signal (LOS) alarm is generated for its port. As a result, alarms are reported for tunnels, pseudo wires, and services at the upper layer. The LOS alarm for the port is the root one.

**Scenario 5: Cross-Network Alarm Analysis**

Transmission is classified into the optical transmission and microwave transmission. An optical transmission node is connected to multiple microwave nodes. If a link is disconnected, one or more sites using this link is affected. If an optical transmission node is disconnected, all downstream sites using microwave transmission are out of service. If a microwave hop is unavailable, all the downstream sites are out of service.

**Scenario 6: Comprehensive Fault Diagnosis**

Faults can be reflected in many ways, such as alarms, abnormal KPIs, or unavailable services. In most cases, alarms cannot reveal all errors, and therefore faults cannot be located just using alarms.

For example, if LTE services are unavailable after a network upgrade, maintenance personnel can view monitoring data based on experience, perform diagnosis actions, and check configurations to locate faulty points.

In IP RAN scenarios, alarms need to be intelligently identified and analysed.

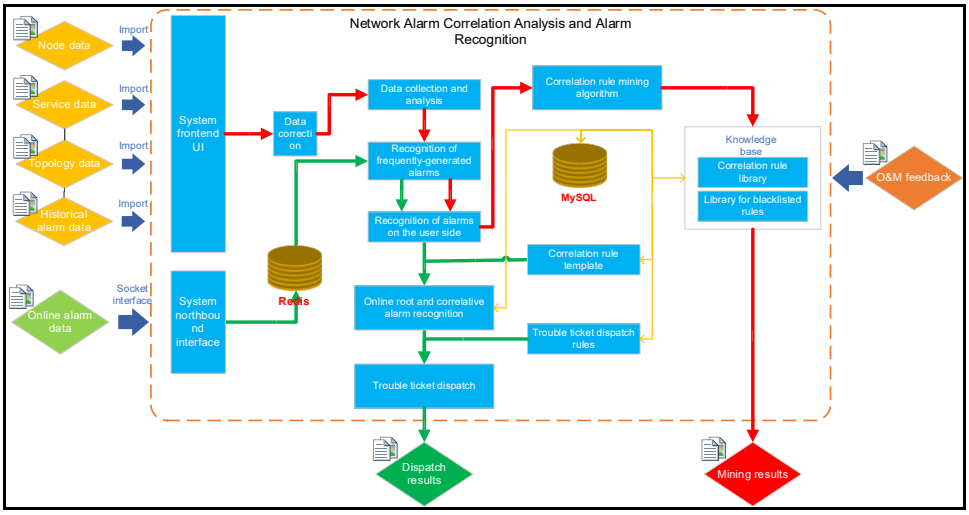



Figure 20: IP RAN alarm compression solution

IP RAN alarm compression consists of four phases:

- Data preprocessing: Data preprocessing includes data import and cleansing, alarm matching on the UE side, and recognition of frequently-generated alarms. The input data includes the historical alarm data, network topology data, and service data captured on the live network. After being cleansed and integrated, the input data is converted into a format that can be processed. Alarm match on the user side refers to removing unimportant or useless alarms. The method of processing frequently-generated alarms is to compress the same alarms generated on the same port in consecutive 10s. Only the first alarm is reported, and other alarms are marked as filterable alarms.
- Correlation rule mining: The core algorithm of correlation rule mining is PrefixSpan time sequence pattern mining algorithm, which is proved to be more suitable after compared with mining algorithms such as Apriori, sequence pattern, and Spatio-temporal pattern. However, the rules obtained using the traditional PrefixSpan algorithm do not have constraints. Experts cannot determine whether the correlation rules (for example, rule A [alarm indicating that optical modules are unavailable -> RRU link disconnection alarm]) is correct. Therefore, the PrefixSpan algorithm is optimized to contain constraints during the mining process. Rule A is optimized to [alarm indicating that optical modules are unavailable -> RRU link disconnection alarm, same NE], making the rules obtained using the algorithm more accurate.
- Correlation rule confirmation and integration into the knowledge base (including the confirmed correlation rule library and blacklist): Multiple experts verify the discovered alarm correlation rules and save the correct ones to the confirmed correlation rule library for future alarm identification. Incorrect and improper rules are automatically imported to the blacklist to prevent similar rules from being discovered next time.
- Root alarm identification: Each alarm is marked with three types of labels, namely the root alarm, correlative alarm, and common alarm. Current alarms are identified and processed based on the following types of constraints: same port, same NE, similar service NE, same service ID correlation, directly-connected peer NE, directly-connected peer port, NEs on the same ring, and corresponding service ID correlation.

IP RAN alarm compression has been piloted in many cities by China Unicom Jiangsu. The test results show that the average compression rate of offline alarms is over 90%, and that of online alarms is over 85%.

| **Alarm Handling Result in Pilot Cities** | **City A (1 month)** | | **City B (1 month)** | | **City C (1 month)** | | **City D (2 months)** | |
|---|---|---|---|---|---|---|---|---|
| Handling method | AI algorithm | Network management rules set by the vendor | AI algorithm | Network management rules set by the vendor | AI algorithm | Network management rules set by the vendor | AI algorithm | Network management rules set by the vendor |
| Total number of original alarms | 5343709 | | 5433432 | | 299341 | | 3304581 | |

| Alarm Handling Result in Pilot Cities | City A (1 month) | | City B (1 month) | | City C (1 month) | | City D (2 months) | |
|---|---|---|---|---|---|---|---|---|
| Filtering rate of alarms on the user side | 84.10% | - | 44.30% | - | 42.40% | - | 11.20% | - |
| Filtering rate of frequently-generated alarms | 10.30% | | 31% | | 54.20% | | 63.50% | |
| Filtering rate of correlative alarms | 1.10% | | 6.30% | | 1.90% | | 7.00% | |
| The proportion of correlative alarms to valid alarms | 19.60% | | 25.50% | | 55.80% | | 27.50% | |
| Total filtering rate | 95.50% | 45.1% | 81.60% | 21.60% | 98.50% | 0.30% | 81.70% | 13.10% |

Table 3: Analysis and handling the results of historical network alarms in pilot cities

**Solution System Optimisation Suggestions:**

- Collection mode optimisation: The resource information can be automatically collected over the northbound interface, or the offline reports can be automatically uploaded periodically. This makes the system functions autonomous and more simplified.
- System maintenance: To make the system valuable for engineering applications, the system can be further evaluated and verified on the live network. The system can be improved and optimised based on issues reported by frontline O&M personnel.
- Multi-dimension alarm analysis: Alarms can be analysed from multiple dimensions based on diversified O&M data sources, including alarm data, service configurations or statuses, KPIs, topology resources, operation logs, and fault rectification records. This makes fault alarm compression more useful and practical and results more accurate and referential.

**Unified and Applicable Correlation Rule Library Building**

Due to differences in vendors and regions, IP RAN alarm correlation rule library needs to be tested for vendors in different cities and sorted up for optimisation. Operators are expected to share mature rule library and rule sorting experience.

**Extended Application of Cross-Domain Diversified Alarm Handling**

When a fault occurs on a cross-domain network, alarms are generated for different domains. These alarms are strictly correlative. Other AI technologies can be introduced to implement fault demarcation and location for cross-domain alarms innovatively using multiple diagnosis technologies.

## 3.2.3 Weak Optical Signal Detection of the Access Network

The following figure shows the general methods and procedures for rectifying the weak optical signal in the industry:

- The weak optical signal of a single collection is affected by factors such as equipment being power-off or busy. The collection rate is low (60%-70%), and the weak optical signal situation cannot be evaluated adequately.
- The weak optical signal data is not analysed or analysed and sent to the front-line maintenance personnel directly, which causes repeated site visits and low efficiency for detection.
- The analysis results cannot be delimited to solve the problems such as the weak optical signal and optical splitting ratio beyond the allowed range, causing disputes between departments.

Difficulties in weak optical signal detection:

- It is difficult to delimit the weak visual fault point of the ODN. The accuracy rate is low because the current method depends on a manual judgment. Poor-quality links need to be checked segment by segment, resulting in low efficiency.
- The unreasonable networking mode of multi-level splitting cannot be found out.
- The timed data collection policy results in incomplete data and weak effect.
- Inaccurate data affect on-site detection.

The weak optical signal processing flow is as follows:

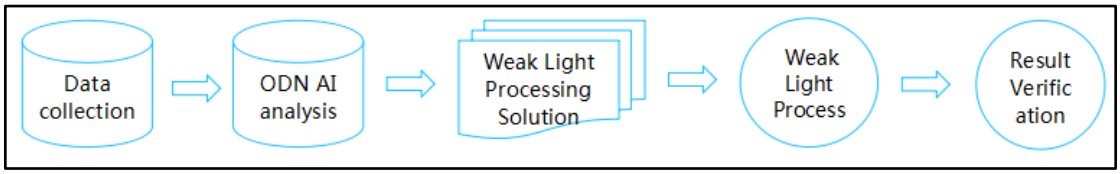



Figure 21: Weak optical signal detection flow

The analysis process uses the AI algorithm to analyse and determine the root cause of a weak optical signal and guides ODN optimisation. The idea is as follows:

- For ODN networks with various optical splitting ratios and downlink networking, the unsupervised learning algorithm is used to convert complex optical power data that is difficult to be directly applied to visualised data.
- Using a classification algorithm to determine the accurate position of weak light and effectively improve the successful detection rate.

Algorithm: The clustering algorithm such as k-means++ is used. SVM, Bayesian Network, and other supervised learning algorithms are used.

Interface: FTP interface between EMS and the work order system.

Data sets: ONU receiving and sending optical power, PON interface receiving and sending optical power, and the distance between ONU and OLT.

With this method, a carrier in China achieves the detection success rate of 85%. It takes 60 minutes per PON interface to check a weak optical node and repetitious on-site visits are required traditionally. After AI analysis and optimisation is used, the total workload of each PON interface is reduced to 20 minutes.

The databases of the transmission power and receiving the power of optical modules at both ends of the optical link and the distance between two ends of the optical link should be built in the industry to obtain better detection results.

This case can be applied to more scenarios，such as WDM PON in the 5G network.

### 3.2.4 Root cause analysis of wireless alarms

With the centralised OMC and 5G construction, the network scale is becoming larger and more dynamic, and alarm monitoring has the following problems:

- A large number of symptom alarms inundate the cause of alarms.
- The network is extensive and complicated, and it is challenging to make alarm deduction and correlation analysis rules.
- The network varies, and unified static rules cannot maximise alarm deduction ratio.

Out-of-service alarms of cells and eNodeBs needs to be processed with high priority in wireless O & M. If such an alarm occurs, O & M engineers need to handle it quickly.

There are various reasons for the service outage. In particular, a large number of the same alarms are reported by a single base station or multiple base stations caused by external reasons such as abnormal power and transmission. The same reason causes this type of alarms, but it takes much time and effort to locate the fault and find out the cause manually.

With the gradual deployment of 5G, the network structure is more complicated, and it is challenging to locate cross-layer faults. Therefore, determining the root cause of out-of-service alarms is more challenging.

The purpose of intelligent alarm root cause analysis is to learn and analyse existing alarms automatically, find out the relationship of various alarms, which then can be confirmed by a human, and implement automatic alarm correlation analysis and alarm compression.

Based on the network topology in the system and the monitoring data including alarms and events, operation logs and fault troubleshooting history records, the AI-based intelligent fault diagnosis outputs a series of rules of fault characteristics and causes. According to fault characteristics and the automatically match with the diagnosis rules, the fault points and related handling suggestions are automatically obtained in actual network operation and maintenance.

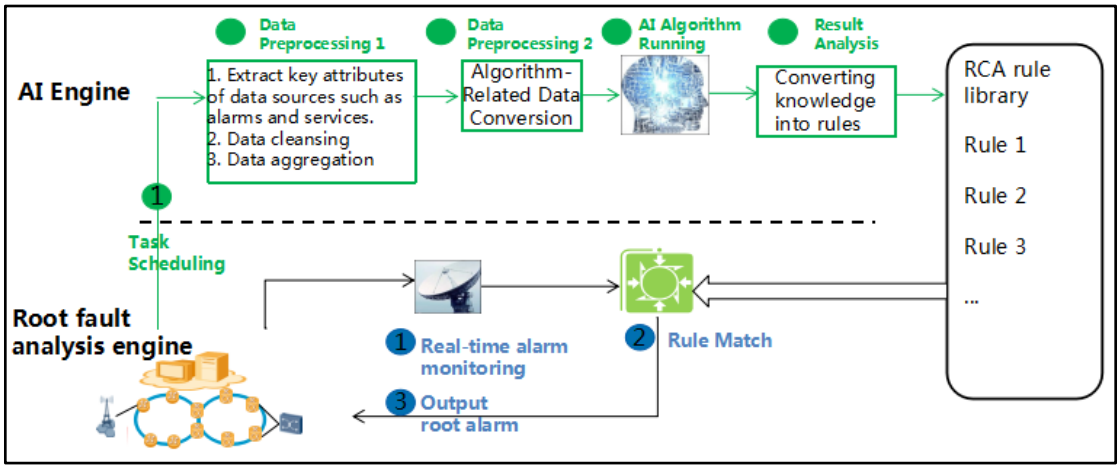



Figure 22: Alarm Root Cause Analysis Process

**Phase 1: Use the AI engine to identify rules.**

Task scheduling: A task can be scheduled periodically or manually following project requirements.

Data pre-processing: Extracts data from the database and converts it into the format required by the algorithm.

An algorithm is running: The algorithm operates based on input data and outputs results.

Result analysis conversion: Converts the result into service rules.

**Phase 2: Use the root fault analysis engine for fault process.**

Real-time alarm monitoring: Monitors real-time alarms of the production system.

Rule matching: Matches the real-time alarms with deployed service rules and identifies the root cause of faults.

Output result: The relationship between root alarms and other alarms is output, and the root cause is identified.

Applicable algorithm: The root cause analysis algorithm (many sequence/item set algorithms, linear correlation algorithm) realises alarm association analysis and root cause analysis. The clustering algorithm implements the topology grouping of base stations.

Interface: The alarm root cause analysis model collects alarm data at the scheduled time through interfaces such as FTP/RESTful.

Data involved: Alarm data, configuration data, resource data, operation logs and O & M knowledge base.

After the trial in China Mobile and China Telecoms 4G/5G live network, the out -of-service alarm reduction rate is higher than 45%. One branch of China Mobile manages over 6000 4G sites and collects 1108 out-of-service alarms within three days. The data is showed as follows:

| Methods | Alarm Automatic Analysis Percentage | Alarm Compression Ratio |
|---|---|---|
| Not Using AI | 0 | 0 |
| Using AI | 77% | 45% |

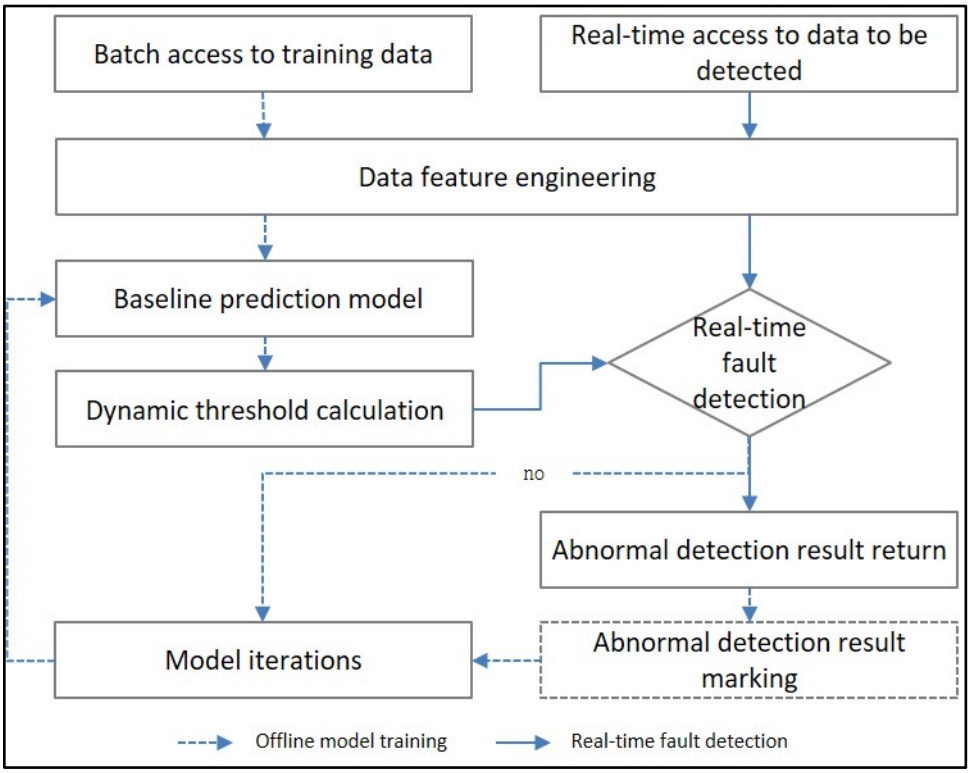
Table 4: Compression Effect of Out-of-Service Alarms

Currently, this solution has a noticeable effect on trials in multiple projects. It applies to both 5G and IoT.

### 3.2.5 Cross-domain Intelligent Alarm Root Cause Analysis

As network technologies continue to develop, the network structure is more complex, and network O&M and troubleshooting become more challenging. After a fault occurs, the traditional method is to manually check and analyse alarms one by one based on experience and preset check rules, which is time- and labour-consuming. For complex situations, multiple departments need to work together, leading to the inefficient fault location. In the 5G era, a new hierarchical-decoupled network architecture is used. The number of alarms to be monitored will increase exponentially. The traditional fault location method with preset check rules may not be applicable any more. Network fault management faces significant challenges. Therefore, advanced technologies need to be adopted for quick root cause location and alarm convergence. This helps improve O&M efficiency, ensure operation quality and lower operating costs.

This solution uses AI algorithms for alarm root cause analysis (RCA). The basic idea is to analyse massive historical alarm data, build models based on resource and topology data, and implement dynamic mining of alarm RCA rules without manual operations, supporting quick fault location and accumulating an O&M knowledge base. The following figure shows the process.



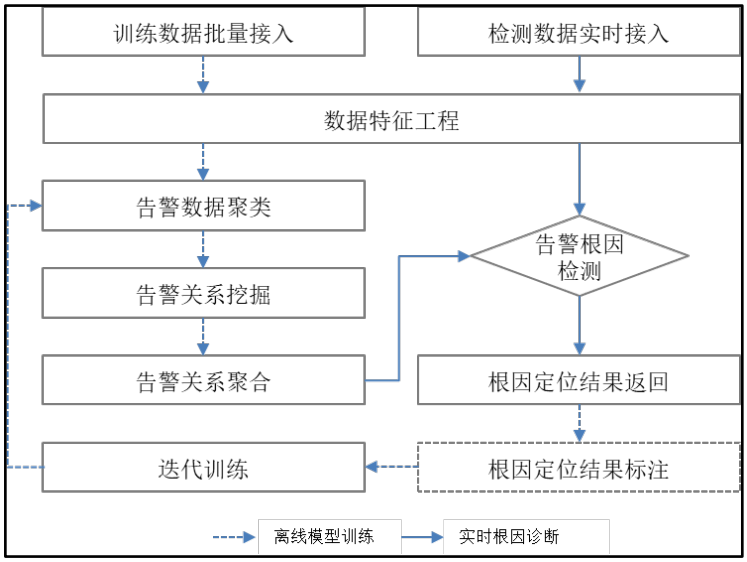




Figure 23: Figure 1 Intelligent alarm RCA process

**Data Access**

During alarm RCA model training, datasets (alarm data and resource and topology data) are imported using files in batches. Alarm data refers to alarm details and related maintenance tables. After correlation, datasets for model training should include at least the alarm subjects, types, and occurrence time. Resource and topology data can be classified into many types, including network topology, system deployment relationship, and service invoking relationship data. Such data can be analysed and processed to build spacial relationships between alarm subjects.

During real-time alarm root cause diagnosis, root cause diagnosis requests are sent through messages or APIs based on real-time alarm information.

**Data Feature Engineering**

For the intelligent alarm RCA solution, feature engineering requires the following operations in addition to routine operations such as data parsing and missing data processing:

- Filter data not needed for model training (for example, alarms that are recovered or manually cleared).
- After topology data parsing, connect alarm objects based on primary keys, create a topology between alarm objects, and divide the topology based on connections.

**Alarm Data Clustering**

Alarm data clustering is the basis of alarm correlation mining. The clustering algorithm is selected from DBSCAN, HDBSCAN, OPTICS, Birch, Agglomerative, and GMM based on the clustering effect, parameter tuning difficulty, and running speed.

Both time- and space-based clustering are taken into account. Time-based grouping refers to clustering based on the initial occurrence time of alarm data. Space-based clustering refers to clustering based on spatial distances between alarm objects after topology division.

**Alarm Relationship Mining**

The correlation mining algorithm is used to find further out correlations between alarms based on the clustering results. In this solution, each clustering result is classified as a project set. Based on these project sets, the correlations between two alarm data records are explored. The correlation results are filtered based on indicators such as support and alarm RCA improvement, and the primary and secondary relationships between alarm data pairs are further determined based on confidence.

**Alarm Relationship Aggregation**

Alarm data relationships need to be further aggregated based on the primary and secondary alarm dependency table obtained through correlation mining to generate an alarm relationship map. This can meet the requirements of scenarios, such as real-time alarm root cause detection.

**Alarm Root Cause Detection**

During alarm root cause detection, root alarms are located from real-time alarm information and processed promptly within a specified window based on the alarm relationship map.

**Result Labeling and Iterative Training**

To ensure the model location effect, incremental training tasks need to be periodically launched during production. Also, O&M personnel can mark and report the located root alarms based on experience and actual conditions. Then, the system will automatically add the marked data to the next iteration.

A primary and secondary alarm dependency test is conducted based on the other cross-domain intelligent alarm RCA solution and alarm data on the cloud management platform of a provincial operator. The algorithm is verified and optimised. The potential alarm relationships among network devices, hosts, databases, middleware, significant data components, and DCOSs are explored. The following shows the details:

- Samples: alarm data of three months, DCN network segment table, and topology of alarm objects
- Algorithm selection: DBSCAN, Birch, Apriori, and FP-growth

- Alarm relationship mining: A total of 200 relationships are discovered when the confidence threshold for alarm data dependency is limited. The accuracy rate is verified to be higher than 60%. The following figure shows part of an alarm relationship map generated according to the primary and secondary relationship table. The dots in four blue circles indicate possible root cause nodes, which can be applied to online real-time alarm root cause detection.

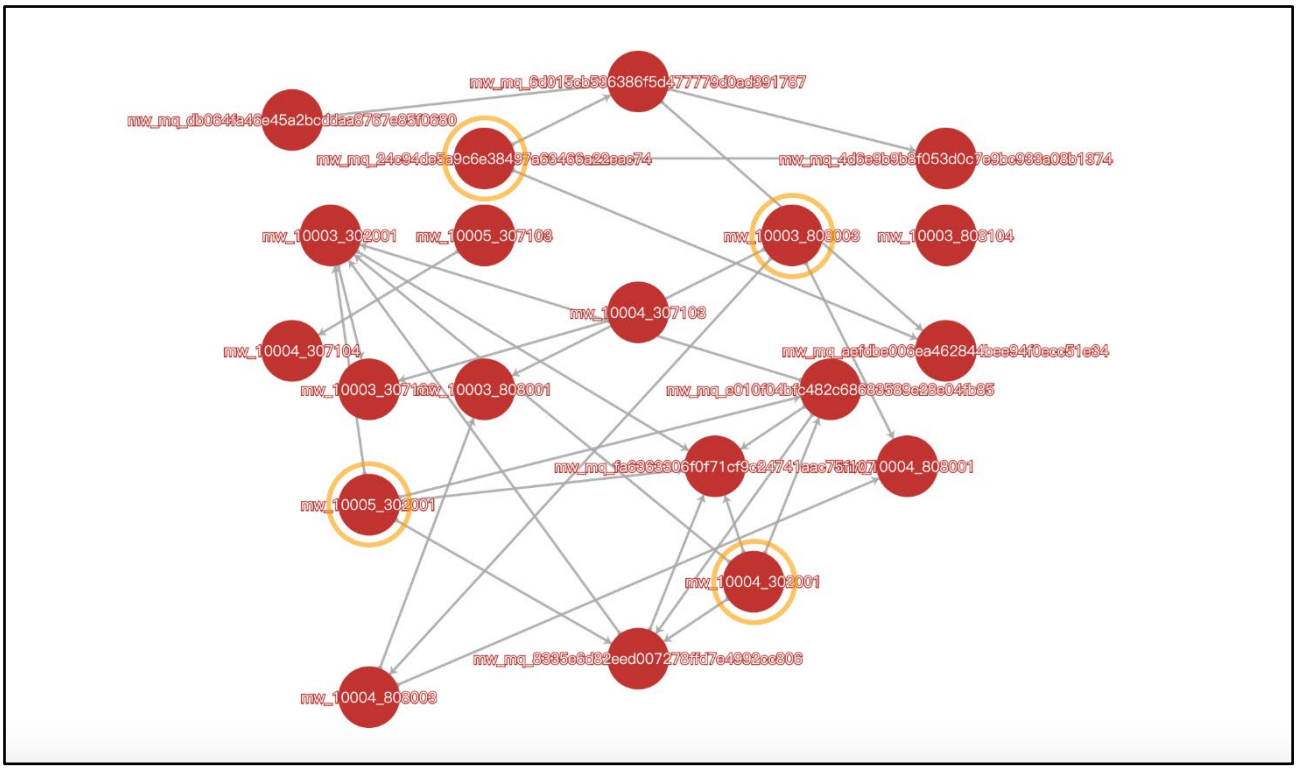



Figure 24: Example of the alarm relationship map

To perform more accurate space-based clustering and avoid interference between irrelevant data, it is recommended that resource data be further standardised. That is, urge each system to further optimise resource data such as network topology, deployment information, and invoking relationships of related alarm objects and standardise data interface specifications of alarm object topology. This will make the solution more adaptive to different scenarios and reduce the workload required for customisation and reconstruction.

Also, the other solution has not been tested in 5G scenarios. In the future, the processing logic and procedure of feature engineering need to be optimised and adjusted based on slice O&M data. In this way, this solution is expected to meet 5G network O&M requirements such as cross-layer cross-domain alarm analysis and fault diagnosis for slice O&M.

## 3.2.6 Dynamic Threshold-based Network O&M Exception Detection

As for network O&M, a standard method to detect exceptions is to check time sequence indicators. Traditional approaches often use fixed thresholds that are manually set. These thresholds need to be set for different types and instances based on experience to make exception detection more accurate. Such methods are simple, straightforward, and easy to operate, but require heavy configuration and maintenance workload and rely on experience. Also, exception detection using fixed thresholds is not sensitive to local exceptions that occur within a period in a cycle. As monitored objects and indicators grow exponentially, methods using manually-set set limits present more disadvantages, such as missing,

incorrect, and massive alarm reporting. In this case, intelligent methods using AI algorithms need to be introduced to improve alarm accuracy and lower manual configuration costs, accurately and automatically detecting exceptions in time.

In this solution, dynamic thresholds are used for time sequence indicator exception detection. The basic idea is to establish a common detection model framework, invoke AI prediction algorithms for model training for different time sequence indicators based on massive historical data, and add threshold ranges based on the predicted values to obtain dynamic thresholds for a later time. During real-time detection, the system checks whether specified indicator values are within the threshold ranges to detect exceptions. In addition, to continuously improve the accuracy of a dynamic threshold–based exception detection, marks can be manually added for automatic reinforcement learning and model optimisation.

The following figure shows the process:

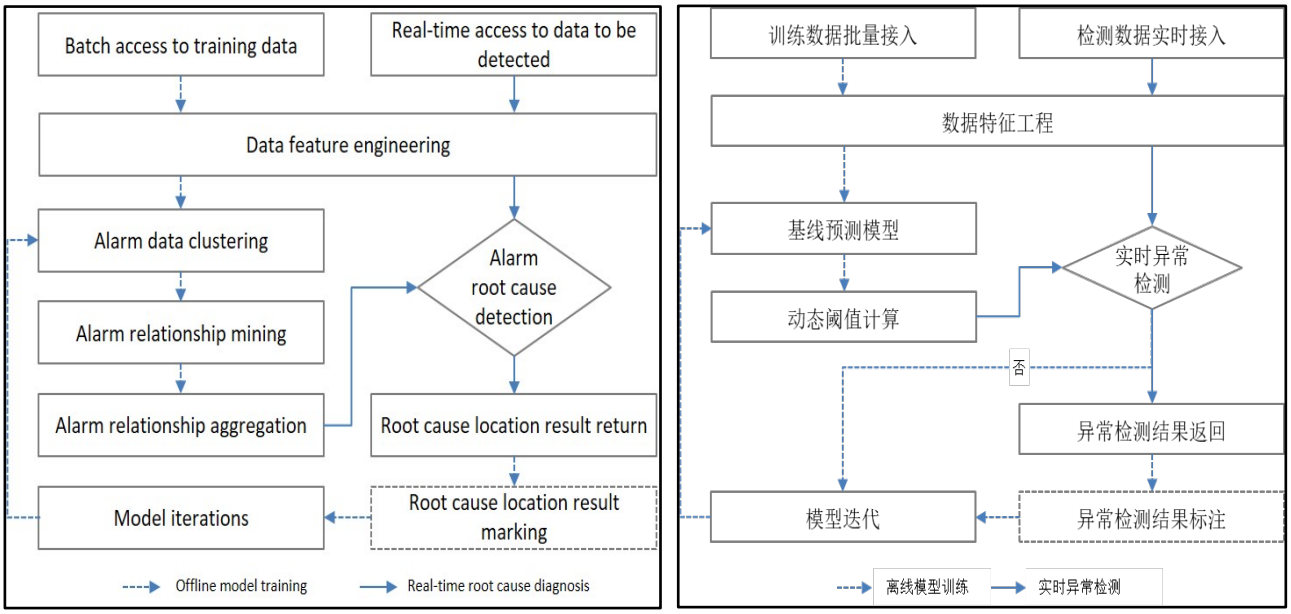



Figure 25: Time sequence indicator exception detection process

**Data Access**

Dynamic threshold-based exception detection supports time sequence data with a time granularity of minute, hour, day, or month. Such data can be obtained from various monitoring platforms such as the 5G NMS, integrated NMS, and professional NMS. The collected data must contain at least instance IDs (which can be a combination of multiple attributes), indicator data for training (for example, numerous training tasks are generated if multiple indicators are accessed at a time), indicator data time, data frequency, and other fields.

During model training, time sequence indicator data is imported using files in batches. During real-time exception detection, exception detection requests are sent through messages or APIs.

**Data Feature Engineering**

The accessed data needs to be parsed, cleansed, and converted, and its characteristics need to be extracted. Operations include missing value processing, deletion of abnormal samples, data time alignment, standard conversion, statistics collection, and characteristic analysis and comparison.

**Baseline Prediction Model**

Prediction model training mainly targets indicator baseline prediction and is divided into initial model training and iterative model training. Initial model training can be initiated upon reception of an offline training request from a third-party system. Iterative model training is undertaken based on a specified training period and continuously optimises models based on real-time accessed standard data and marked data.

This solution does not specify any algorithm. An optimal algorithm is selected from machine learning and deep learning prediction algorithms including ETS, ARIMA, LSTM, FB-PROPHET, and TBATS through automatic hyperparameter selection and effect evaluation, to make the framework more adaptable to dataset exception detection requirements in different scenarios.

**Dynamic Threshold Calculation**

Dynamic thresholds at different time points can be obtained by adding threshold ranges based on the indicator baseline generated by the prediction model. The threshold ranges can be calculated using the fluctuation rate set according to experience or automatically calculated based on the fitting effect. This solution considers residual indicators with different confidence levels as value ranges by analysing statistics characteristics of the remaining sequence and obtains threshold ranges for different levels at different time points based on the predicted baseline. This helps measure exceptions in the past and future time sequences more accurately.

**Real-time Exception Detection**

After the time sequence indicator data is accessed and preprocessed in real-time, intelligent exception judgment is performed according to whether the detected indicator values are within the dynamic threshold ranges at the current time. The detection results (including exception detection results, causes, fitted values, upper thresholds, and lower thresholds) are returned for upper-layer alarm policy management.

**Abnormal Detection Result Marking**

O&M personnel can mark the diagnosis results and provide feedback based on experience or actual conditions. The system automatically initiates training based on modified data.

**Model Iterations**

To ensure the accuracy of baseline prediction, the number of data points predicted each time is limited. Therefore, incremental training tasks need to be periodically initiated. During iterations, incremental data can be periodically accessed, and non-abnormal sample data obtained after real-time exception detection can be automatically stored and added to the next iteration. Besides, whether corresponding data is integrated into training can be adjusted based on manual marking results to implement intensive training.

This solution is piloted for testing forwarding duration time sequence data during load balancing for a type of devices of a provincial operator. The exception detection rate is proved to exceed 90%. The following figure shows the detection effect.

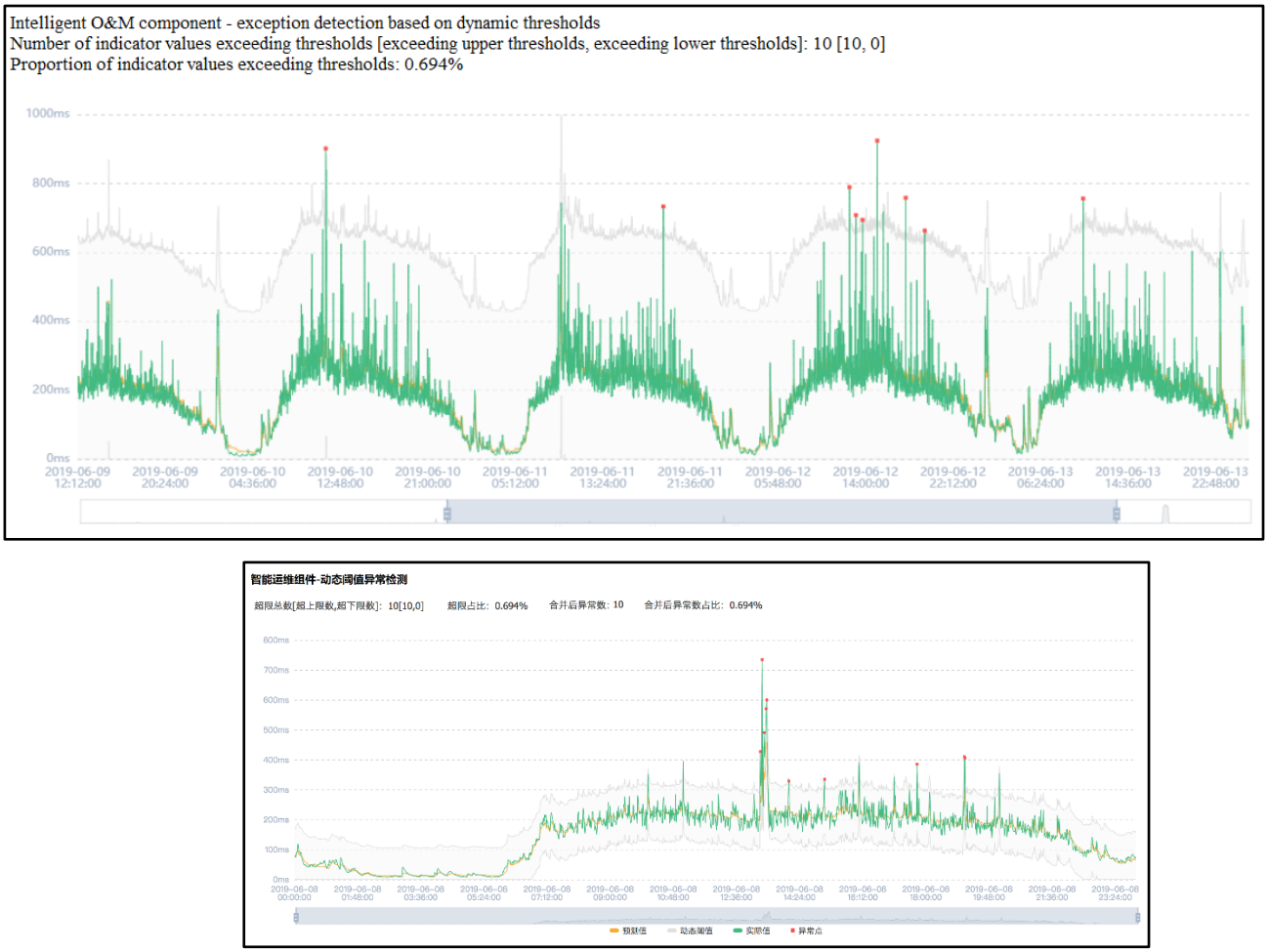



Figure 26: Effect of a dynamic threshold–based exception detection

As shown in the preceding figure, ten exceptions are detected for a specific instance within a day. Compared with the original method of setting the threshold to 500 ms, dynamic threshold–based exception detection is more autonomous and flexible, which effectively reduces the proportion of missing and incorrect exception reports.

There are a large number of time sequence indicators in the monitoring indicators of various network systems and network management platforms. To improve the efficiency of exception detection and avoid unnecessary investment in model calculation resources, it is recommended that monitoring KPIs of network domain systems including the 5G NMS be clarified and standardised, and that core time sequence indicators be detected using AI algorithms and framework. This could reach a much better exception detection effect.

In addition, exception detection in this solution is based on thresholds. As indicators such as network performance and traffic fluctuate considerably, it is recommended that the network management system uses corresponding alarm policies together with exception detection, determine whether to deliver an alarm based on how much a single indicator value exceeds the threshold, number of consecutive exceptions, and exception severity. This prevents false alarms caused by single-point exceptions.

## 3.2.7 Gene Map-based Intelligent Alarm

With more complex networks, diversified services, large-scale customers, and various terminals, requirements for network service quality are higher. The traditional methods (through complaints, alarms, and performance indicators) for detecting network faults present the following issues:

- The complaint process is time-consuming. Frontline maintenance personnel cannot be notified of network faults upon their occurrences.
- Network faults cannot be accurately identified, and their impacts on services cannot be evaluated from massive alarms.
- Performance indicator warnings are delayed, and user experience issues cannot be accurately located.

A new solution, intelligent monitoring using gene maps, is designed to make breakthroughs in scenario-based control, achieve an intelligent production process, explore next-generation network monitoring, and comprehensively improve the centralised monitoring capability.

Signalling that controls the setup, monitoring and removal of voice or data services is called network genes. This solution utilises network genes, reduces intermediate processes, proactively detects potential network risks based on big data analysis, and takes optimisation measures before the service impact increases. This prevents critical faults and improves network O&M efficiency and user experience.

Network genes are massive, uncertain, and fluctuating. Therefore, network gene data has big data characteristics of large volume, variety, velocity, and complexity. When network or service exceptions occur, the gene structure always changes dramatically in a short period, especially for genes indicating service failures. Such symptoms are called "gene mutation". To implement gene-based real-time network monitoring, gene data is monitored for fault detection.


Figure 27: Characteristics of network gene data

This solution uses the provincial performance big data platform to collect data from significant gene interfaces on the entire network, draw gene maps for each critical service process based on studies on gene characteristic identifications (release code, cause value, and status code), and analyse data to explore new methods for detecting network faults. The following figure shows status code–based distribution.

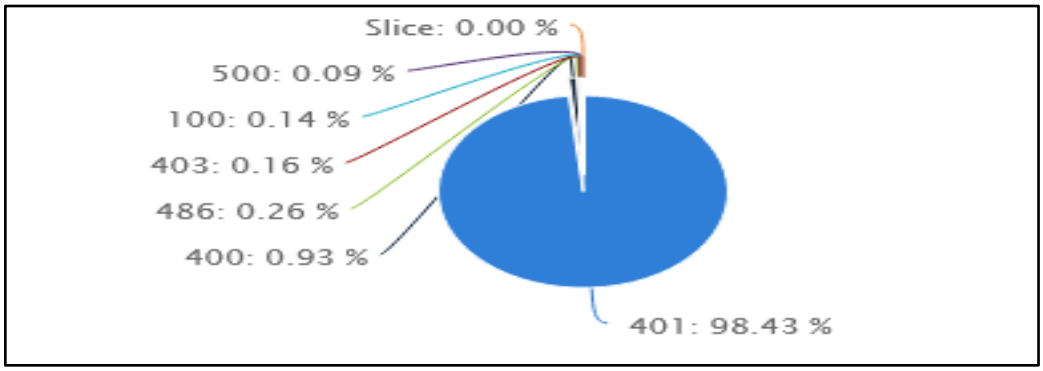



Figure 28: Interface status code distribution

The following takes the IMS status code as an example to describe the procedure.

1. **Selecting Status Codes**

   Obtain periodic status codes through manual statistics and apply them to time sequence algorithms. Collect data of at least one week and select status codes whose integrity rate is higher than 70%.

2. **Differentiating Status Code Data**

   Perform time sequence differentiation for data corresponding to the status codes, as shown in the following figure.

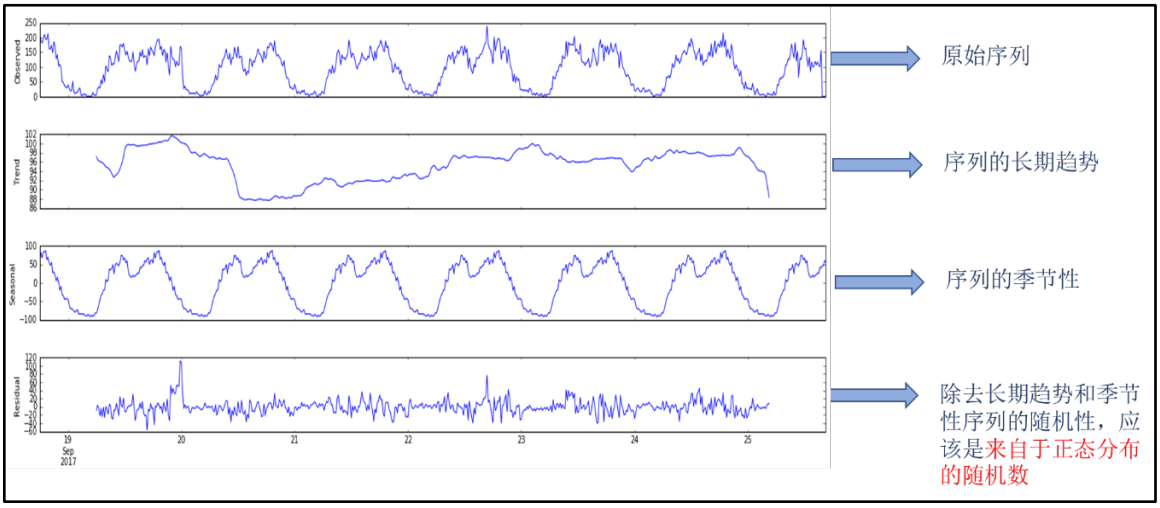



Figure 29: Differentiation for data corresponding to the status codes

3. **Setting the Alarm Threshold**

   Select a random sequence (R sequence) after differentiation and verify its distribution. R sequence is proved to be a stable sequence and meet the normal distribution. During parameter commissioning, it is found out that **±4σ** can be set as the alarm triggering condition as its fault detection rate can reach as high as 92%. The following figure shows a status code–based alarm reference with **±3σ** and **±4σ** as the alarm triggering conditions. The red and blue lines indicate the thresholds.

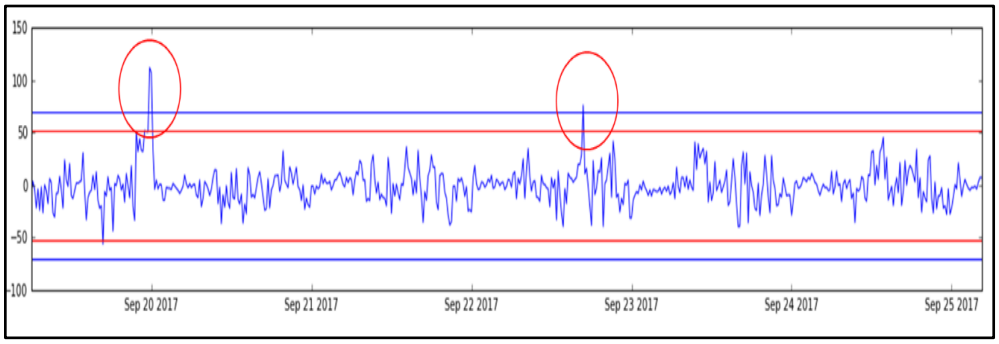



Figure 30: Example of alarm thresholds and faulty points

4. **Analysing Genes and Drawing Maps Based on Alarms**

   Obtain alarm data corresponding to the status codes and generate alarms. Then, analyse indicators for network gene data for these alarms and identify abnormal indicators.

   A network gene map is drawn based on failure causes determined in the analysis result, as shown in the following figure:

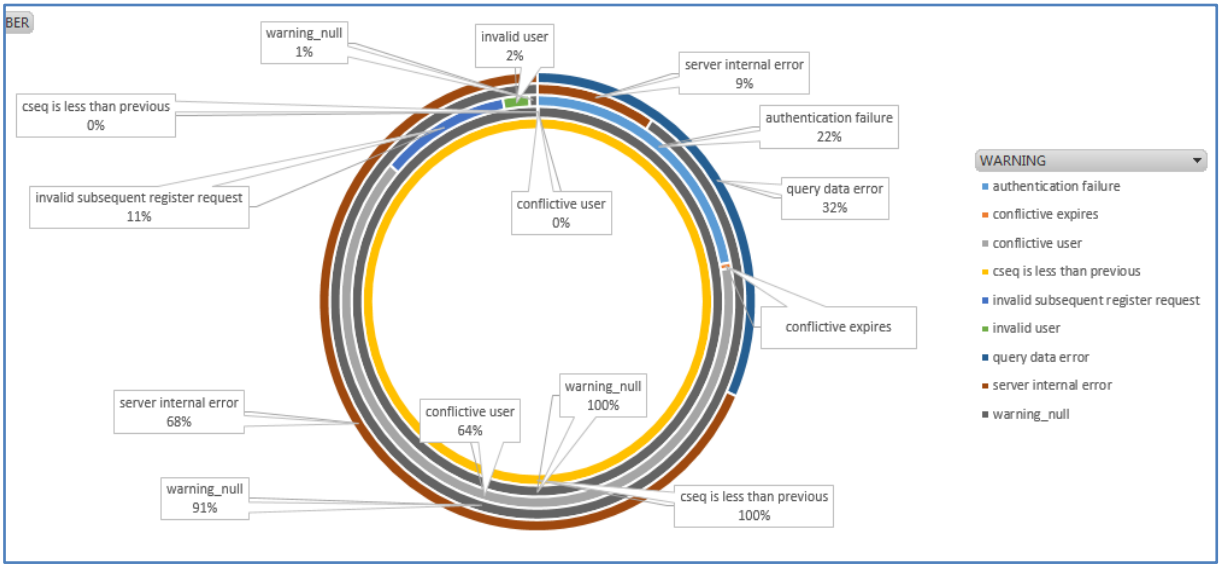



Figure 31: Example of a network gene map

A live-network rule library in each region is summarized based on accumulated experience and detected faults to automatically demarcate and locate faults, identify affected users, and implement one-click analysis of complaints.

5. **Case of CE Replacement on the IP Bearer Network**

   On the next day after CE replacement on the IP bearer network, users complained that VoLTE calls occasionally failed. The indicator value for the VoLTE network connection rate decreased by 0.04%, which does not reach the threshold for triggering the performance indicator deterioration alarm. The intelligent monitoring system found out that the number of reported interface status code 480 increased by 15 folds. After inspections, the cause was confirmed to be a faulty port during CE replacement to HSTP equipment, as shown in the following figure.

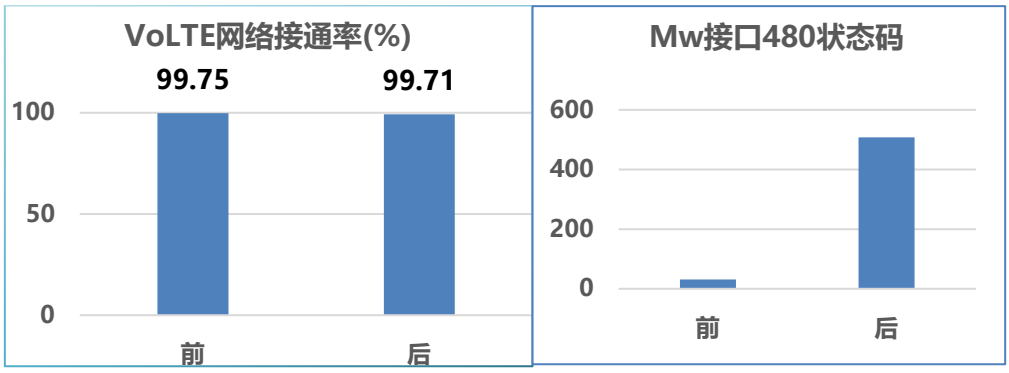



Figure 32: Changes in the network connection rate and status code before and after the VoLTE network is faulty

6. **Case of a Board Fault Alarm**

An SAE GW board fault alarm was reported on the live network provided by an operator. After confirmation, the board was manually replaced. The whole handling process took about 10 minutes. However, the fault impacts on services and users cannot be quickly evaluated. The number of status code 503, indicating VoLTE call connection failures increased sharply. It can be rapidly summarised that user call connection was affected during the board replacement, as shown in the following figure.

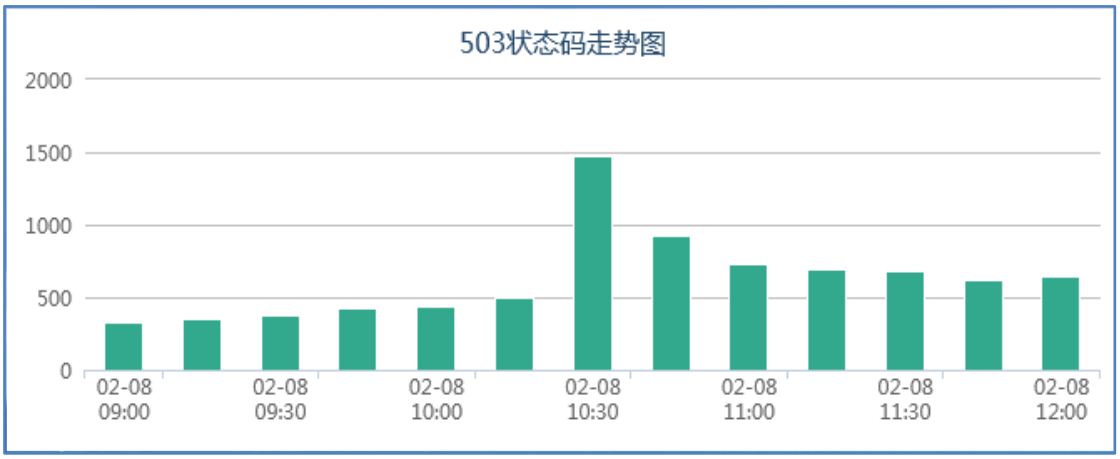



Figure 33: Status code trend chart

7. **Case of PCRF Upgrade Tool Bugs**

In December 2017, bugs in the PCRF upgrade tool of a vendor caused the active and standby boards to restart at the same time, affecting the subscribers who were initiating registration requests. The number of status code 503 was detected to be abnormal. Its actual value was **1842**, 205% more than that of the threshold (**603**), as shown in the following figure.

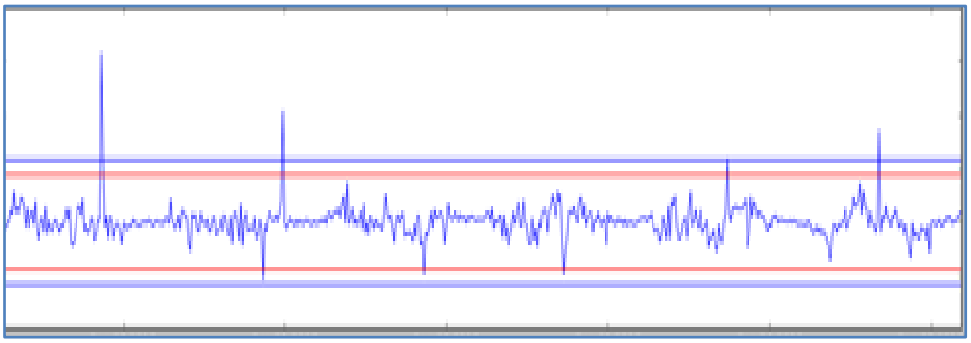


Figure 34: Time sequence differentiation of status code 503 - R sequence

8. **Case of a Fault in January 2017**

   On a day in January 2017, a network fault occurred at 16:25, causing VoLTE call connection failures. Five minutes later, numbers of reported status codes 487 and 504 were detected to be abnormal. An alarm message was delivered at 16:38, and user complaints were received at 16:50. The alarm message indicating a decrease in the VoLTE call completion rate was received at 17:20. The following figures show the status codes that trigger alarms.

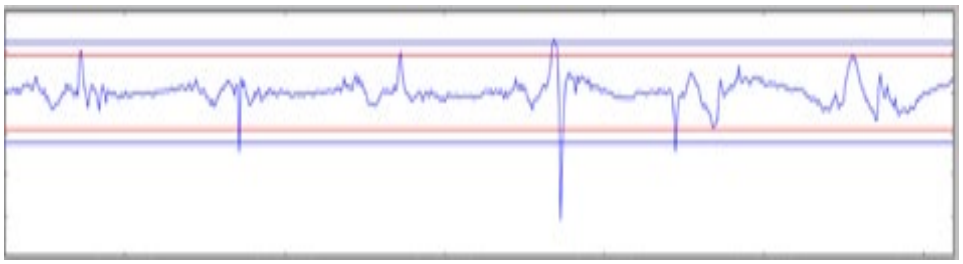


Figure 35: Status code 487 (the calling party hangs up after not connected for a long time) triggering alarms

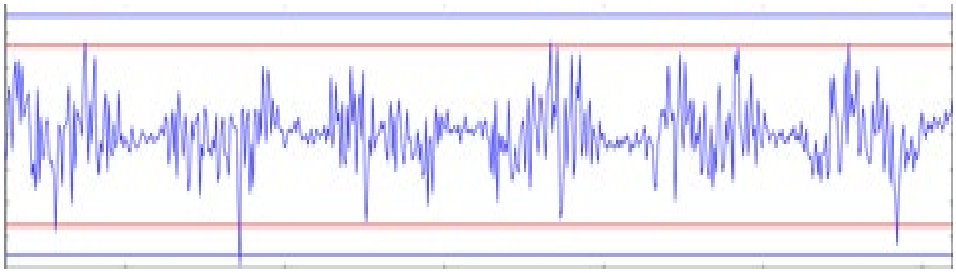


Figure 36: Status code 504 (the calling party hangs up after not connected for a long time) triggering alarms

**Improving Alarm Timeliness**

The database performance and algorithm model can be optimised to improve alarm timeliness. The monitoring and warning granularity needs to be shortened to 5 minutes from 15 minutes.

**Optimising Alarm Correlation Rules**

The network gene correlation map can be optimised to make automatic fault demarcation and location more accurate and gradually build the intelligent autonomous O&M capability for wireless networks.

## 3.3 AI for Network Optimization and Configuration

### 3.3.1 5G Intelligent Broadcast Parameter Adjustment

Massive-MIMO technology is one of the critical 5G physical layer technologies. Compared with the traditional antenna, the Massive-MIMO equipment has the 3D beamforming capability, and can flexibly adjust the weight (power and phase) of each antenna array, thus significantly improving the beam direction accuracy of the system. It concentrates the signal on the specific area and specific user group while enhancing the user signals and considerably reducing the interference inside the cell and among adjacent cells.

To achieve the optimal coverage of Massive-MIMO equipment, a set of initial network parameters cannot meet the requirement. Therefore, the network broadcast parameters need to be adjusted in a differentiated manner. Manual adjustment takes time and effort. For 5G networks that support broadcast multi-beam scanning, thousands of beam parameter combinations increase the adjustment complexity. In addition, the coordinated adjustment between multiple cells is also difficult. Therefore, it is necessary to adjust broadcast beam parameters by using intelligent algorithms.

Intelligent broadcast beam parameter adjustment is applicable to the following scenarios:

- Hot-spot areas covered by multiple Massive-MIMO cells, such as stadiums, school and CBD areas;
- Massive-MIMO cell continuous coverage scenario.

The following figure shows the intelligent weight adjustment solution.

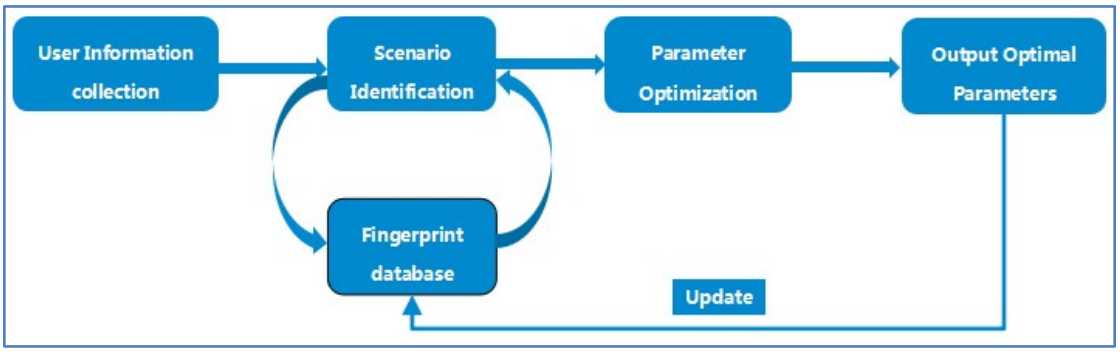



Figure 37: Intelligent Parameter Adjustment Solution

The solution includes four modules: User information collection, scenario identification, weight optimisation, and optimal weight output. The functions of each module are described as follows:

- User information collection module: Collects UE information data and provides data for subsequent scenario identification.
- Scenario recognition module Queries the historical fingerprint database based on MR information and recommends the optimal parameters for current scenario. This module classifies scenarios, and the parameter optimisation module uses the optimal parameters as the initial value for global optimisation.

- Parameter optimisation module: It is mainly used to get the optimal parameter based on the recommended parameter, to realise the self-adjustment and optimisation of Massive-MIMO parameters.
- Output optimal parameter module: Evaluates and compares the recommended optimal parameter output by the scenario identification and parameter optimisation module, and selects the best parameter. Give feedback about the information back to the fingerprint database, and updates and improves the fingerprint database.

Applicable algorithm: Considering complexity and overhead involved, the scenario identification module in the solution uses classical machine learning algorithms such as KNN algorithm, decision tree and logistic regression for scenario prediction and classification. In this solution, the parameter optimisation module can use AI algorithms such as the genetic algorithm, PSO algorithm, and ant colony optimisation (ACO) algorithm for parameter optimisation.

Data sets involved: UE periodical MR data, UE location information etc.

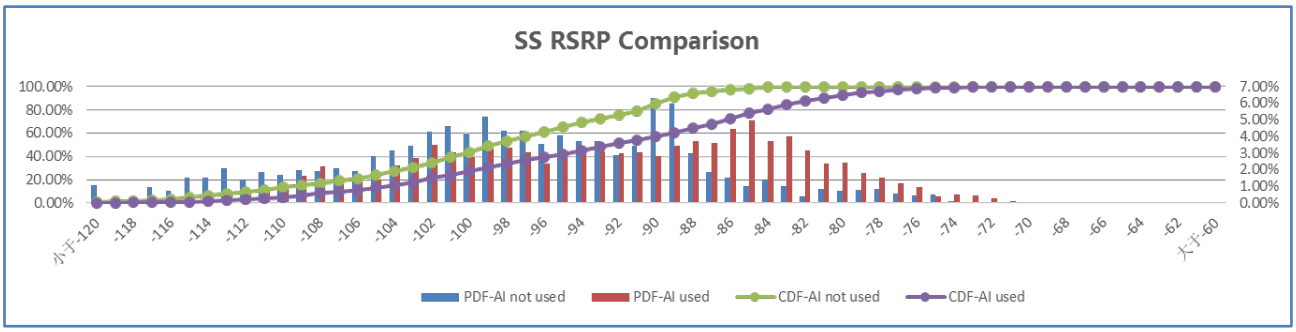



Figure 38: Test Result of Intelligent Parameter Optimization-SS RSRP

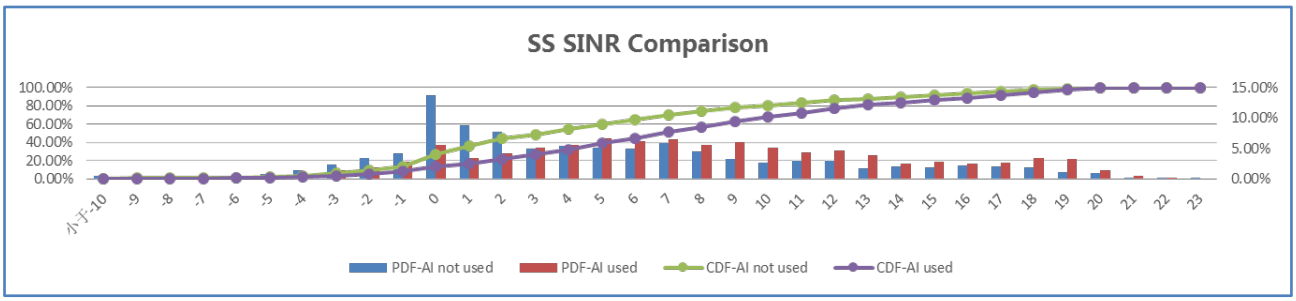



Figure 39: Test Result of Intelligent Parameter Optimisation-SS SINR

The data of base station height, latitude and longitude, and actual scene information (high building, railway and subway, etc.) are transferred from carriers' system to wireless network management system and serve as the input for Massive MIMO parameter optimisation.

## 3.3.2 RF Fingerprint Based Load Balancing

The load balancing based on RF fingerprints applies to the scenario where the load is high or unbalanced in the multi-frequency networking scenario. For example, in the four-carrier networking scenario shown in the following figure, F1 is the access carrier frequency. When there are a large number of UEs in the network, if all the UEs access the network through F1, the load of cell F1 will be increased, and F1 will become congested, even inaccessible. In this case, it is necessary to properly balance users in F1 to the

F2~F4 carrier frequency with the same coverage, balance the load among cells, make full use of system resources and improve user experience.

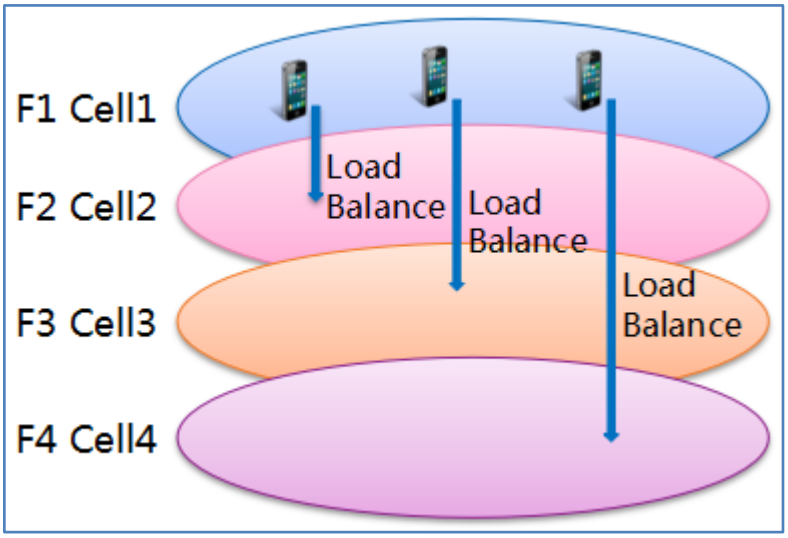



Figure 40: Load Balancing Scenario

In a multi-frequency network situation, load balancing between cells at multiple frequency layers is implemented quickly and accurately based on RF fingerprints to improve resource usage and user experience.

The RF fingerprint-based balancing solution includes the following aspects:

- The EMS first obtains historical UE measurement reports and handover information and constructs an RF fingerprint database to obtain the relationship between UE and the radio coverage of surrounding cells. Then, the system makes a validity judgment on the fingerprint database information and transmits a valid RF fingerprint database to the base station.
- The eNodeB monitors a load of each cell in real-time and triggers load balancing when the load unbalancing conditions are met.
- The eNodeB determines the load balancing target and performs load balancing based on the cell load, cell characteristics, UE characteristics, and the relationship between the UE and neighbouring cell radio characteristics (RF fingerprint database information). The load balancing target includes load balancing source and targets cell; the traffic volume needs to be balanced to each target cell, and the relationship between the migrated UE and the target cell.

While performing load balancing based on RF fingerprint, the base station assesses the RF fingerprint database. When it finds that UE handover successful rate is low, it originates a fingerprint database update request to the network management system.

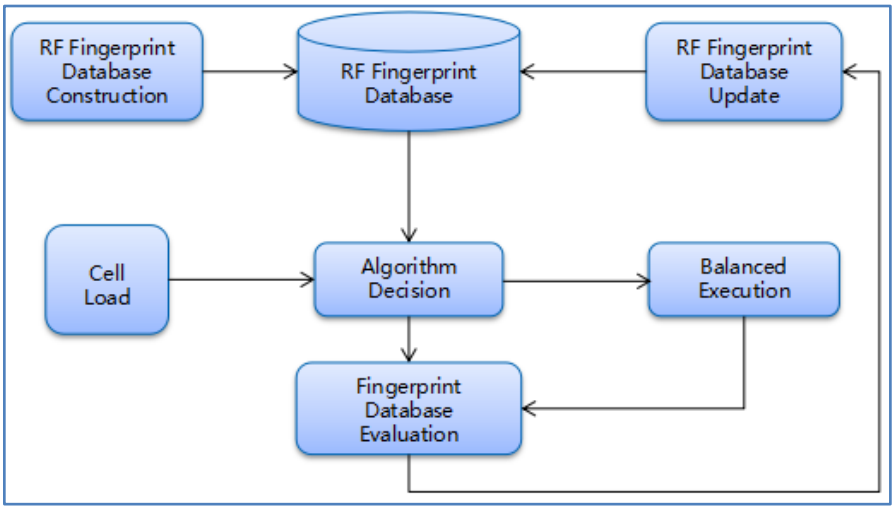



Figure 41: Load Balancing Solution Procedure

Algorithm: Clustering algorithm such as K-means

Interface: The interface between the EMS and the eNodeB is involved. The base station initiates a fingerprint database update request to EMS, and the EMS downloads the RF fingerprint database to base station.

Involved data sets: UE measurement reports (including intra-frequency periodical MR, intra-frequency event MR, inter-frequency periodical MR, and inter-frequency event MR), inter-frequency handover information data, and cell load data.

This solution monitors the load of each cell in real time, triggers load balancing actively, and balances loads more quickly. It has a comprehensive understanding of the load of each cell and an understanding of the coverage of surrounding cells. When deciding for load balancing target, it can set the load-balancing ratio more reasonably, and accurately select the target UE and target cell for load balancing to reduce unnecessary load balancing and measurement. Thus it can improve the balancing efficiency, and finally improve the resource usage ratio and user experience.

Compared with traditional load balancing, the RF fingerprint-based load balancing improves the accuracy of target UE and cell selection to avoid unnecessary UE measurement, improves the load-balancing efficiency, and makes inter-cell load balancing much faster.

The test is handled about load balancing in the lab and comparison are made between the traditional load balancing and the load balancing based on RF fingerprint in the environment with intra-frequency and inter-frequency neighbour cells. In the following scenario, Cell1, Cell2 and Cell4 are F-band cells, and Cell3 and Cell5 are D-band cells. Cell1 and Cell2 are at the same site, which is different from the site where Cell3, Cell4 and Cell5 are. The high-load Cell1 covers the target neighbouring Cell3, high-load Cell1 is a neighbour of Cell5.

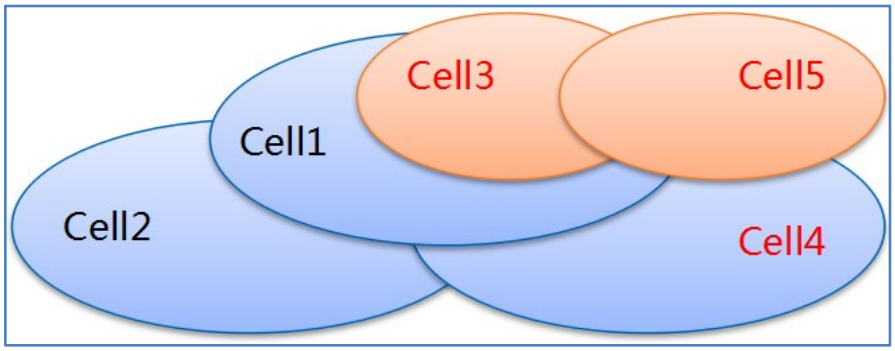



Figure 42: Load Balancing Test Scenario

The test result is as follows:

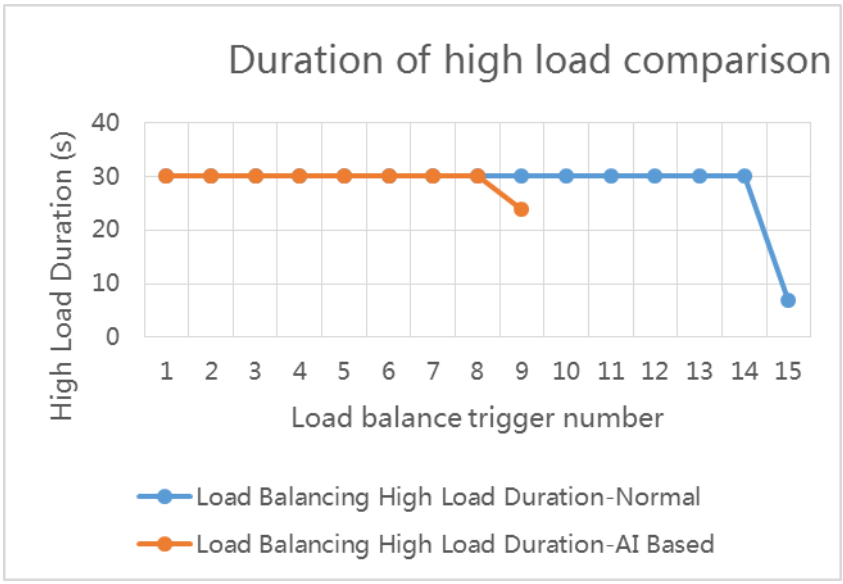



Figure 43: Test Result of Load Balancing – Comparison of Duration of High Load

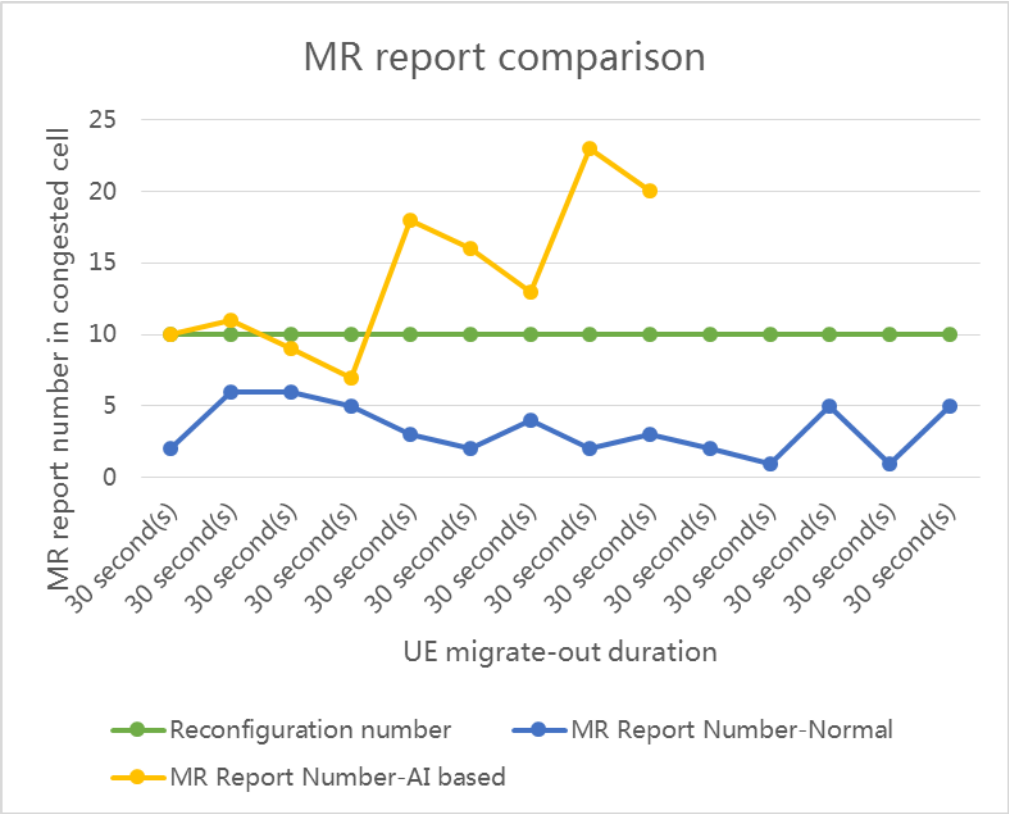



Figure 44: Comparison of MR Report Number in Load Balancing Test Result

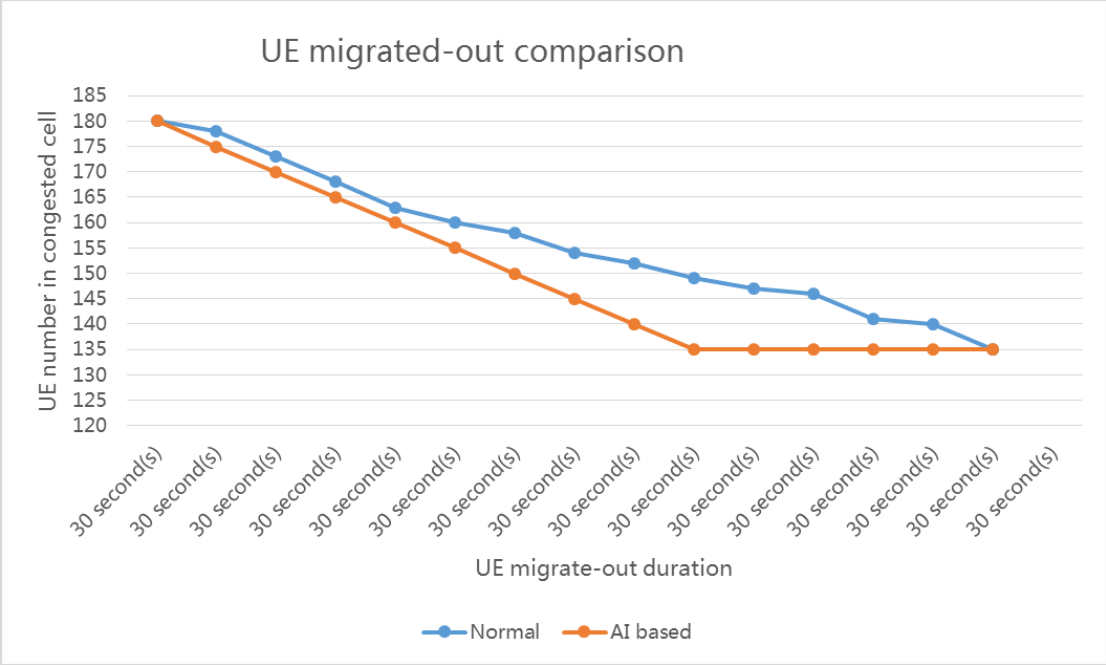



Figure 45: Test Result of Load Balancing -UE Migrate-out Speed

Test conclusion: After the RF fingerprint function is used, the load balancing target can be achieved in 10 load balancing periods, saving five load balancing periods, reducing the number of load balancing, moving out the UE faster, and improving the load balancing efficiency. In addition, the number of MR reported is higher than the number of reconfiguration issued, which indicates that the ratio of useful measurement performed by the selected UE is high and the measurement efficiency is also improved.

Standardisation suggestion: This function is closely related to the load balancing policy of NEs. Because the policies of different vendors are different, it is challenging to interwork between different vendors. However, this feature can be extended to 5G, and the construction and application of RF fingerprint database can be further optimised based on 5G features, for example, beam factors.

## 3.4 AI for Service Quality Assurance and Improvement

### 3.4.1 Intelligent Transport Network Slice Management

Current telecommunication services are diversified in different scenarios. If a dedicated network is established for each service, the cost is high. Network slicing enables multiple logical networks to share the same physical infrastructure through cloud and virtualisation technologies, effectively reducing the cost. Such sharing provides a new business model for flexible network services. In vertical industries, the network architecture with elastic resources will dynamically change according to service requirements. Compared with the traditional network, this mode is more flexible. However, flexible and dynamic requirements bring new challenges to the current network operation based on man-machine interaction.

Network slicing is a key enabling technology for 5G networks. A network slice is an end-to-end logical subnet, which involves the core network (control plane and user plane), radio access network, and transport network and requires coordination of multiple domains. Different network slices can share resources or be isolated from each other. Network slicing helps users implement desired functions and features, quickly deploy services, and reduce service rollout time. Considering the limited network resources and the network status of different network slices, operators need to reuse physical network resources as much as possible while ensuring the fulfilment of the service level agreement (SLA). To improve operation efficiency, operators need to optimise resource allocation for network slices.

On a transport network, to avoid resource shortage in peak hours, slices are generally deployed to meet peak user requirements. However, this also causes redundancy and waste of dedicated resources such as network bandwidth and QoS in most off-peak hours. Therefore, accurately predicting traffic usage, dynamically configuring slice resources on-demand, and intelligently managing transport network slices are essential to properly allocating network resources and ensuring service quality.

Intelligent slice management based on the purpose of serving users is an urgent requirement for automatic slice deployment.

The following figure shows the basic system structure of the Transport Network Slice Manager (TNSM), which consists of three subsystems:

- Slice manager: supports slice computing and device capability abstraction and mapping. It abstracts the forwarding behaviour of each device and various implementation processes of resource management, that is, it models the slicing function as a mathematical constraint, and describes the underlying physical network as an abstract network. In this way, the TNSM can perform network management without distinguishing a device type and a device management interface. It is also responsible for slice topology management, including node selection, physical interface selection, and subinterface bandwidth configuration.

- Slicing controller: verifies and delivers configurations of physical devices on the underlying network. It ensures that the configuration of slice creation and adjustment complies with the user intent and slicing rules, and then delivers the configuration to the corresponding device.
- Slice analyser: monitors real-time end-to-end information through a traffic generation device, including the throughput, delay, packet loss ratio, and sub-interface bandwidth usage.

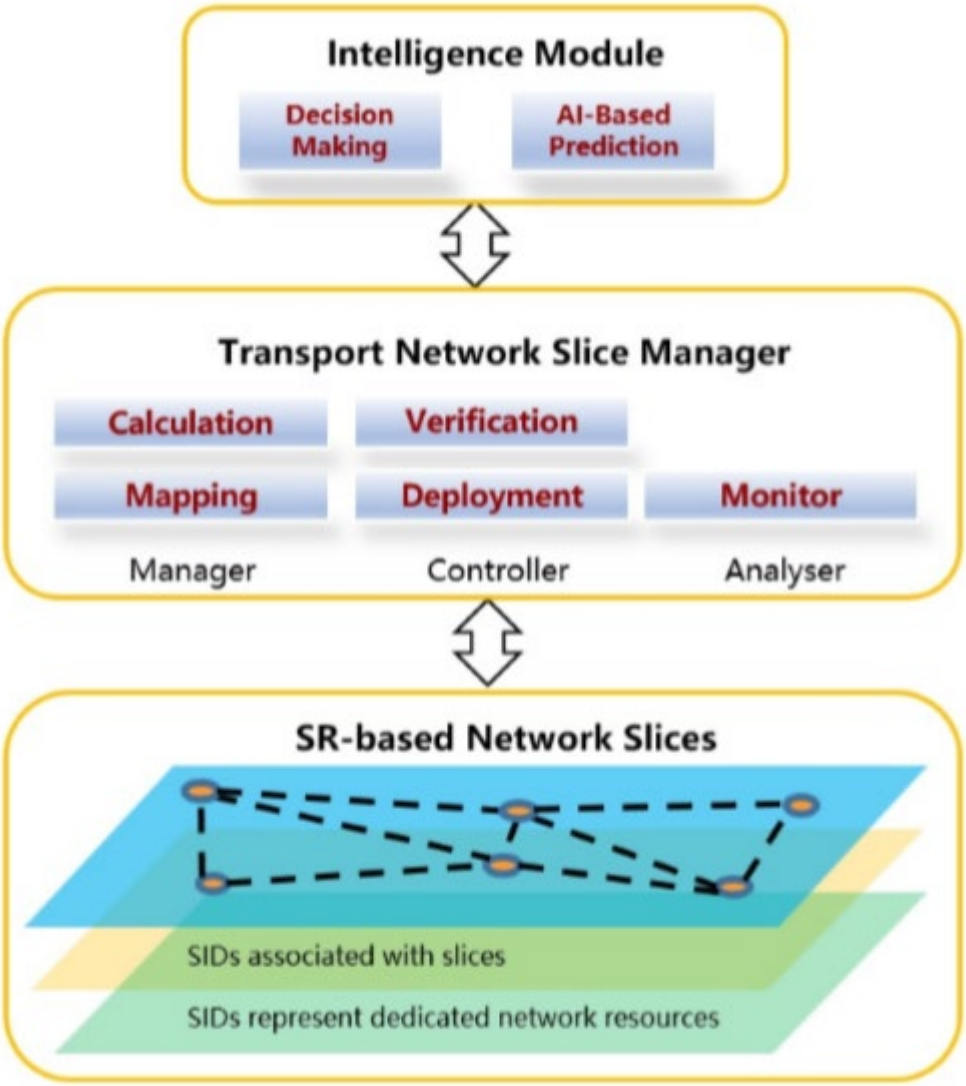



Figure 46: System structure of the TNSM

The following figure shows the functional architecture of the intelligent transport network slicing system. The AI predictor first performs training using historical traffic throughput data. The TNSM collects the real-time traffic throughput data of transport network slice instances and sends the data to the AI predictor. The predictor predicts the traffic throughput values in the next several time segments based on the training model and real-time data and sends the traffic throughput values to the smart policy generator. The smart policy generator determines the scaling and bandwidth adjustment policies for transport network slice instances based on the prediction result and delivers the policies to the TNSM if necessary. Finally, the TNSM implements the corresponding scaling policy by reconfiguring the port bandwidths of the two transport network nodes.

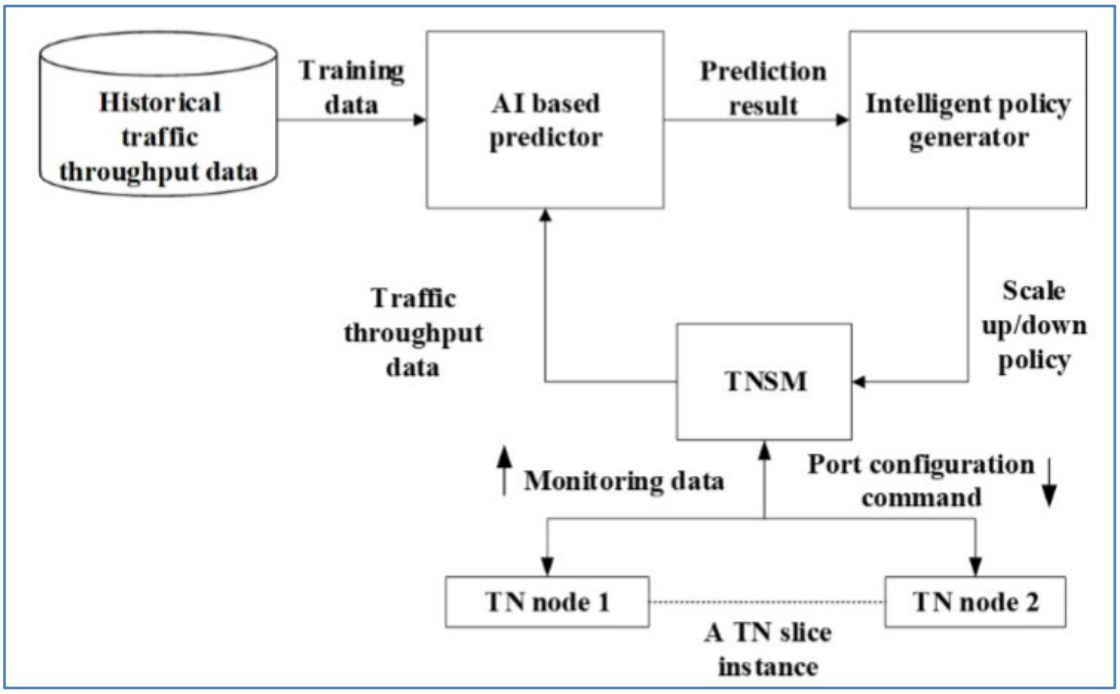



Figure 47: Functional architecture of the intelligent transport network slicing system

In this case, the public data packets of the backbone network in the WIDE project (https: //mawi.wide.ad.jp/mawi/) are used as the data source to capture the data transmitted every day over the transmission link from the WIDE to the upstream ISP. The traffic data changes periodically within a week. Therefore, the training and test are performed weekly in this case. During data processing, if the data is incomplete and is lost for more than three consecutive time points, all data of the week is deleted from the data set. After data cleaning is complete, data is normalised (that is, average value = 0, std = 1), and the data set is classified into the training set, test set, and verification set based on the proportion of 8:1:1.

Experience shows that the autoregressive method of predicting the future value by using the past value of time series has apparent advantages in real-time policy adjustment in various fields. Therefore, in this case, related algorithms and integration models are studied. During algorithm comparison, some of the most commonly used methods are tried, including the ARIMA model, long short-term memory (LSTM) model, gated recurrent unit (GRU) model, temporal convolutional network (TCN) model, and the integration model of these models. Through comparative research, it is found that the model integrating multiple single algorithms can obtain higher accuracy than a single algorithm through the synergistic effect. For a given data set, the hardware platform, in this case, is used. The LSTM, GRU, and TCN models have similar prediction accuracy. Therefore, the GRU-based integration model is used.

The main task of creating a slice is to allocate all requirements on a network by using available bandwidth on a physical interface, and the slice can be successfully created only when all required bandwidth and delay conditions are met. The TNSM uses the greedy algorithm to allocate paths. In the startup phase, the greedy algorithm sorts the requirements based on the delay requirements and then sets the available bandwidth and delay for each link. Starting from the requirement of the shortest delay, the algorithm filters out the links whose available bandwidth is smaller than the required bandwidth and then calculates the shortest path. If the shortest path meets the delay requirement, the algorithm subtracts the bandwidth resources occupied by the link and updates the network capacity. After all requirements are successfully calculated, path aggregation is performed by reducing the delay to maximise the bandwidth usage of the

FlexE subinterface. Finally, a topology formed by the selected node and the FlexE subinterface is used as a result of slice creation.

There are two types of slice adjustment requirements. One is a new requirement, that is, a requirement that does not exist in the previous slice calculation. The other is a legacy requirement, that is, a requirement that has been successfully allocated in last slice creation or adjustment policy execution, and whose bandwidth will be expanded, reduced, or kept unchanged in current policy execution. Most operators are very cautious about any adjustment in the existing services. Therefore, slice adjustment is preferentially performed on the existing paths when there is available bandwidth. Therefore, the slice adjustment algorithm first checks whether all legacy requirement adjustments can be completed on the existing paths. If yes, only the slice creation algorithm needs to be invoked for new requirements. Otherwise, the adjustment algorithm first determines the number of legacy requirements that can be met, and then combines the legacy requirements that cannot be met with the new requirements for which the slice creation algorithm is invoked, and outputs a slice topology that meets all the legacy requirements and new requirements as a result.

In the test phase, the Intel Xeon Gold 6148 dual-slot system with 192 GB DDR4 2666 memory is used in the lab. When the batch size is 64, a single model is used. The predicted delay is 0.08 ms, and the precision is 91.17%. Then, the six models are integrated. The predicted delay of the integrated six models is 0.58 ms. for the integration of multiple models, the precision is slightly improved. The integration of six models achieves the following prediction precision: R square value = 0.9338; precision = 91.75% (The precision is defined as 1 – Average prediction error rate. The higher the precision is, the better the performance is.)

Using AI to enhance and optimise network slice management and control operations is a typical case of ENI. In June 2018, ETSI ISG ENI launched its first Proof of Concept (PoC) project. When the alarm rate is acceptable, the resource efficiency of the test set can be improved by about 30% by using the intelligent policy based on traffic prediction. In the implementation and deployment process of the intelligent network slicing system, an additional resource unit can be added to reduce the alarm rate further.

The AI technology is introduced to better process data and optimise algorithms, helping accurately predict the network service volume and resource requirements. Then, automatic scaling and adjustment of virtual networks and network slices can be implemented based on the prediction results. After the optimised resource allocation policy is used, the network running status can be iterated to the prediction model again to complete closed-loop feedback, maximising resource efficiency. This reduces operation difficulties of virtual networks and network slices while improving resource efficiency.

It is recommended that international standardisation organisations further standardise the intent-driven AI module function, TNSM function, and interfaces between the AI module and the TNSM, including the interface type, interface API, and information model. In addition, it is recommended that the GSMA AI in Network working group establish and maintain contact with other standardisation organisations to further complete relevant standardisation work according to the actual requirements of global operators in terms of network AI, and promote the application of standardisation achievements in more operators networks.

Slicing is a critical technology for 5G networks. In the future, a large amount of 5G data on live networks will be used, and E2E network slice management will be streamlined based on standard specifications, significantly improving network resource efficiency and reducing costs for operators.

## 3.4.2 Intelligent Service Identification

User service type identification is a process of identifying a corresponding type to which TCP or UDP flows that have the same quintuple in a network belongs. The type can have different dimensions depending on the needs. For example, it can be an application layer protocol carried by the TCP flow, such as FTP and P2P, a corresponding app, such as WeChat and QQ, and even a refined action in an app, such as sending a red envelope or sending an image in WeChat. Currently, the DPI system identifies user service types mainly through rule library matching. The maintenance of the service identification rule library requires a large amount of workforce, and the update period is as long as three to six months. The intelligent service identification solution uses deep learning technology to automatically obtain service data, train and release models, and identify online services. This solution reduces workforce costs for rule library maintenance and supports encrypted protocol identification. This solution provides accurate and efficient user service identification capabilities with a wide range and fast iteration, supporting analysis applications of various upper-layer production systems. With the development of AI technologies, the network brain with high reliability proactively plans network paths. The optimisation efficiency of highly dynamic networks is likely to exceed that of traditional network algorithms. In this way, service identification is more complete and accurate, ensuring user experience.

The following figure shows the overall framework of intelligent service identification, which consists of the offline part and an online part. The offline part needs to implement APK management, data management, model training, and model evaluation. The online part inputs the source code stream of the user to be identified and outputs the service type tag. The model training data is packet data captured when an app is used on a mobile phone. Conventional algorithm models include random forests and neural networks. Model training specifically includes algorithm design, parameter adjustment continuously performed on the model with reference to the training data, and multiple rounds of iterative optimisation until a model with high identification accuracy is obtained.

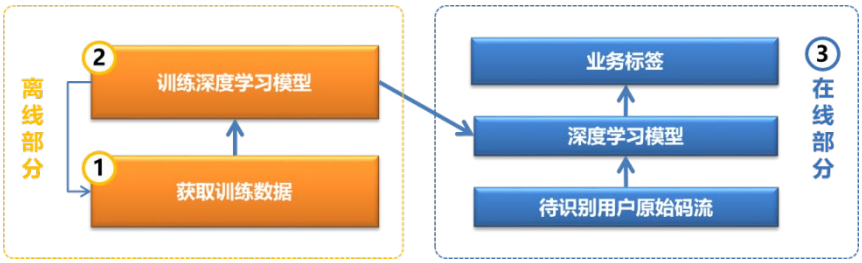



Figure 48: Overall framework of intelligent service identification

Currently, the latest intelligent service identification model has identified more than 1000 apps, and the identification accuracy in the lab reaches 91.4%.

In 2018, intelligent service identification was piloted in the live network environment of a province. The overall accuracy reached 81.3%, and the HTTPS accuracy reached 75.7%. In terms of performance, it is estimated that one GPU server can handle 2 TB/s traffic of the entire province, and the accuracy, efficiency, and stability meet the availability requirements of the live network.

For refined action identification in apps, verification has been completed on some apps in a lab environment, for example, identification of red envelope sending, image sending, and voice call actions in WeChat and training accuracy is 90%.

In terms of standardisation, work with industry partners to develop the service identification architecture and unified interfaces, and promote large-scale application through cooperation or technical authorisation. In terms of technology, the solution needs to be further improved, innovation needs to be made in aspects

such as algorithm optimisation and model migration, and the solution needs to be gradually verified. In terms of scalability and high adaptability, this scenario may be further extended to IoT service identification.

### 3.4.3 Intelligent Service Experience Evaluation

With the evolution of mobile networks and the change of user requirements, the traditional voice and data service KPI-based network quality evaluation system cannot fully reflect real user experience and cannot meet the requirements of the rapid development of vertical services. Operators need to accurately perceive and evaluate the service quality of various services to ensure service quality. Different services have different requirements on network quality, and therefore the methods for evaluating service quality are different. If the real service experience SLA of the service party can be obtained and associated with related network indicators, the correlation between the user experience SLA and network indicators can be analysed based on big data and AI methods to establish customer experience perception models in different service scenarios. Based on the perception models, operators can evaluate the user experience SLA value through network indicators, accurately perceive the real user experience, and then correctly orchestrate and control network resources to ensure service quality.

User experience indicators and associated network indicators vary depending on service scenarios. Third-party content providers (CPs) are service owners who are most concerned about service experience and can accurately understand and provide feedback on service experience. Therefore, you can collect SLA experience data of service from third parties, analyse network indicator data related to the service, and establish a correlation model between the SLA and network indicator data. The service means opinion score (MOS) data of third-party services can be collected to the network through the open interface that is defined in the 5G network and oriented to the third-party application function (AF). The service MOS data can also be collected through the built-in SDK of the third-party service client.

Video experience evaluation is used as an example. Extract experience indicators such as initial buffering and frame freezing in user experience, analyse correlation with the SINR, RSRP, and xDR data on the network, and establish a correlation model between video access experience data with network indicator data by using a regression algorithm model such as a neural network or XGBoost. After the correlation model is established, you can collect network indicator data online to evaluate service experience. The solution consists of offline training and online evaluation, as shown in the following figure.

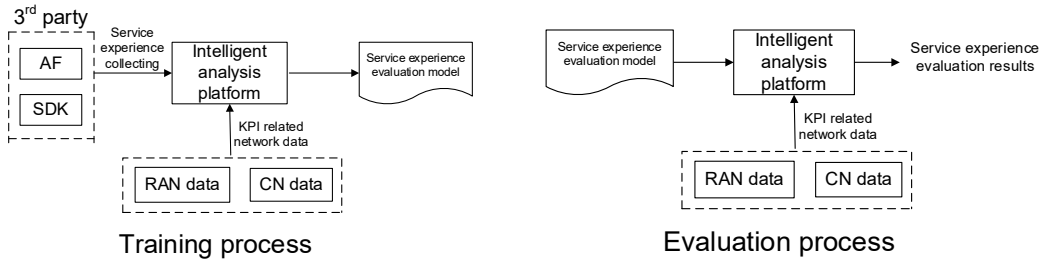



Figure 49: Process of offline training and online evaluation

Because different types of services, such as video, game, payment, and Internet of Vehicles (IoV) services, have different correlation models, modelling and training may be performed for these types of services separately, and migration or generalisation processing may be performed on services of the same type.

This case is piloted in a province. The following is the evaluation result of a video APP initial latency experience indicator based on the live network DPI data, and the evaluation accuracy can reach more than 84%. The model is still in the process of continuous improvement in different scenarios.

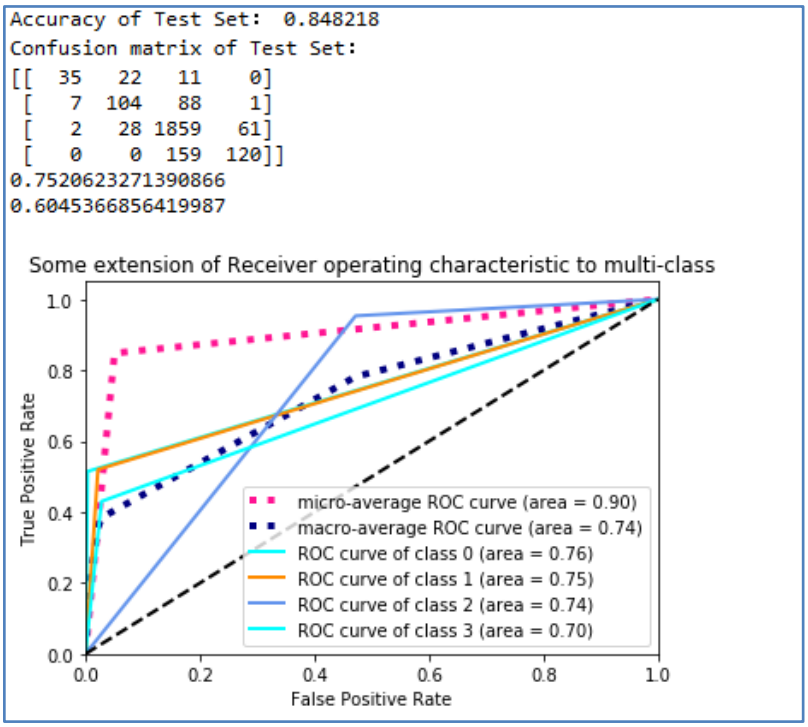



Figure 50: Evaluation result of a video APP initial latency experience indicator

In terms of standardisation, promote the definition of standard collection interfaces for service experience data and network data in typical application types. In terms of technical solutions, work with industry partners to further improve the solution, and cooperate with vertical industry users to promote the implementation of the solution based on the 5G network architecture and the open interfaces oriented to third-party AFs. In terms of solution promotion, extend this pattern to other industry application types.

## 3.4.4 Intelligent MOS Evaluation

With the rapid growth of VoLTE traffic, the voice quality experience score is important for identifying VoLTE quality problems and improving user experience. Currently, the voice quality MOS is analysed using the drive test method and the perceptual objective listening quality analysis (POLQA) algorithm, and the coverage is limited. The DPI system can be used to monitor and analyse user experience on the entire network. However, the MOS evaluated using the no-reference E-model has a large deviation from the actual user experience, and the accuracy is low.

In this case, the machine learning algorithm is used to establish the correlation between VoLTE service user experience and network indicators. This helps accurately evaluate the voice service quality and detect exceptions. The DPI data supports user-level analysis, implementing user-level full coverage evaluation on the entire network.

Intelligent MOS evaluation builds an intelligent evaluation model to accurately evaluate the voice quality MOS based on the VoLTE media-plane XDR data of the DPI system. This enables detection of VoLTE experience exceptions for all users on the entire network. The solution consists of two parts: offline training and online application.

Offline training collects a large amount of dialling test data on the live network, designs and extracts multidimensional RTP packet header characteristics, and obtains the corresponding POLQA MOS as the training tag. POLQA is a voice quality evaluation standard recognised in the industry and is also a common evaluation method used in drive tests. Reference-based acoustic evaluation can well reflect user experience. Model training uses neural network or XGBoost machine learning algorithms to perform regression analysis and obtain the intelligent evaluation model.

The online application generates segment xDRs and constructs RTP characteristics that are the same as those in model training, such as the encoding rate, packet loss rate, delay, jitter, and continuous packet loss, based on the VoLTE media stream data of the DPI system. This part invokes the evaluation model to perform 5-second segment-level MOS evaluation.

During model training, radio network data, such as the RSRP, RSRQ, SINR, RSSI, CQI, downlink throughput at different protocol layers, and handover conditions, can also be considered as characteristics to construct a mapping model between network environment indicators and user experience.

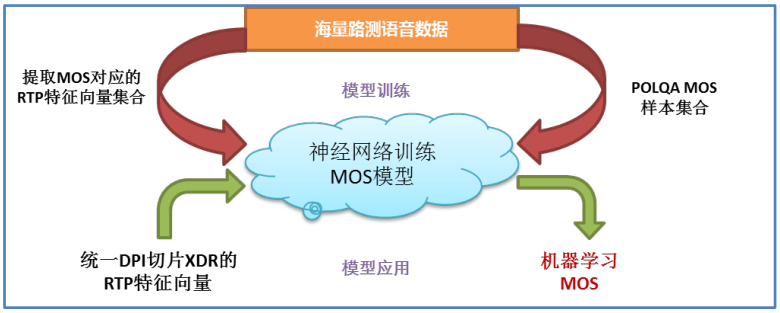



Figure 51: MOS model training by using a neural network

The following figure shows the simulation performance of 8000 sample sets in different RSRP and SINR grouping scenarios.

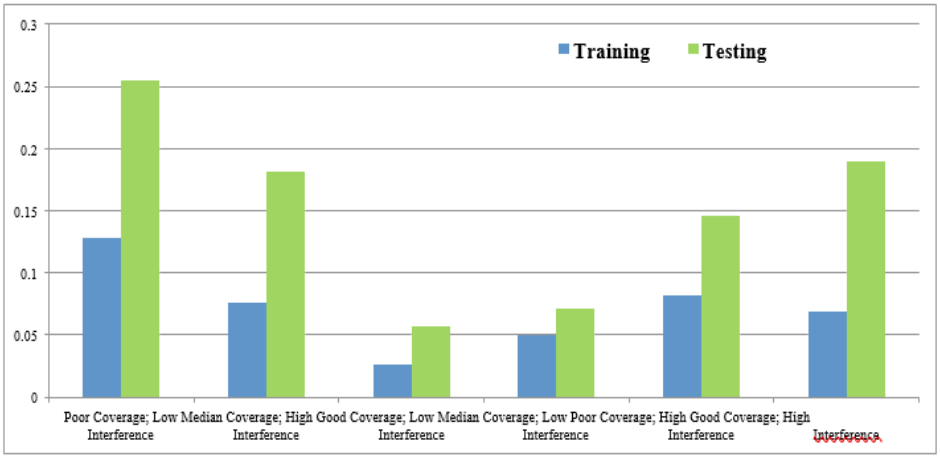



Figure 52: Data simulation performance

The blue bar in each group indicates the mean percentage error between the POLQA MOS in the training phase and the score calculated by the machine. The green bar indicates the error in the test phase. The

two groups in the middle indicate good quality scenarios, where there are a large number of grouping samples on the live network, the performance is good, and the error is less than 10%. The groups at both ends indicate poor quality scenarios, where the number of sample sets is small, the error is large, but the actual effect is within 25%. With the increase in the number of collected sample sets, the error of the groups at both ends may be reduced, and the performance may be further improved.

In 2018, the intelligent VoLTE MOS evaluation model was piloted in a province. The evaluation accuracy was 80%, the mean square error was 0.3, and the accuracy at the medium and poor quality points was 70% and 78% respectively compared with the POLQA MOS evaluation with the error within ±0.25. The evaluation accuracy was 40% higher than that of the ITU E-model evaluation model, greatly improving the efficiency of detecting VoLTE poor quality problems. It is planned to deploy this solution in the entire province to support voice service quality analysis.

Promote solution improvement with industry partners. Data standards must be formulated for the DPI interface on the VoLTE media plane. By defining unified interface standards and encapsulating models into capability APIs, the solution can be promoted for network-wide deployment and application.

The model can be further optimized. Model generalization can be enhanced to ensure that the model is highly accurate even in various poor-quality scenarios and coding schemes.

This solution can also be used to evaluate the quality of multimedia services such as VoIP, streaming media video, VoLTE video, and AR/VR on 4G and 5G networks.

## 3.5 AI for Network Energy Saving and Efficiency Improvement

### 3.5.1 Wireless Network Energy Saving

As operators  network energy consumption keeps increasing, reducing the energy consumption of main equipment is key to energy saving. Reducing the power consumption of main equipment of wireless sites has become the top priority for all. For a typical operator, the power consumption of wireless sites accounts for about 45%, and the power consumption of wireless base stations as main equipment accounts for 50%. In the power consumption of a wireless base station, the power consumption of radiofrequency units (RRUs) accounts for a large proportion, and that of the power amplifiers in the RRUs also accounts for a large proportion. In actual networks, traffic has obvious tidal effect in most cases. When the traffic is light, the base station is still running, which causes a great waste of energy.

Reducing unnecessary power consumption is a key measure of energy saving but is faced with many challenges The network traffic volume varies greatly during peak and off-peak hours. The equipment keeps running, and the power consumption is not dynamically adjusted based on the traffic volume. As a result, a waste of resources is caused. The capability of "zero bits, zero watts" needs to be constructed. However, in a typical network, the features of different scenarios vary greatly. How to automatically identify different scenarios and formulate appropriate energy saving policies becomes the key to energy saving.

- Business district: high requirements on user experience, obvious tidal effect, and light traffic at night
- Residential area: high requirements on capacity, heavy traffic in a whole day, and no obvious traffic fluctuation
- Suburban area: low requirements on capacity, light traffic, sparse sites, and long site coverage distance

In traditional energy saving mode, a large amount of data needs to be manually analyzed, including common parameter data, network inventory data, feature adaptation data, site co-coverage data, and

multi-frequency and multi-RAT network identification data. Therefore, unified shutdown parameters need to be manually set. However, these parameters are not differentiated and cannot automatically match different scenarios to adapt to the traffic volume of a single site. During peak hours, services are affected, and KPIs are affected due to inappropriate parameter settings. During off-peak hours, the power saving effect cannot be maximised due to inappropriate parameter settings.

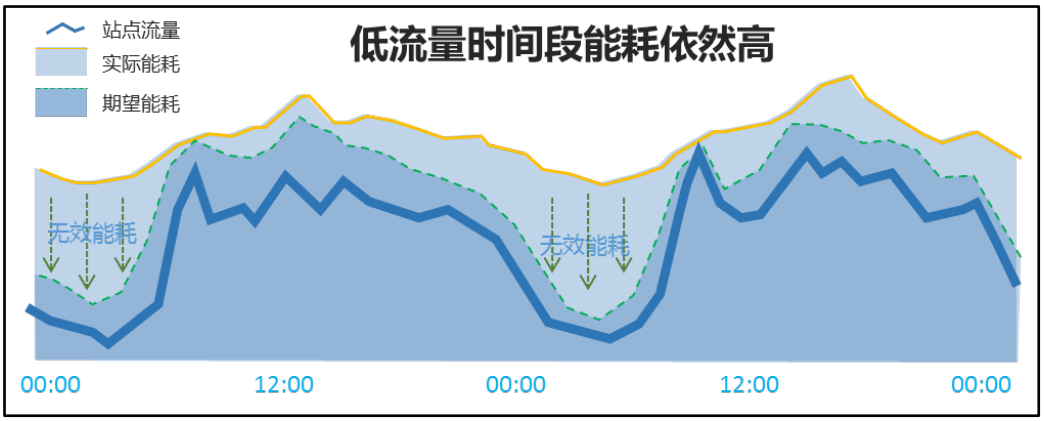



Figure 53: Challenges of the traditional mode

The mobile network energy saving solution uses AI technology to achieve intelligent energy saving in different scenarios, sites, and time. This also enables multi-network collaboration in energy saving. This solution maximises the network energy saving effect and achieves the optimal balance between power consumption and KPIs while ensuring stable KPIs.

The solution consists of four phases: evaluation and design, function verification, energy saving implementation, and effect optimisation.

In the evaluation and design phase, the system automatically sorts out mainstream scenarios on the live network based on big data analysis, analyzes energy saving scenarios based on service models and base station configurations, evaluates energy saving effects in different feature combinations, network environments, and scenarios, and automatically estimates energy saving effects and designs solutions.

During function verification and solution implementation, the network management system automatically monitors and analyses power consumption in all scenarios, provides accurate power consumption reports, and verifies the deployment and effect based on automatic energy saving policies and parameter design. The energy saving policy can be customised for each site, enabling customers to quickly and efficiently start network-wide energy saving.

In the effect optimisation phase, the system automatically adjusts threshold parameters, monitoring items, and power consumption based on the traffic model, energy saving effect, and KPI trend analysis in all scenarios and the AI algorithm. In this way, the energy saving effect and KPIs are balanced.

During the entire process, three key technologies play an essential role: cell co-coverage learning, multi-mode coordination, and AI-based parameter optimisation.

**Cell Co-coverage Learning**

The traditional method of identifying co-coverage cells is to select the capacity and coverage of cell pairs with the same latitude and longitude and azimuth based on standard parameters. This method cannot accurately identify co-coverage cells. The intelligent energy saving system provides the co-coverage learning algorithm. It collects statistics on the inter-frequency support rate of UEs in a capacity cell and

initiates inter-frequency measurement. If both the support rate and measurement success rate of UEs exceed a specified threshold, the co-coverage relationship exists. Otherwise, coverage holes exist. After the learning results are periodically updated, the system automatically establishes intra-site and inter-site frequency band co-coverage relationships. This increases the scenarios where energy saving takes effect by about 20%.

By analysing massive measurement reports and service information, the intelligent energy saving system can detect energy saving cells and compensation cells on the network and predict the service change trend. When an energy saving cell is in the light-load state, the system migrates the services of the energy saving cell to the compensation cell and lets the energy saving cell enter the sleep state. With the real-time monitoring function, the system can wake up the sleeping energy saving cell in time during peak hours to ensure network quality.

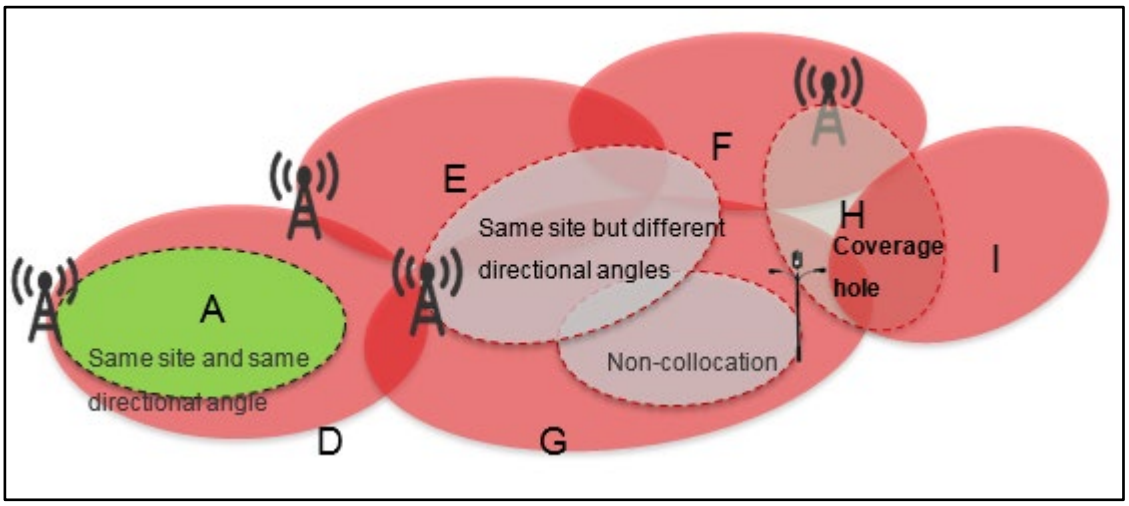



Figure 54: Co-coverage scenario

**Automatic Multi-mode Energy Saving Policy Coordination**

The traditional energy saving mode uses independent energy saving for a single RAT. Symbols for all RATs on a frequency band cannot be shut down simultaneously, and RRUs cannot enter the sleep mode at the same time. Therefore, energy saving gains are insignificant. Energy saving parameters are manually configured, which is inefficient. Energy saving parameters are not differentiated. The intelligent energy saving system uses automatic multi-mode energy saving policy coordination to implement multi-mode and multi-band coordination, intra-band and inter-RAT coordinated shutdown, and inter-band multi-carrier shutdown. The intelligent energy saving system can also implement cell-level automatic differentiated energy saving parameter configuration without manual intervention. This technology saves energy by more than 5% for multi-mode base stations.

**AI-based Carrier Shutdown Threshold Optimization**

For energy saving, a higher threshold for entering the shutdown state indicates a better energy saving effect. However, traditional energy saving solutions cannot automatically identify different scenarios and formulate appropriate energy saving policies because of the diversity of network scenarios and feature differences. Therefore, the threshold for entering energy saving mode is conservative, and the energy saving effect is limited. The AI-based carrier shutdown threshold optimisation technology is used to achieve a tradeoff between the load threshold and performance, maximising the energy saving effect.

The system automatically obtains the optimal shutdown threshold based on traffic prediction and enhanced learning, and performs online iterative optimisation. This increases the shutdown duration by more than 10% while ensuring that KPIs are not affected.

By using the historical data of a large number of cells on the network, such as time, load information, neighbour relationships, and other external factors including weather data and specific events as input, the system performs AI modelling on the cell, cell cluster, or area level. In this way, the system can predict a load of a cell, cell cluster, or area in a coming period, and determines the optimal energy saving time for different energy saving functions (such as carrier shutdown, channel shutdown, and symbol shutdown) in a cell within the range.

In the prediction modelling, the system monitors the network KPIs and provides feedback for prediction modelling according to the changes of the KPIs, achieving iterative prediction modelling and optimal energy saving and system performance.

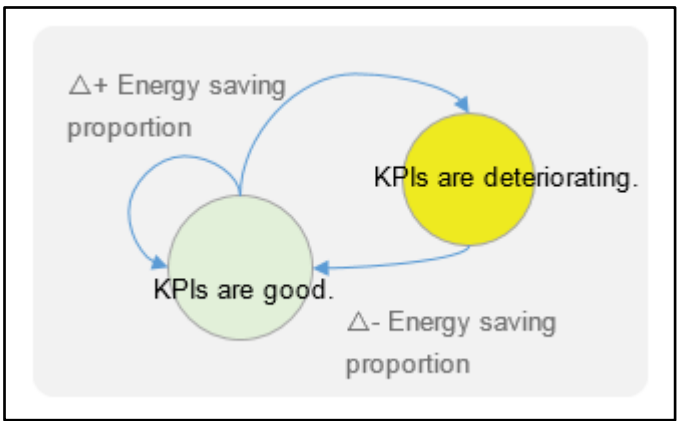


Figure 55: Online iterative optimisation

Algorithm: Load prediction can be implemented based on time sequence algorithms such as ARIMA or regression algorithms such as the random forest and neural network algorithms.

Involved data set: The data includes historical load data, historical performance KPIs, neighbour relationships, and other external information such as weather and event information (such as special party activities).

When AI can accurately predict the load trend in a coming period, if the load is low, the network side pre-determines an energy saving function that can be implemented in the period and estimates energy saving effect in the period. This effectively improves energy saving efficiency in the energy saving period. The AI method can be used accurately predict a sufficient period of an energy saving application. This reduces the impact on performance KPIs caused by inappropriate energy saving period configuration in manual configuration. In critical assurance scenarios, you can set a whitelist to disable the energy saving function in the scenarios to avoid the impact of energy saving.

In typical network configurations, the power consumption of base stations can be reduced by 10%–15%, and the emission of about 2 million kg carbon dioxide can be avoided for every 1000 base stations in one year. Intelligent energy saving has been deployed in 15 provinces and cities in China and a total of more than 500,000 cells. The system can save 400,000 kWh power per 10,000 cells in one year.

Energy saving threshold control can be further refined. AI policies can be used to determine fine-grained thresholds for different energy saving functions, and the energy saving threshold configuration can be customized, improving the energy saving efficiency. The hardware system represented by large-scale

antenna systems and micro base stations (transmit power is less than 250 mW) will bring a vast energy consumption challenge to widespread 5G deployment. Therefore, 5G new scenarios need to be introduced to the existing network-level energy saving algorithm of the system.

## 3.6 AI for Network Security Protection

### 3.6.1 Advanced Threat Defense

Human life and activities have been deeply integrated with the Internet, and major attacks such as APT and ransomware are becoming more and more frequent. The core idea of traditional threat defence is to implement attack behaviour detection based on attack signatures matching. New advanced attacks, such as APT, are good at using 0-day vulnerabilities and new types of malware. This makes it impossible for security devices relying on known features and behaviour patterns to detect advanced attacks such as unknown threats. Also, traditional detection methods are too difficult to cope with the increasing number of anonymous threats.

Research shows that the total number of malware and computer viruses in the world has reached 100 million, and the average number of malware and computer viruses have reached millions daily. Faced with such a vast amount of unknown threats, traditional threat detection devices will be exhausted. Besides, with the development of mobile devices, cloud storage and the IoT, the possible attack surface has grown exponentially. The unknown threat, the massive characteristics and complexity of the data to be analysed make it impossible to be processed with by the manual analysis of security experts only. The objective and effect of this case is to accurately detect unknown threats and provide situational awareness to improve the effectiveness and efficiency of security responses.

In addition, the 5G's high bandwidth, large-scale, and ultra-low latency functions greatly promote the interconnection between everything, including smart home security monitoring systems, vehicles, UAVs, medical devices, and other IoT sensors. However, as highlighted by the findings of the threat intelligence report, many current IoT devices are lagging in security protection, and technology complexity is increasing, which gives cybercriminals a higher chance to successfully launch IoT device attacks. To cope with increasingly sophisticated threats, organisations must integrate all security elements into the security architecture to find and respond threats in the broader range at a faster speed. The system uses machine learning and AI to form future defence strategies, automate analysis and detection of possible advanced threats, reduces MTTD(Mean Time To Detection), and provides quick remedies. Integrating the single-point products deployed on distributed networks will help to deal with increasingly intelligent and automatic attacks.

By processing and analysing traffic, logs, and other information, the system separates and identifies abnormal traffic behaviour and anomalous user behaviour in the network. Based on AI, the system designs a traffic behaviour pattern identification model and uses a dynamic analysis engine to detect malicious codes to catch potential risks (for example, unknown code/files, phishing, likely trojan horse, and abnormal operations). The application of behaviour models can also discover the types of user roles and the associations between attack behaviours, which helps understand attack events. The AI can be used to automatically mine potential features when the sample set is large enough, breaking the restriction of artificially designed detection features. In this way, hidden cyberattacks can be identified.

Compared with traditional traffic detection solution, the AI-powered traffic detection solution has following highlights: by combining multi-dimensional semantic modelling with AI, the solution can discover deeper hidden features and different granularity of data can be used to find abnormal behaviours. User behaviours can be modelled, and unusual access to service resources can be detected.

**Malicious code detection**

By introducing samples into the secure virtual environment, the system can monitor and analyse the dynamic processing of the samples and detect hidden malicious codes. Based on the hardware simulation technology, the dynamic analysis engine can prevent malicious software from detecting the operating environment. It has the characteristics of anti-detection, anti-tampering, anti-penetration, and anti-interference, and can adequately simulate the behaviour of malicious samples. Because it extracts raw data directly from the simulated hardware layer and runs samples in a pure VM machine environment without any modification, it can avoid the detection of the running environment by malware, and is more likely to capture malicious behaviour.

In the case of capturing the complete behaviour of malware, the malicious behaviour determination model is continuously optimised by AI algorithms to improve the detection capability and cope with the escape of malicious code. Because the detection method of the dynamic behaviour analysis engine does not depend on signature, it can support the detection of malware variants and unknown threats and effectively avoid the rapid spread of unknown attacks and the loss of enterprise core information assets. AI algorithms are used to realise the intelligent judgment of malicious behaviour, including using traditional machine learning multi-classification algorithms such as SVM and random forest and large labelled samples to train models and determine malicious categories and converting samples into static image to recognise and classify by using CNN(convolutional neural network). By acquiring a large number of malware sample file and traffic behaviours information, a malicious behaviour judgment model can be trained and established, which can solve the problems of high costs of manual analysis, long time, and rigid judgement rules.

**Account Compromise**

This solution can detect whether hackers have accessed network users' credentials regardless of the attack vector or malware used. It includes the detection of attacks such as passing the hash, passing the token, and brute force attacks. For successful account compromise detection, the technology needs to recognise indicators of compromise on any asset the user touches including endpoints and networks. Potential signs for account damage are as follows:

- Unusual authentication mode (e.g., dormant account access)
- Lateral movement after an attack
- Concurrent login from multiple locations
- Blacklisted account activities

**Insider Threats**

Insider threats include malicious insiders, compromised insiders and negligent insiders and they often result in the destruction of data, risk of data, and so forth. By establishing baseline behaviours for users, the solution can detect and alarm on unusual, high-risk behaviours that fall out of baseline profile based on several factors. Potential indicators of insider threats include:

- Deviated behaviour from those in peer group
- Unusual system access(for e.g. unusual login time)
- Disabled account logins

- Unusual file access and modifications
- Abnormal password related activities
- Excessive authentication failures
- Multiple accounts lockouts

Taking too many identity authentication activities as an example, behind the anomalous authentication activities, attackers try to crack sensitive information such as a users username and password by systematically combining and trying all possibilities. Attackers often use automated script tools to launch brute force attacks. According to the brute force principle, when the system has security risk, it is generally login failure that is significantly increased before a certain login success. At present, the security detection scheme generally adopts a manual threshold setting method, and cannot dynamically model based on different objects. AI can be used to detect whether a user has login intrusion behaviour based on security logs (including the user name, login success or not, accessed source IP address, login failure reason, etc.). If the number of login failures of a source IP address exceeds the set baseline and there is a record of successful login, the brute force attack event is alarmed; if the current login failure number exceeds the set baseline but there is no record of successful login, the suspicious login failure event is alarmed. Therefore, security auditors can be notified of login intrusion behaviour in the system in time, and trace the origin and take corresponding security protection measures.

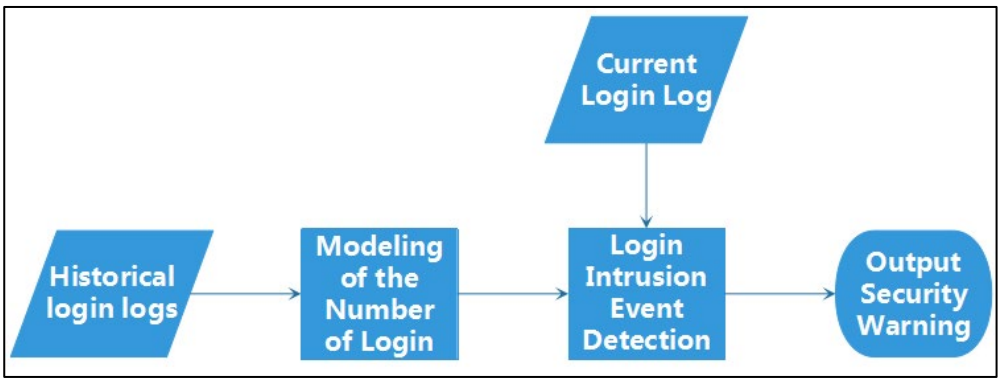



Figure 56: Login Intrusion Detection Flow

Applicable algorithms: Probability distribution estimation algorithms (such as kernel density probability estimation and parameter probability density estimation), periodic detection algorithm and periodic sequence modeling algorithm

Interfaces involved: Internal interfaces (security log reading interfaces and security event output interface)

Data sets: Security logs related to the EMS and NEs

**Privileged account abuse**

The solution can identify specific attacks on privileged users who have access to sensitive information by detecting compromised credentials and lateral movement of systems that contain these privileged data. In addition to privileged accounts, the solution can monitor high-value assets to generate high-priority alarms.

Other risk indicators for an account can also be monitored, such as account lockouts, new account creation, account sharing, and dormant accounts activities.

Potential indicators for privileged account abuse include:

- Suspicious temporary account activities
- Abnormal account management
- Unusual privilege upgrade

**Data Exfiltration**

The solution can monitor in time on indicators that data exfiltration appears to be happening to investigate and stop exfiltration before damage occurs. This is where automated responses can be valuable in lowering mean time to respond and ultimately protecting organisations from data exfiltration.

Potential indicators include:

- Suspicious data transfers
- Abnormal traffic patterns
- Blacklisted user communication

Since 2016, the system has been applied in government agencies (including smart cities), operators, finance, energy, and other industries. Hundreds of major cybersecurity incidents were detected during the application. For example, in 2016, when global ransomware out broke, the latest global ransomware incidents and their variants were first discovered, the early warning ensured information security and avoided the consequences of large-scale attacks. Targeted attacks against enterprise executives were discovered, and regional directional attacks against telecommunication enterprises were found. The internal and external situations were traced and analysed, assisting the emergency response of large-scale WannaCry attack events, and continuing to discover the latest variants of WannaCry. The latest variant of 2019 ransomware GandCrab5.2 was found, which greatly improved the capability of early warning and response to unknown threat and advanced attacks.

- For the Malware detection result, different features are used for detection. The latest comprehensive model is the best one:
    - Normal method, false-positive rate: 8.32%, false negative rate: 17.48%, F1: 0.1912
    - Using AI, false positive report rate: 2.8%, false negative rate: 0.8%, F1: 0.957

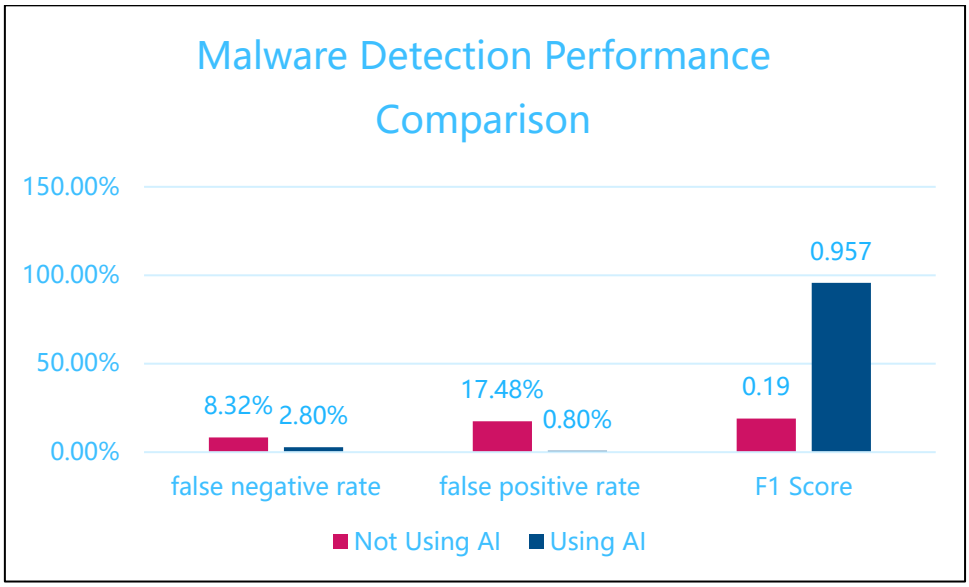



Figure 57: Malware Detection Method Comparison

- For Login intrusion detection

By analysing the login behaviour of 4G network management security logs, the attempt to crack the admin user by NsFocus and other simulation software can be detected accurately, and a large number of login failures caused by login script bugs can be accurately found out. The successful detection rate is not less than 90%.

Standardisation recommendations: Definitions and elaborations of security situational awareness requirements, application scenarios, goals, and functions are required in the standard. The threat recovery response indication is interconnected to the response platform through the API, including a. threat events information to be fixed, raw data and metadata related to threat events. Manual and automatic operation instructions necessary to fix threat events. The message can be carried by security protocols such as HTTPS.

Application suggestions: With the further development of Internet applications, the global network security situation becomes increasingly severe, which brings significant challenges to enterprise operation. Meanwhile, the emergence of cloud computing, 5G, and IoT technologies makes the network boundary fuzzy. The passive protection idea with the security boundary as the core and the deployment of traditional security devices has become more and more limited, and the construction of security protection systems urgently need breakthroughs. The idea of building a security protection system has evolved from passive defence to active and intelligent self-adaptive protection, from simple defence to active countermeasure, and from independent protection to collaborative protection. This means that after basic defence capabilities are established, it is necessary to increase the investment in non-characteristic detection capabilities and focus on building event analysis and response capabilities. Through in-depth analysis of events and sharing of intelligent information, the forecasting and early warning mechanism can be optimised, and the security system is improved accordingly. The target of effective detection and defence against new types of attacks and threats is achieved finally.

- In terms of time, more training data (day, week, or even more extended period) is collected to improve the learning ability of the model, so that abnormal behaviours can be discovered and analysed.
- On the dimension of detection contents, AI-based detection needs to cover three aspects: network traffic, terminal behaviour, and content payload detection.
    - Traffic-based exception analysis mechanism
    - Static and dynamic content analysis mechanism
    - Abnormal analysis mechanism based on terminal behaviour logs

In addition, statistical models and machine learning are used to discover deeper relationships between detection data and behaviour, and correlation analysis is used to identify hidden advanced threats.

To apply advanced threat defence to 5G and the IoT, the corresponding threat intelligent information is required to ensure the effect.

## 3.6.2 Intelligent Junk SMS Analysis and Optimization

SMS messages are cheap and can be sent to a group of users or strangers. Therefore, SMS messages are the most important part of illegal information transmission. Currently, the service systems of provincial and municipal companies and professional companies contain a large amount of illegal and nuisance messages, affecting service running and user experience. The existing SMS message filtering system has the following disadvantages:

- A high false-positive rate of automatic filtering and lack of methods for automatic policy extraction and optimisation
- Lack of intelligent analysis capability of suspicious information

To solve the above problems, this solution uses a self-developed algorithm to identify and filter out illegal SMS messages on the live network of China Mobile. This improves user experience, enhances operator image, and ensures national information security.

The following figure shows the overall architecture of this solution.

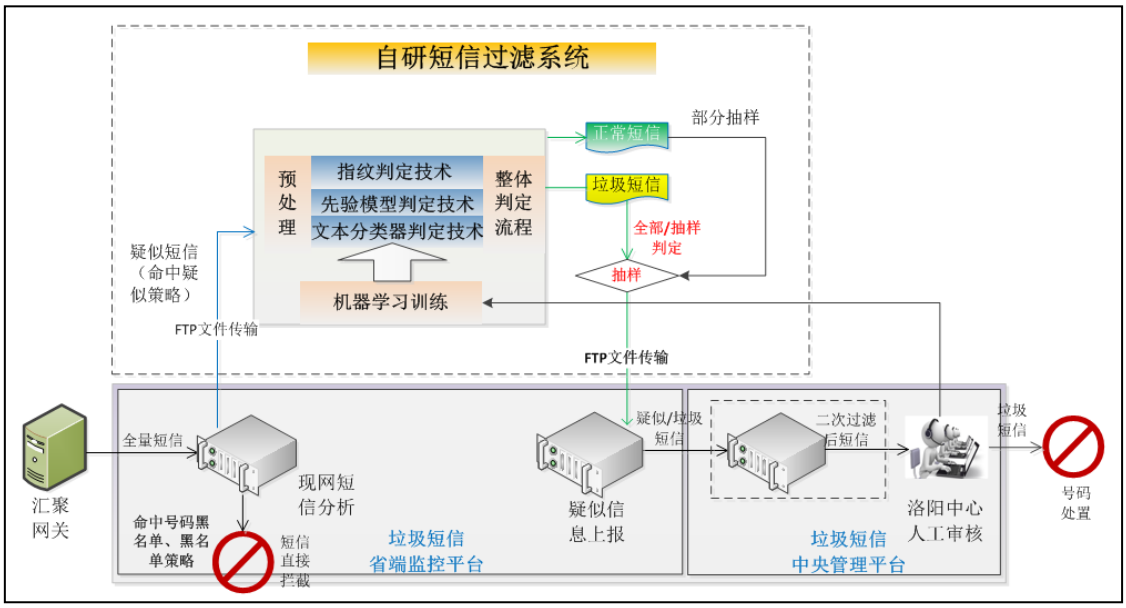



Figure 58: The overall architecture of the intelligent junk SMS message analysis and optimisation system

The overall solution consists of the following four steps:

- The province-level junk SMS message monitoring platform analyses all SMS messages in real-time and sends suspicious SMS messages to the SMS message filtering system through FTP.
- The SMS message filtering system filters out suspicious SMS messages and injects the illegal SMS messages and some sampled normal information back to the province-level junk SMS message monitoring platform.
- The central junk SMS message management platform regularly obtains suspicious SMS messages from the province-level monitoring platform and submits the SMS messages to the Luoyang centre for review.
- After manual review, the data is sent to the training module of the SMS message filtering system for training.

The following figure shows the technical model of the SMS message filtering system.

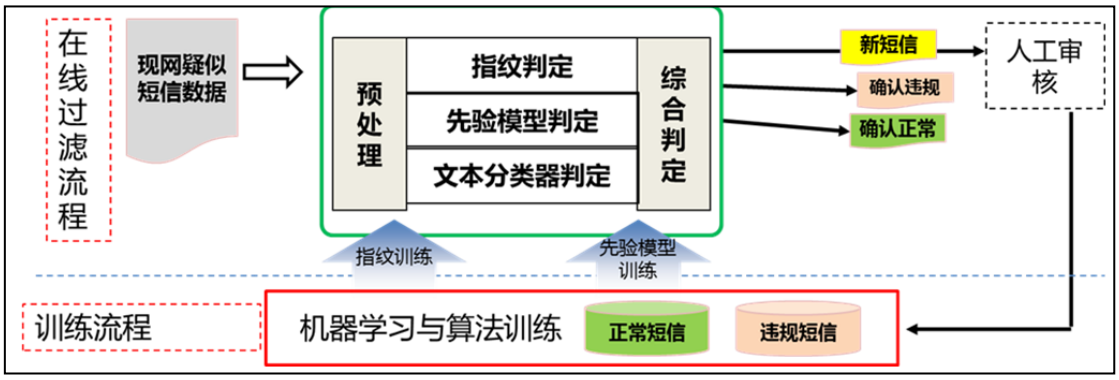



Figure 59: Technical model of the SMS message filtering system

The technical model consists of two layers of intelligent models:

- Training process: Training is performed based on the manually reviewed SMS message data by using the AI model, and training libraries such as the black/white fingerprint library and prior model library are generated, providing a basis for online filtering.
- Online filtering process: Based on the fingerprint model, prior model, and support vector machine (SVM) model, the system accurately determines suspicious junk messages on the live network. The judgment results are as follows: junk, normal, and unknown SMS messages, which can be filtered based on the production system requirements.

This intelligent model has the following advantages:

- Preprocessing technology of short message text

  Based on the summary and study of a large number of manually reviewed SMS messages and user complaint SMS messages, this proposal can propose the character processing methods, such as deleting special characters, replacing special characters, comparing simple and complex characters, processing character disorder, scrambling special characters, and performing DBC/SBC conversion. Exclusive special processing methods can be used to identify the SMS messages that are related to politics, law violation, and fraud and evaded from supervision using various means, such as interference change in prefixes, middle parts, and suffixes, typesetting change, reverse order, and sample truncation. Change the contents that cannot be identified by the machine to common and processable messages.

- Comprehensive SMS message text determination model

  In this solution, a model integrating the fingerprint identification (Simhash) technology, self-developed prior model technology, big data clustering analysis technology, SVM algorithm, and policy mark is proposed, and advantages and disadvantages of the three methods are combined to design and train a comprehensive SMS message determination model, which is used to determine junk SMS messages. The models capability of determining normal and illegal SMS messages is far beyond the basic determination level expected at the beginning of the project. The determination accuracy and recall rate are all industry-leading. In the 10:1 data model, a good determination effect (the suspicious message confirmation probability is 95% for normal SMS message and 65% for illegal SMS messages) is achieved. The following figure shows the overall model.

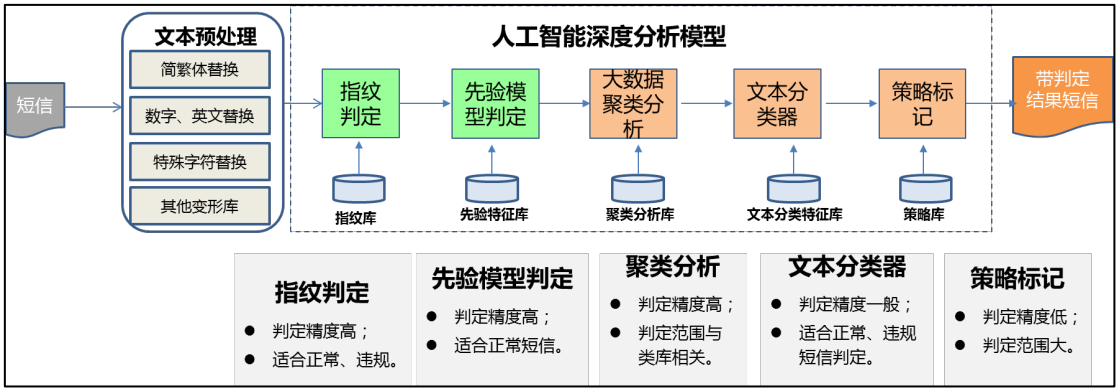



Figure 60: AI-based comprehensive junk SMS message determination model

- Combination of online determination and offline training

  Make full use of existing manually reviewed SMS messages and user complaint SMS messages to construct feature databases, such as the fingerprint library, text feature library, clustering analysis library, template library, and blacklist library, to support the comprehensive determination model. Manually sample and check some determination results, and continuously train and optimize the determination model to avoid a decrease in the model determination accuracy.

This solution was deployed on the cloud MAS platform of China Mobile Government and Enterprise Customers Branch in August 2015. An average number of 5 million SMS messages are detected per day, reducing the error rate from 1‰ to less than 1/1000000.

This solution went live in China Mobile Jiangsu in July 2017. By the end of 2017, 100,000 suspicious illegal SMS messages were monitored, about 20,000 unauthorised SMS messages were determined, and 20 unlawful IoT network adapters were detected on average every day. This solution effectively helps provincial and municipal companies with their O&M services.

Track the latest illegal SMS message samples in time and optimise the algorithm. After the junk SMS message determination model is used for a certain period, the algorithm determination accuracy decreases as more and more violation users master the algorithm determination rules. Therefore, the latest illegal SMS message samples must be trained, and the algorithm model must be updated.

Extend the application of the algorithm. With the massive application of SMS messages in IoT devices and the possible application in 5G networks, the algorithm model can be updated to adapt to new development. For example, in the IoT application, different types of illegal information may be detected based on application scenarios of different devices.

## 3.6.3 Sensitive Data Protection System

Refined analysis based on network and service data not only optimises operator network and service quality, improve user experience, but also enriches mobile Internet services. However, operator network data and service data contain a large amount of user privacy information. Once the information is disclosed or illegally used, the user privacy and personal information security will be severely affected, and even the security of national critical network infrastructure will be seriously affected. Therefore, how to ensure the security of service and user sensitive data becomes a key issue during the promotion of data resource sharing.

Currently, sensitive data identification in the industry mainly depends on regular expressions, dictionary matching, and manual sorting. The former two capabilities are limited by the quantity and quality of regular expressions and dictionaries. Especially when regular expressions and dictionaries are incomplete, or dictionaries are created incorrectly, the accuracy and coverage are low. Manual sorting takes a long time and requires processing personnel to be highly qualified in the case of big data.

To meet the security requirements of big telecom data, this solution provides sensitive data identification and protection service and designs a data security intelligence product system. Based on AI algorithms such as machine learning and neural network, the system implements accurate detection of sensitive data. Based on the analysis of user permissions, this system adaptively recommends algorithms and desensitises data efficiently.

### Refined Data Protection in Dynamic and Static Modes

This system is an algorithm and tool for automatically identifying and desensitising sensitive data based on AI. It implements refined protection of data resources in an open environment (development, test, and data analysis) through three steps: formulation of sensitive data classification and grading principles, automatic detection and extraction of sensitive data, and adaptive recommendation of desensitisation algorithms. The following figure shows the technical architecture of the system.

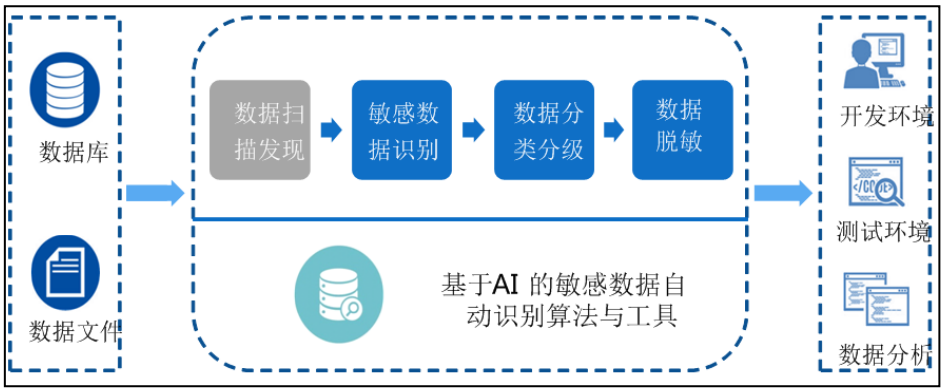



Figure 61: System architecture of AI-based automatic sensitive data identification and desensitisation

### Developing Classification and Grading Principles Based on Industry Standards

To facilitate unified data management, promotion, and application, data is classified into the following types based on the internal data management and external data openness scenarios: user identity-related data, user service content data, user service derived data, and enterprise operation management data. Data is classified into four levels based on data sensitivity and preset leakage loss characteristics: extremely sensitive level, sensitive level, relatively sensitive level, and low sensitive level. Subsequently, classification and grading tags are set based on the rules such as regular expressions, policy sets, and models.

### Extracting Sensitive Data in Dynamic and Static Modes

In a typical production environment, hundreds of TB data in the business (B), operation (O), and management (M) domains are waiting to be processed every day. The data contains millions of sensitive data tables and fields. To quickly and accurately identify sensitive data, different sensitive data discovery capabilities need to be used based on existing data tags.

This project proposes the mechanism of distinguishing between structured data (such as database fields and fixed formats) and unstructured data (such as SMS message content and service content) by using dynamic and static handling modes. The static mode ensures accuracy, and the dynamic mode improves system discovery capabilities. The two modes complement each other. When scanning the database metadata and sample data, the structured data first analyses the static metadata and identifies sensitive data by regular expressions, fuzzy match, and keywords. Unstructured data and content that cannot be identified by some metadata are sampled. Most of such unstructured data is short text data (such as short messages). Short text data has special problems such as sparse data and semantic gap in natural language analysis due to its features such as a small number of words and random expression. This project uses the subsemantic space-sensitive rule mining algorithm and model corresponding to convolutional neural network training, as shown in the following figure.

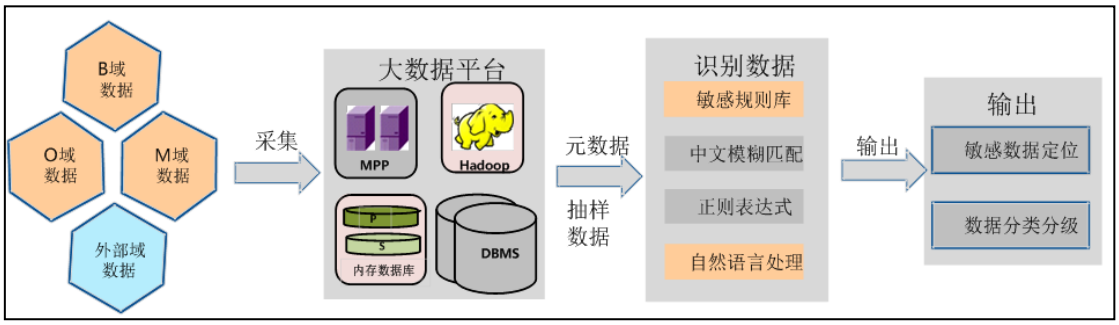


Figure 62: Figure 1 System processing flow

**Adaptive Recommendation of Desensitisation Algorithms**

Based on the desensitisation validity, configurability, consistency, and transparency principles, the desensitisation algorithms are adaptively recommended from the service scenario, service requirement, and applicable party dimensions. The desensitisation algorithms include reversible algorithms such as encryption, format-preserving encryption (FPE), and rearrangement algorithms, and irreversible algorithms such as relationship mapping, offset rounding, hash, and random replacement algorithms. These algorithms retain the original data format and some attributes without degrading the security level of sensitive data and ensure that the desensitised data can still be used for data analysis, mining, and testing, achieving reliable protection of confidential and private data, as shown in the following figure.

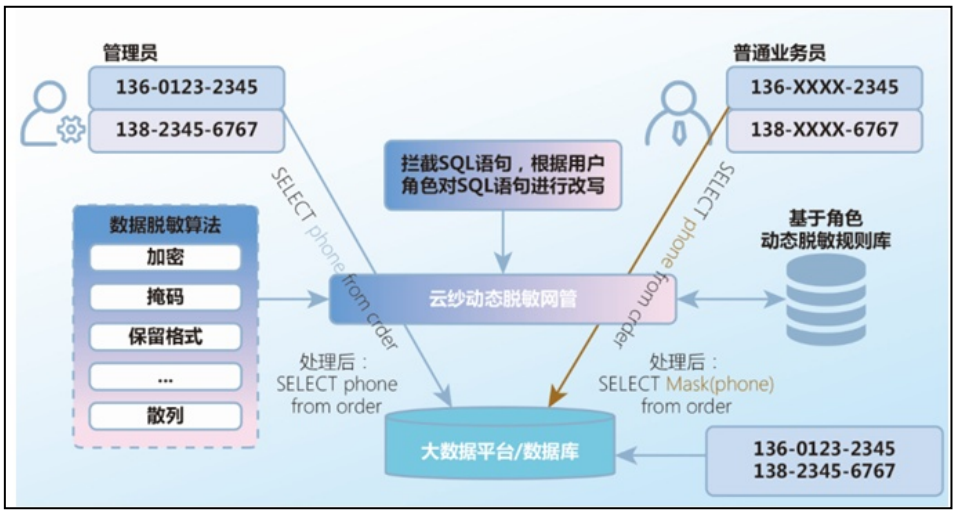


Figure 63: Figure 3 Adaptive desensitisation algorithms

This project uses the sensitive rule library and models to identify sensitive data. The sensitive rule library is constructed by combining regular expressions and sensitive semantic rules, and more than 30,000 rules are accumulated. Eleven identification models are constructed by using the convolutional neural network. The capability of discovering sensitive data is improved from 70% of the initial capacity that relies only on the keyword and regular expressions to 96%. Also, models can be trained for daily incremental data in real-time, achieving high scalability.

Currently, the system has been incorporated into China Mobiles data security products and put into use on the Big Data platform in four provinces and cities. The system provides 24/7, real-time sensitive data monitoring and desensitisation services. More than 800 TB data is processed per day. More than 100,000 sensitive data items have been identified. The sensitive data detection period is shortened from per month

to per day (within 24 hours). This effectively prevents leakage of a large amount of sensitive information, protects service data and user privacy data, and promotes data resource openness and sharing.

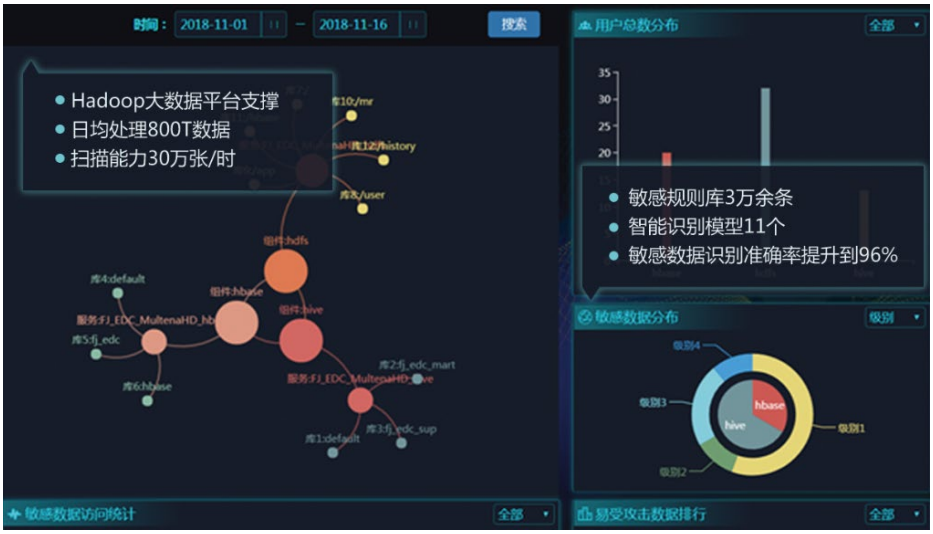



Figure 64: Application effect of the system

The classification and grading principles of telecom data have been promoted as standards in the group and provincial branches. In the future, the system needs to meet the requirements of more data scenarios and deployment environments, expand algorithm capabilities and system functions, and support internal and external requirements of the company. Key technology tackling includes the following three aspects:

- Decouple capabilities from tools. Based on different application scenarios and requirements, sensitive data discovery, sensitive data classification and determination, and desensitisation algorithm recommendation capabilities are decoupled to better serve applications.
- Enhance the intelligent algorithm matching capability. Currently, the algorithm matching capability is still insufficient in complex desensitisation scenarios. The next step is to analyse data association relationships to accurately match algorithm requirements.
- Extend the process streamlining capability for applications. Currently, sensitive data is identified by a single operation. Data security issues that may be caused during multiple operations are not fully considered. The next step is to research and improve data security based on the data life cycle.

## 3.6.4 Botnet Domain Name Detection

With the rapid development of IoT technologies, the number of smart devices increases sharply. Some smart devices have weak security protection measures and are vulnerable to attacks and become controlled devices of botnets. As the number of controlled devices increases, botnets become larger and bring more and more threats. Illegal users can use botnets to initiate various illegal activities, such as DDoS attacks and competition, fraudulent activities, and ticket swiping. These illegal activities affect not only the attacked but also the quality of the entire network. The core of a botnet is a command and control (C&C) server. It is a central computer that sends commands to and receives information from zombie hosts. Botnet malware typically has a built-in set of methods to find C&C servers to keep in touch with C&C servers and reconnect to them after disconnection. However, as IP addresses are used, botnets are easy to detect. Therefore, the C&C main control server usually registers some random domain names in place of IP addresses, to avoid blacklist detection. Domain Generation Algorithm (DGA) is the most

common method for generating C&C domain names by using random characters. In this case, AI technology is applied to botnet detection. By learning the known DGA domain name of a botnet, the unknown C&C server domain name is detected. The domain name of a new C&C server in the network can be recognised actively, which breaks through the limitation that the domain name of the C&C server can be detected only through reverse or blacklist matching and significantly improves the domain name detection efficiency.

**AI-based Botnet Detection Technology**

After the AI technology is introduced and the model is optimised, the accuracy of botnet C&C server domain name detection reaches 99.38%, the false positive rate is 0.28%, and the false-negative rate is 0.95%, achieving accurate detection of botnet C&C servers. The following figure shows the working process of AI-based botnet C&C server detection.

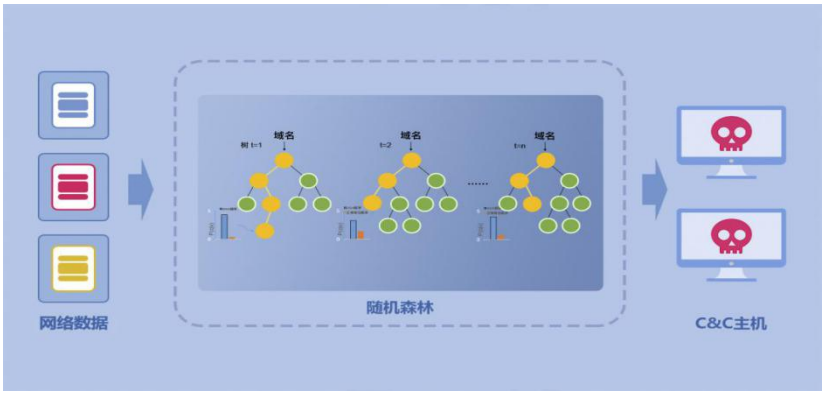



Figure 65: Working process of AI-based botnet C&C server detection

In this case, the characteristics of domain names are extracted, and the supervised machine learning algorithm is used to accurately detect the existing botnet domain names. In addition, the unsupervised learning algorithm (clustering) can be used to discover unknown botnet domain names in advance based on the discovered botnet domain names.

The characteristics of botnet domain names are diversified. The detection methods are continuously upgraded, and the means of evading botnet domain name detection are also changing. To achieve the optimal detection accuracy, the detection model is optimised for multiple times in this case. First, by comparing the machine learning algorithms such as SVM and random forest, the random forest algorithm is selected as the C&C domain name detection algorithm. Second, based on the possible differences between normal domain names and C&C domain names, the domain name entropy, number, letter probability (n-gram), and other characteristics are selected to establish a botnet C&C domain name machine learning detection model. Then, the model is optimised in terms of training data, feature quantity, and model parameters (training data is expanded from initial 1.78 million pieces of training data to 11.41 million pieces of training data, the feature quantity gradually increases from initial 17 to 71, and model parameters are expanded from an initial number of 30 decision trees to 300 decision trees). This ensures the accuracy and recall rate of the AI algorithm. After optimisation, the system detection accuracy is improved from 88.23% to 99.38%, the false positive rate is reduced from 13.51% to 0.28%, and the false-negative rate is reduced from 10.16% to 0.95%. Finally, a pilot project is conducted on the live network. The machine learning model is used to detect network data. Suspicious C&C domain names are discovered, and the number of IP addresses accessing the suspicious C&C domain names and the access time are collected based on the characteristics of botnet activities to confirm the C&C domain

names of botnets. The following figure shows the comparison of feature vector modeling data between some C&C server domain names and normal domain names.

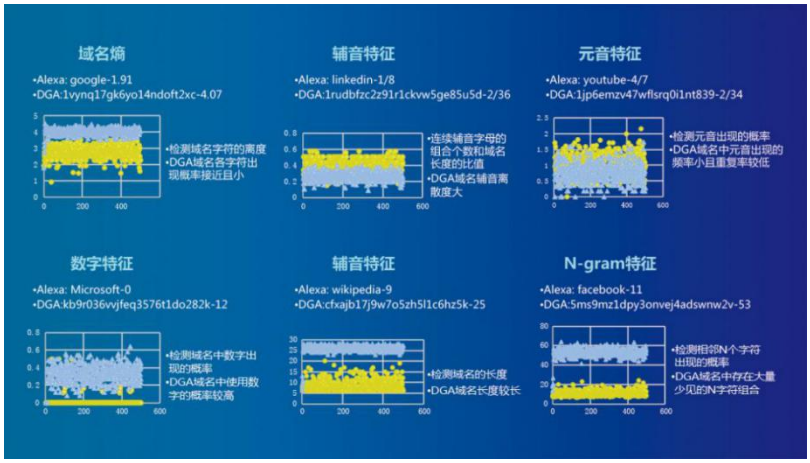



Figure 66: Comparison of feature vector modeling data between C&C server domain names and normal domain names

In the experiment, deep learning neural networks such as CNN, RNN, and Attention Mechanism are used as the classifier model. The detection accuracy is higher than 98%, and the false-positive rate and false-negative rate are lower than 1%.

In terms of prediction, this case uses the confirmed botnet domain name based on the domain name association relationship and implements the prediction of unknown botnet domain names through IP address and domain name clustering. This solves the problem of delayed botnet domain name discovery. In terms of data selection, the inexistent domain names are filtered out based on the number of accessed IP addresses and the activity features of the controlled devices accessing the C&C server on the botnet. When selecting an algorithm, the system performs horizontal comparison on multiple clustering algorithms such as Levenshtein and kmeans, shows vertical comparison on clustering effects of the same algorithm in different similarities, and finally selects Levenshtein in a 0.6 similarity as the clustering algorithm.

**Integrated Detection and Handling**

This case integrates data collection, intelligent analysis and mining, and source tracing and handling to construct a complete defence system for botnets from detection to source tracing and handling. In terms of collaborative processing, the detected C&C server domain names are processed by collaboration with devices such as a gateway. In terms of source tracing, the source tracing module searches for the IP address of the requesting host based on botnet communication data, and confirms and warns the controlled device.

The accuracy of detection based on the AI algorithm is greater than 98%, and the false-positive rate and false-negative rate are lower than 1%. If AI is not used, the accuracy is only 91%, and the false-positive rate is 8%.

This case has been verified and applied with China Mobile Fujian. More than 200 Internet DGA domain names are detected every day. A total of more than 3000 hosts have accessed the C&C server. This issue is handled in a timely manner. Botnet detection and governance have achieved a remarkable effect.

Optimise the detection algorithm. Collect more training data to improve the model learning capability, and improve the model classification capability by adjusting parameters and optimising algorithms.

Extend the applicability of the system. Extend collaborative processing devices. The discovered botnet domain names can be disposed by using a domain name system (DNS), which is a distributed database used on the World Wide Web to create mappings between domain names and IP addresses. In the future, the botnet domain names can be disposed of by means of collaboration with the DNS.

## 3.7 AI for Operational Services

### 3.7.1 Intelligent Customer Service

With the steady growth of telecommunication services, the scale of telecommunication call centres is continuously expanding, and the labour cost is continuously increasing, it is a trend to use automated and intelligent technologies to replace labours. A series of intelligent service scenarios can be found for customer service centres in the telecommunication field, so that the work that used to be done manually can be done by machines, including intelligent customer service, intelligent IVR navigation, and intelligent knowledge base. Intelligent services can release the workforce to a large extent, save labour cost, make enterprise customer service online at 24×7, improve user experience, and implement multi-channel interaction with users. The intelligent knowledgebase can help the enterprise customer service centre reduce the knowledge management cost, improve the knowledge management efficiency, support the full-channel knowledge application, improve the customer service quality, reduce the dependence on training, and standardise the accumulation of enterprise knowledge.

Artificial Intelligence (AI) is also widely used in telecommunication-related vertical industries. For example, by using AI and IoT, big data, and automated control technologies, home devices, home environments, and home consumption can be interacted and controlled in natural languages, creating a natural, comfortable, low-carbon, and convenient personalised home life. With existing network service advantage, telecommunication operators can complement traditional telephone and broadband services by providing end-to-end smart home solutions. In this way, telecommunication enterprises can consolidate existing markets, seize opportunity in emerging markets, and provide high-quality products and services.

Applicable algorithms: Semantic identification of different user statements that have the same meaning, text representation of the users semantics, classification of the problems domain intent, and calculation of the semantic similarity based on the multi-feature integration can be implemented. Multiple rounds of dialogues can be implemented through context-based reference resolution and omission recovery, and thus multiple rounds of dialogues can be implemented. Multi-round dialogue management such as information confirmation, questioning and interaction with the user can be implemented, and user reply

information can be generated. In addition, knowledge is the cornerstone of intelligent customer service, and it is needed to convert multi-source heterogeneous data into available knowledge. Based on knowledge service, advanced knowledge expression methods such as knowledge graph and standardised technical methods can be realised. A knowledge representation model can be established to systematically organised and use knowledge resources.

Interfaces: Interfaces are divided into three types: text representation interface, text classification interface and text generation interface. Text representation interface includes word2vec-based word representation and bert based sentence representation, providing support for various subsequent tasks. Text classification interface includes machine learning algorithms and various deep learning models and supports classification (including intent identification, classification, and text matching) in various scenarios. Text generation interface is based on the seq2seq model, and the text generation function includes text abstract, retelling, machine translation and chat.

Data sets involved: At present, there is no public data about multi-round of dialogues in intelligent customer service aspect in Chinese. The training data used in the existing network is the non-public training data collected through various legal channels.

This solution is used in two branches of China Telecom for an intelligent IVR system. It is applied in the call centre to improve the quality of call service, reduce the working load of the attendant and save expense. It is an important portal for the call centre to implement human-machine interaction. The system has solved more than 70% of user problems since it went online; that is, it has helped the telecommunication call centre reduce 70% of the labour cost.

In the intelligent knowledge base system launched by one branch in China Telecom, the integration of the intelligent knowledge base and the call centre provides agents with real-time intelligent services, thus reducing the cost of training on customer service labours due to the faster update of service knowledge. According to the carrier, the intelligent knowledge base system has improved the working employees efficiency about 75% and greatly saves the cost.

The big video voice assistant of one branch of China Unicom provides powerful voice control and command delivery capabilities, helping users to easily and conveniently implement control. In addition, the AI-based user portrait and content recommendation function achieve a precise match between contents and users. The AI-based voice assistant has doubled the searching efficiency.

The telecommunication related corpus needs to be improved, the data format and interface should be standardized, and the corresponding model usage specification should be formulated (for example, the accuracy and performance requirements based on a certain benchmark corpus). Intelligent customer services are highly expandable, and their service forms need to be designed for 5G, IoT, and other scenarios.

## 3.7.2 Intelligent Complaint Handling

With the development of Internet technologies and rapid expansion of telecom services, operators support systems become more and more complicated, and the scale keeps increasing. The complaint handling process of the traditional 4G network faces three major difficulties: low efficiency, difficult demarcation, and manual operations. The specific problems are as follows:

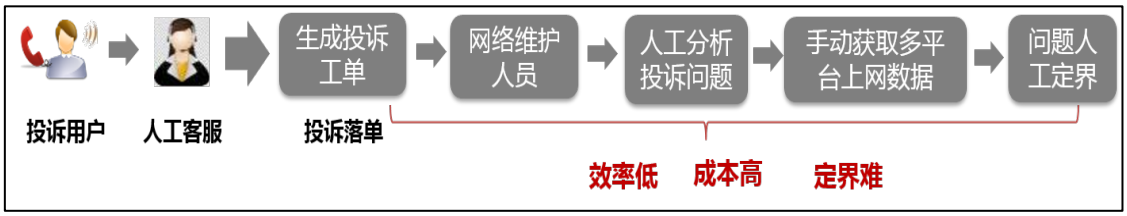



Figure 67: Complaint handling process of the traditional 4G network

- From the time when a complaint ticket is submitted to the time when a fault is rectified, data such as subscription data, radio network data, alarms, and xDR signalling need to be queried on multiple platforms, and an appointment for onsite tests needs to be made. On average, it takes 1–2 days to handle a fault, and more than half of the network operation personnel in the provincial branch are involved.
- Problem demarcation involves network optimisation, monitoring, performance, and professional knowledge in multiple domains and therefore faces many difficulties.
- The entire complaint handling process is completed manually, and no end-to-end (E2E) intelligent demarcation method integrating multiple professional domains is available.

Such partial, extensive, and fragmented O&M management mode cannot meet the actual requirements of operators. It is urgent to introduce AI into the entire complaint handling process to achieve the goal of using AI in place of humans, reducing cost, and improving efficiency.

Intelligent complaint handling (network self-service robot) introduces AI technologies into the E2E process of communications network operation and aggregates powerful service data and network data of operators for multi-source and multi-dimensional comprehensive analysis. The solution integrates multiple AI technologies, such as voice recognition, natural language processing, knowledge graph, deep learning, and intelligent reasoning, and provides E2E self-service, including user intention perception, automatic network data association, network fault location, and fault solution. The solution implements one-click intelligent processing of problems such as complaints, in place of traditional manual operations, improving complaint handling efficiency and user experience.

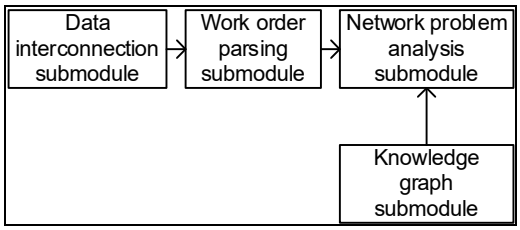



Figure 68: Relationship between intelligent complaint handling submodules

The data interconnection submodule is used for interconnection of work orders and user and device data. The work order parsing submodule is used for complaint work order processing and complaint voice analysis. The network problem analysis submodule is mainly used for network fault location, intelligent resource scheduling, and user intention mining. The knowledge graph submodule is mainly used to construct the knowledge graph in the network domain. AI is introduced into the following procedures: network fault location based on the AI model, network domain knowledge graph construction based on the knowledge graph technology, complaint information processing based on the natural language understanding technology, complaint voice analysis based on the voice recognition

technology, and user intention mining and intelligent resource scheduling based on intelligent reasoning and deep learning. These are the core innovation points.

Involved interfaces: data transmission interface and result feedback interface

Involved data: complaint work order data, xDR data, performance data, alarm data, and network optimisation data

Intelligent complaint handling (network self-service robot) has been put into commercial use on the entire network of a provincial branch of China Mobile. The accuracy of locating and demarcating complaints reaches 70%, and the complaint locating duration is shortened from 1–2 days to 15 minutes. This significantly improves the complaint handling efficiency and user experience. The following figure shows the locating duration distribution of more than 200 work orders located using the network self-service robot.

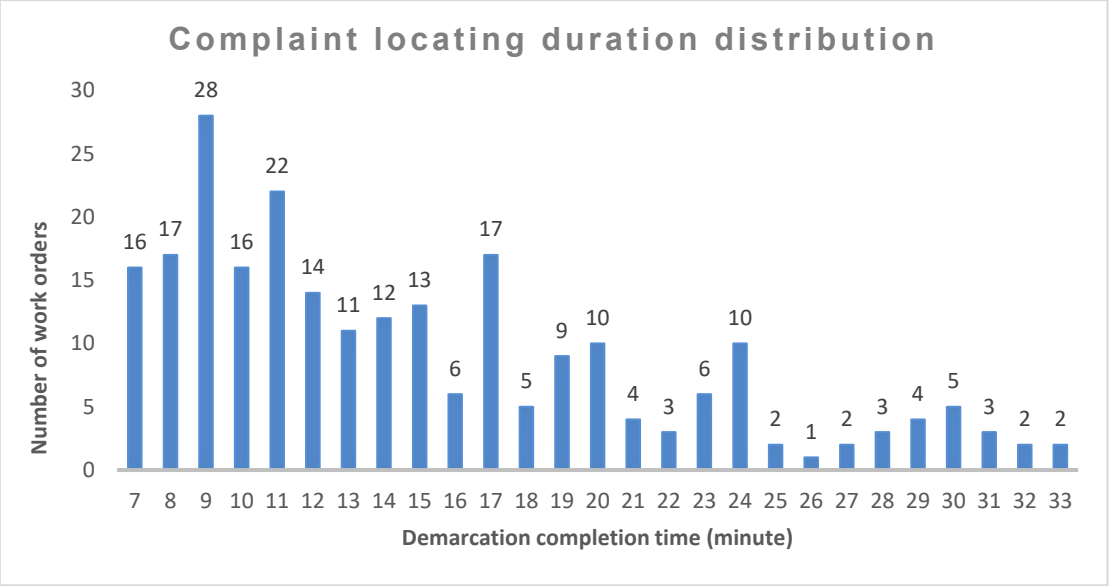


Figure 69: Complaint locating duration of the network self-service robot

Improve the capability of locating and demarcating complaint problems as the general network intelligence capability and continuously optimise the capability, expand the application scope, and continuously improve the industry service level. At the same time, promote the preceding technical solution to 5G network construction and operation.

## 3.7.3 Batch Complaint Warning

Once a regional network problem occurs in the communications network, the Internet access and call experience of a large number of users are affected, causing a large number of user complaints. When the number of complaints in a certain area or about a certain service exceeds a certain threshold, the customer service centre determines this type of complaints as a top complaint issue. Such an issue affects a large number of users in a large scope. Therefore, the processing efficiency needs to be improved, and the impact scope needs to be reduced as much as possible. In practice, the network operation department usually uses expert experience to analyse NE counters, cell counters, and signalling data of a single complaint user to locate network problems after receiving top complaint issues. However, this method belongs to post-processing based on complaint information and is of low efficiency, and top issues cannot be identified in a timely manner.

In this case, the AI algorithm is used to detect the affected user groups and areas in real time. This method is efficient and accurate and does not require much expert experience. This method overcomes the disadvantages of the existing method, and can identify potential batch complaint users and complaint areas promptly and generate warnings so that proactive O&M and customer care are performed before user complaints, improving user experience. In addition, related achievements can be embedded in the complaint preprocessing phase to improve complaint handling efficiency.

The overall framework of the batch complaint user warning model consists of five parts: data selection, data preprocessing, feature extraction, model building, model training, and optimization. The following figure shows the model construction process.

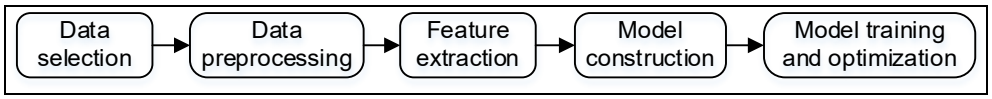



Figure 70: Process of constructing a batch complaint user warning model

- Data selection: User experience is affected to some extent when top issues occur on a network. The impact is reflected in the signalling data, and service data flows of users in real-time. Therefore, select data that can reflect user experience, including user xDR data and user-plane data.
- Data preprocessing: The data contains invalid, incorrect, duplicate, and blank values. Therefore, you need to clean the data based on the network service logic and professional knowledge of the user service process.
- Feature extraction: Raw data reflects different degrees of user perception status. To accurately describe the user perception status, the features of raw data need to be extracted concerning factors such as the type and meaning of data fields. The purpose is to find out the features of invariance in the same type of samples, distinguishability in different samples, and strong robustness, such as statistical features and differential features.
- Model construction: Due to the particularity of network problems and signalling data, classification model selection faces the following challenges: First, to help professionals with subsequent related analysis, the classification model needs to be interpretable. Second, the raw data contains a certain amount of noise. The classification model needs to have strong generalisation ability. Third, the positive and negative samples are not balanced, and the classification model needs to be robust. Considering the other challenges, an appropriate algorithm model is selected to construct the batch complaint user warning model.
- Model training and optimisation: Use the training set to train the model. The accuracy of the output result is used for model evaluation. Optimise the model by adjusting model parameters until the model accuracy meets the requirements.

This solution has been deployed on the live network of a province in China. The batch complaint warning model is introduced to the complaint preprocessing phase to preprocess complaints. The prediction time can be shortened to less than 5 minutes. Compared with the traditional method, this method can provide early warnings about potential complaint users, improve the network service capability, simplify the work process, implement proactive O&M before complaints, and continuously improve user satisfaction.

Continuously supplement data samples, update and optimise models, develop unified interfaces, and promote large-scale application.

# 4 Summary

The mission of technology is to serve human beings. As one of the most important technologies in human history, mobile communications connects huge amount of people, makes the overall productivity of human beings possible to increase by squares, and greatly enriches peoples lives. However, due to the dramatic increase in the number of connections, the explosive growth of diversified services, and extreme requirements for network performance and reliability, the complexity of the network will far exceed that of the current network, and the mobile communications network is facing unprecedented challenges.

In recent years, AI technologies have some tremendous breakthroughs. With the advancing of algorithms and computing capabilities, great achievements have been made in all fields that can provide a large amount of training data. AlphaGo and Alpha Go Zero can easily beat human Go players. The accuracy of voice recognition and face recognition can reach 98% or higher, which is also higher than the accuracy of the ears and eyes of ordinary people. AI technologies have entered many aspects of peoples lives and various industries including communications. AI, 5G, and IoT are the three essential elements of GSMAs Intelligent Connectivity vision. The convergence of AI and mobile communications networks will inject new technological vitality into communications networks and unleash unprecedented possibilities, helping to achieve this vision.

AI in Network has become the key to success. The AI-enabled intelligent autonomous network is an important development trend of 5G and post-5G networks and will bring fundamental changes to networks. The network management mode will gradually change from the current human-driven passive mode to the network self-driven autonomous mode. In the future, intelligent networks will provide more intelligent and flexible network policies based on network data, service data, and user data perception and AI-based intelligent analysis. In this way, networks will function autonomously, significantly improving the lifecycle efficiency of mobile networks and reducing the E2E operation costs.

However, the path towards fully intelligent autonomous network will be a long-term process of gradual evolution. To achieve this goal, the entire industry needs to have a unified understanding of the intelligent autonomous network and its development path. Based on the concept of "hierarchical autonomy with vertical coordination", the industry needs to continuously clarify the connotation and extension of the intelligent autonomous network through various cases. In addition, all parties in the industry need to pay attention to the key problems found in the practice of these cases, such as insufficient training data, incomplete standardised interfaces, non-optimized algorithms, and poor model portability. All parties should work together to solve these problems and promote the development of intelligent autonomous networks.

This report collects innovative use cases of AI in network planning and construction, maintenance and monitoring, configuration optimization, service quality assurance and improvement, energy saving and efficiency improvement, security protection, and operational services. It is expected that this report shares experience, promotes cooperation, and encourages industry partners to gradually clarify the objectives, architecture, and development phases of intelligent autonomous networks and focus on technical difficulties. Although these use cases are still in their initial stage, they have achieved good results. These innovation use cases are growing and interconnected. Experience and knowledge in different fields are being effectively aggregated and integrated. The picture of intelligent autonomous networks is emerging step by step. More and more opportunities for win-win cooperation are being made available.

It is believed that AI and mobile communications technologies, the two that have changed the world, will reshape the future of mankind together.