

Journal of Electronic Research and Application

Editor-in-Chief

Joselito A. Dolot

Lyceum of the Philippines University-Batangas, Philippines

BIO-BYWORD SCIENTIFIC PUBLISHING PTY LTD

(619 649 400)

Level 10

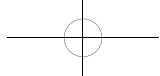
50 Clarence Street

SYDNEY NSW 2000

Copyright © 2026. Bio-Byword Scientific Publishing Pty Ltd.

Complimentary Copy





ISSN (ONLINE): 2208-3510

ISSN (PRINT): 2208-3502



Journal of Electronic Research and Application

Focus and Scope

Journal of Electronic Research and Application is an international, peer-reviewed and open access journal which publishes original articles, reviews, short communications, case studies and letters in the field of electronic research and application.

Topics covered but not limited to:

- Automation
- Circuit Analysis and Application
- Electric and Electronic Measurement Systems
- Electrical Engineering
- Electronic Materials
- Electronics and Communications Engineering
- Power Systems and Power Electronics
- Signal Processing
- Telecommunications Engineering
- Wireless and Mobile Communication

About Publisher

Bio-Byword Scientific Publishing is a fast-growing, peer-reviewed and open access journal publisher, which is located in Sydney, Australia. As a dependable and credible corporation, it promotes and serves a broad range of subject areas for the benefit of humanity. By informing and educating a global community of scholars, practitioners, researchers and students, it endeavors to be the world's leading independent academic and professional publisher. To realize it, it keeps creative and innovative to meet the range of the authors' needs and publish the best of their work.

By cooperating with University of Sydney, University of New South Wales and other world-famous universities, Bio-Byword Scientific Publishing has established a huge publishing system based on hundreds of academic programs, and with a variety of journals in the subjects of medicine, construction, education and electronics.

Publisher Headquarter

BIO-BYWORD SCIENTIFIC PUBLISHING PTY LTD

Level 10

50 Clarence Street

Sydney NSW 2000

Website: www.bbwpublisher.com

Email: info@bbwpublisher.com

Table of Contents

- 1 Performance Optimization Strategies for Porous Carbon Supercapacitors**
Fei Ren
- 7 Exploration of the Application Value of Flexible Carbon Nanotube Electromyography Electrodes in Health Monitoring**
Li Song, Jingjing Sun
- 15 LSTM-Based Operating-State Forecasting and Risk Warning for Small-Scale Battery Energy Storage Systems Using Multi-Source Aging Data**
Qian Wang, Xing Wan, Lihua Luo
- 22 Electrochemical Oxidation Dispersion of Graphene and Improvement of Its Compatibility with Coating Matrices**
Qian Zhang, Kaijie Yuan, Manjiao Deng, Hailin Wu
- 34 Prediction of Grain Yield in Anhui Province Based on Multi-Source Remote Sensing Data Introduction**
Mengxin Ji, Haifeng Jiang
- 46 Analysis of the Problems Existing in Electronic Product Testing and Corresponding Countermeasures**
Haibo Liang
- 52 Research on Medium- and Long-Term Wind Power Forecasting Technology Based on Ensemble Forecasting**
Li Liu, Xiaolin Xu, Qian Liang, Jiayu Chen, Sen Duan
- 64 Research on Equipment for the Preparation of PAN Photocatalytic Nanofiber Membranes**
Zhifeng Yang

- 71 Research on Teaching Resource Generation for the Theory-Practice Integrated Curriculum in Higher Vocational Education from the Perspective of Human-Machine Collaboration**
Liming Feng, Ning Pan
- 80 Human-AI Collaborative Decision Model for Low-Voltage Diagnosis and Closed-Loop Mitigation in Distribution Transformer Areas**
Biwei Li, Caiyu Zhang, Xianwen Zhang, Yexuan Liang
- 94 Research on the Application Path of Dual-Mode API Data Market for Enabling Cross-Domain Multi-Scenario Data Circulation**
Mutong Chen, Xiong Ni, Chuqiao Xiao
- 102 Research on a Trustworthy Collaborative Computing Mechanism for Cross-Domain Sensitive Data Based on Secure Multi-party Computation**
Hao Yang, Mutong Chen, Xiong Ni
- 110 Current Status and Development Trends of Biomedical Big Data**
Juan Zhang
- 116 Research on Plateau Cattle and Sheep Detection Based on QHCSM-GAN Image Completion and Improved YOLO-OC Network**
Xin Zhang
- 132 Research and Practice on Construction Pathways of Emerging Engineering Programs for the Intelligent Manufacturing Industry**
Miaomiao Xin, Wanzhen Wang, Na Li, Ting Wang, Man Qiu, Xichao Yan, Dandan Zhang, Yuanyi Liu
- 141 A Review of Path Planning Algorithms for Mobile Robots**
Xinrui Li
- 161 Discussion on Challenges and Development of Electronic Warfare under Distributed Operations**
Bing Luo, Tianwei Zhou, Yuanming Zhou, Guofeng Shao

Performance Optimization Strategies for Porous Carbon Supercapacitors

Fei Ren

Hongde College, Beijing University of Chemical Technology, Beijing 102202, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Porous carbon materials are carbon-based materials with well-developed pore structures and high specific surface areas. This unique structure endows porous carbon with excellent physical properties, adsorption capacity, and catalytic activity, making it a popular material for supercapacitor electrodes. However, low energy density has become a major bottleneck that limits its development. This article reviews the optimization methods from the two perspectives of the template method and the heteroatom doping method by sorting out its energy storage mechanism. The results show that the hard template method can accurately regulate the pore size, but the process of removing the template involves strong corrosive reagents; The soft template method does not need to use acid/base etching, but it is difficult to regulate the pore size accurately; The salt template method can remove the template by washing, which is the most environmentally friendly strategy among the three; Heteroatom doping is mainly represented by nitrogen doping, which can effectively improve the electrochemical properties of porous carbon materials, but the underlying mechanism still needs to be studied. Finally, the possible development prospects of porous carbon supercapacitors are discussed at the end.

Keywords: Porous carbon; Supercapacitors; Template method; Heteroatom doping

Online publication: Sept 3, 2026

1. Introduction

A supercapacitor is a new type of energy storage device with the characteristics of high power density and long cycle life, so it has attracted much attention in energy storage. However, compared with commercial lithium-ion batteries, their energy density is lower and their endurance is poor, which seriously limits the development of supercapacitors ^[1].

In order to overcome these limitations, researchers tried to introduce porous materials into the electrode materials of supercapacitors, and porous carbon has attracted much attention because of its low cost, simple preparation process, and excellent stability ^[1]. The researchers used the hard template method to prepare and successfully achieved the regulation of the pore size, but the preparation process involves the use of strong corrosive acids or strong alkalis to remove the template, which does not meet the requirements of green chemistry ^[1]. The emerging salt-template method mainly uses water for cleaning, which is cheaper and more

environmentally friendly ^[2]. Although heteroatom doping can effectively improve the conductivity of carbon materials, thereby improving the energy density, the underlying mechanism has not been fully exposed and hinders its practical application ^[3].

This article will start with the energy storage mechanism of supercapacitors, review the two processes of the template method and heteroatom doping to optimize porous carbon materials, and compare and analyze the advantages and disadvantages of different strategies in terms of pore size control, green preparation, performance improvement, etc., which can provide literature support and theoretical basis for subsequent researchers to understand this field quickly.

2. Energy storage mechanisms of porous carbon supercapacitors

According to the different energy storage mechanisms, the current supercapacitors can be roughly divided into three categories: Using electric double-layer capacitors, Faradaic pseudocapacitors that rely on surface redox reactions, and hybrid capacitors that combine the two mechanisms. The porous carbon supercapacitor mainly uses the first principle: placing two non-reactive electrode plates in an electrolyte solution, and then applying voltage at both ends. At this time, under the drive of the external voltage, the positive electrode plate adsorbs anions, the negative electrode plate adsorbs cations, and the potential difference of the two polar plates realizes the storage of energy. Because it only relies on physical adsorption to store energy, this kind of capacitor has a long cycle life. The energy storage is directly proportional to the capacitance of the capacitor. The energy density can be improved by increasing the capacitance. Carbon porous materials have a large specific surface area due to their developed pore structure, so they can adsorb more ions under the same conditions and have greater capacitance than conventional materials. Nevertheless, porous carbon supercapacitors are far inferior to today's commercial lithium-ion batteries in terms of energy density. In order to further reduce this gap, researchers have given multi-angle optimization strategies around the preparation of porous materials: First, use the template method to regulate the pore diameter structure of porous carbon electrodes; second, use the chemical doping of porous carbon electrodes by introducing nitrogen, oxygen, sulfur, and other heteroatoms.

3. Applications of template methods

When the pore diameter of the carbon material is slightly larger than the diameter of the desolvated ion, the ion can undergo desolvation, so as to enter the hole in a denser way, make full use of the huge specific surface area, and improve the capacitance. The idea of the template method is to design the structure of the template in advance, reversely synthesize the material according to its configuration, and then control the shape and size of the finished product. For porous carbon materials, the template method can be used to regulate their size effectively. According to the template category, it can be divided into three types: hard template method, soft template method, and salt template method that has emerged in recent years.

3.1. Hard-template method

The hard template method mainly uses ZnO, MgO, SiO₂, CaCO₃, and other rigid inorganic materials as templates, fills the carbon precursor into its voids, calcines and carbonizes at high temperature, removes the template by acid immersion or alkali immersion, and finally obtains carbon porous materials with inverse

replica structure ^[4].

The traditional hard template method often includes these four steps, and the steps are more complicated and tedious. In response to this problem, the team of Zhejiang Sci-Tech University proposed a more efficient improvement method - mixing tetraethyl orthosilicate as a silicon precursor, resorcinol and formaldehyde as carbon precursors, and mixing with catalysts to complete the template synthesis and material coating in one step, and finally obtain mesoporous hollow carbon nanospheres through hydrofluoric acid etching ^[5]. After experimental verification, the carbon nanosphere prepared by this method has a specific capacitance of 240F/g at a current density of 1A/g, and the energy density is significantly improved compared with ordinary carbon porous materials; The cyclic stability is also extremely excellent - after 10,000 charges and discharges, the specific capacitance retention is still as high as 93.75%, reflecting the good application potential.

The team's method greatly simplifies the complex and cumbersome preparation process of porous carbon materials. At the same time, the preparation of templates and material coating is carried out. The excellent performance of the preparation of carbon nanospheres also proves the rationality of this method, providing feasible ideas for the low-cost and large-scale preparation of porous carbon materials in the future; However, this method still has its limitations. In the preparation process, it relies on strongly corrosive hydrofluoric acid and fails to find a good green substitute; At the same time, the research has not quantitatively explored the relationship between the template size and the product pore size, and the process stability and the accuracy of pore size control have yet to be verified.

3.2. Soft-template method

The soft template method uses organic molecules such as polymers and surfactants as templates. After they are self-assembled to form a stable structure in the solvent, carbon precursors are added. The two are combined through intermolecular forces such as hydrogen bonds and van der Waals forces. After high-temperature carbonization, the soft template is removed, and finally a carbon porous material with a reverse replication structure is formed ^[4]. However, the method also faces some existing challenges, where the pore size is difficult to regulate accurately, the self-assembly system is difficult to control fully, and the cost of soft templates is high.

In order to solve these problems, Liang et al. proposed a method of integrating a multi-component collaborative system to synthesize porous carbon spheres ^[6]. The team used corn straw alkali lignin as the carbon source, CTAB as the soft template, zinc ions as the cross-linking agent, and potassium bicarbonate as the activator. After mixing the four components, hydrothermal synthesis and high-temperature carbonization were carried out, and finally, the finished porous carbon spheres were prepared. Its electrochemical performance is good, and it is tested with a current density of 0.2 A/g. The material has a specific capacitance of 293 F/g, which is also at a high level in biomass derivatives; When the current density increases to 40 A/g, the initial capacity has a retention rate of 70.8%; after 10,000 charging and discharging cycles, the specific capacitance retention rate is 86.7%, demonstrating excellent cycling stability.

The team's innovation points in the synthesis process are mainly as follows. First, use low-cost and easily available corn straw as a carbon source to realize the effective use of biomass resources; The second is the introduction of zinc ions to form stable coordination bonds with hydroxyl groups and phenol hydroxyl groups in lignin, thus enhancing the stability of the complex and facilitating the control of the synthetic

self-assembly system; Finally, potassium bicarbonate is added, which will release carbon dioxide during decomposition, which can enrich the pore structure of porous carbon spheres and solve the problem that the conventional soft template method is difficult to regulate the pore size. This work shows that it is completely possible to prepare high-performance porous carbon materials using renewable biomass resources, and also provides new ideas for the rational utilization of biomass resources.

3.3. Salt-template method

The salt template method mainly uses NaCl, KCl, LiCl, and other water-soluble salts or their mixtures as templates. The carbon precursor is fully mixed with salts. After the crystal is precipitated, it is carbonized at high temperature under inert gas conditions. Finally, the salt template is removed with deionized water, and the product can be obtained after drying ^[2]. Unlike etching agents such as hydrofluoric acid used in the hard template method, the salt template method only needs deionized water to be washed, which is more environmentally friendly. However, the salt template method still has defects such as long washing time and difficulty in completely removing the template.

To address this issue, Feng et al. pioneered a self-sacrificial salt-template method, utilizing ammonium thiocyanate (NH_4SCN)—which readily decomposes upon heating, as the self-sacrificial salt ^[7]. In the process of high-temperature carbonization, ammonium thiocyanate will decompose and evaporate by itself, thus eliminating the subsequent tedious washing steps. At the same time, the decomposition of ammonium thiocyanate also promotes the formation of rich micropores inside the carbon material. The pore diameter is distributed at 0.6–1.1 nm, and the specific surface area of micropores is as high as 770 m^2/g .

This method fundamentally solves the problem of residual template in salt-template-synthesized porous carbons, and this strategy is worth adopting. However, it should also be pointed out that it has limitations; at present, it is only applicable to ammonium thiocyanate, which is easy to decompose, and it is difficult to generalize to conventional salt templates. In addition, the size control of crystal growth has not been solved. Therefore, although the salt template method has significant advantages from the perspective of green synthesis, it still faces great challenges in applying it to industrial production and commercialization.

4. Applications of heteroatom doping

The raw materials doped with heteroatoms contain precursor substances of non-carbon elements such as nitrogen, sulfur, phosphorus, and boron. In the process of high-temperature carbonization, heteroatoms will form a coordinate bond or covalent bond with the carbon skeleton, thus partially replacing the carbon atoms in the carbon lattice, and finally obtaining porous carbon materials with electron-deficient or electron-rich characteristics.

Heteroatom doping can effectively improve the performance of porous carbon electrodes. Take nitrogen as an example - research shows that nitrogen atoms can provide delocalized electrons for the carbon skeleton, thus improving the overall conductivity of the material; nitrogen doping also introduces defects in carbon crystals, effectively increasing its specific surface area, and the more doped functional groups, the greater the energy density ^[8]. At the same time, nitrogen-containing functional groups (such as amino groups, imino groups, and pyridinic nitrogen, etc.) introduced after doping are often hydrophilic, which can promote electrolyte penetration into the pores through capillary action.

Cui Guangfu's team at Northeast Petroleum University uses longan shell as a carbon source. Under the condition of high-temperature activation and hydrothermal doping, after using urea and nitric acid to dope porous carbon with N and O, the specific capacitance at 0.5 A/g is 463 F/g, and the retention rate of 10,000 cycles under 20 A/g is 89% ^[9]. The main reasons can be attributed to two points: first, the carbon derivatives prepared from longan shells themselves have a graded porous structure, which can promote the transportation of ions in the pores; the doping of N and O also contributes to part of the pseudo-capacitance, which further improves the surface wettability of carbon materials and is conducive to the full infiltration of the electrolyte solution. The work of the research team provides a green and feasible path for the preparation of high-performance porous carbon, and realizes the value-added utilization of biomass resources.

The doping of other heteroatoms also has a similar modification effect, and there is a certain synergy between them, but the doping mechanism has not been fully exposed. For example, the influence of doping concentration, doping position, doping method, and other factors on the electrochemical properties of materials is still a direction worth exploring in depth.

5. Conclusion

This paper reviews the performance optimization strategy of porous carbon supercapacitors from two aspects: template method and heteroatom doping method. In terms of template method, the hard template method can more accurately regulate the pore size, but the process of removing the template involves the use of strong corrosive substances such as hydrofluoric acid; The soft template method does not need to be etched to remove the template, but it is difficult to regulate the pore size; the salt template method only needs to be washed to remove the template, and even when introducing self-sacrificial salt, it can be exempted from washing, so it is the one with the greenest development potential among the three methods. In terms of the heteroatom doping method, doped heteroatoms can improve the overall electrochemical properties of carbon materials by providing delocalized electrons, improving the wettability of materials, and introducing carbon skeleton defects.

This article summarizes the typical strategies to improve the energy density of porous carbon supercapacitors, which can provide ideas and references for researchers to optimize supercapacitors.

Porous carbon supercapacitors have excellent electrochemical properties, so they are widely used in portable electronic equipment, new energy vehicles, and other fields. Although there is a large difference in energy density compared with commercial lithium-ion batteries, with the advancement of green synthesis technology, a deeper understanding of its energy storage mechanism is likely to become a major trend in future development.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Wang T, Liu D, Wang J, et al., 2026, Progress in the Application of Coal Tar Pitch-Based Porous Carbon in Supercapacitors. *Journal of China Coal Society*, 51(4): 2684–2699.

- [2] Jiao S, Yang L, Wu T, et al., 2021, Preparation of Nitrogen-Doped Hierarchical Porous Carbon Nanosheets for Supercapacitors by a Mixed-Salt Templating Method. *CIESC Journal*, 72(5): 2869–2877.
- [3] Kong Z, Li X, Xu R, et al., 2024, Research Progress in the Preparation and Application of Heteroatom-Doped Porous Carbon. *Journal of Jiangnan University (Natural Science Edition)*, 52(4): 15–26.
- [4] Liu W, Jiang L, Rong M, 2024, Progress in the Preparation and Application of Porous Carbon Materials. *New Chemical Materials*, 52(8): 8–14.
- [5] Zhu Y, Wang S, Ji L, 2021, Preparation and Supercapacitive Performance of Mesoporous Hollow Carbon Nanospheres. *Journal of Zhejiang Sci-Tech University (Natural Sciences Edition)*, 45(5): 575–580.
- [6] Liang Y, Yu Y, Qi X, 2025, Synthesis of Hollow Hierarchical Porous Carbon Spheres from Lignin by Soft Template and Hydrothermal Method for Supercapacitors. *International Journal of Biological Macromolecules*, 307: 141938.
- [7] Zhang Z, Feng J, Jiang Y, et al., 2018, Self-Sacrificial Salt Templating: Simple Auxiliary Control over the Nanoporous Structure of Porous Carbon Monoliths Prepared Through the Solvothermal Route. *Nanomaterials*, 8(4): 255.
- [8] Li L, 2024, Synthesis of Nitrogen-Doped Hierarchical Porous Carbon Nanospheres and Investigation of Their Application Performance, thesis, Dalian University of Technology.
- [9] Cui G, Guan Y, Huang N, et al., 2025, N,O-Codoped Porous Carbon Derived from Longan Shells for High-Performance Supercapacitor Electrodes. *ACS Applied Energy Materials*, 8(11): 7002–7012.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Exploration of the Application Value of Flexible Carbon Nanotube Electromyography Electrodes in Health Monitoring

Li Song, Jingjing Sun

Chongqing Energy College, Chongqing 402260, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: The development of wearable devices has posed new requirements for surface electromyography (EMG) electrodes, including long-term, comfortable, and high-fidelity signal acquisition. Traditional Ag/AgCl wet electrodes suffer from limitations such as skin irritation and signal attenuation. Carbon nanotubes (CNTs), with their excellent conductivity, flexibility, and biocompatibility, have emerged as ideal materials for constructing flexible dry EMG electrodes. This paper systematically reviews the material properties, preparation strategies, and signal acquisition performance of flexible CNT EMG electrodes, analyzes their application value in scenarios such as muscle function assessment, gesture recognition, motion monitoring, and multimodal integrated monitoring, discusses challenges such as scalable preparation, long-term stability, and standardized evaluation, and looks forward to directions such as self-powered systems, artificial intelligence integration, and smart textiles. This paper suggests that the electrode is expected to become the core sensing element of the next generation of wearable health monitoring systems, promoting the clinical translation of non-invasive monitoring technologies.

Keywords: Carbon nanotubes; Flexible electrodes; Electromyography signals; Health monitoring Classification Code: G712 Document Code: A

Online publication: Sept 3, 2026

1. Introduction

Electrophysiological signal detection, including electrocardiography (ECG), electromyography (EMG), and electroencephalography (EEG), represents a critical clinical testing method for assessing human health status and enabling real-time monitoring of patient well-being^[1]. With the advancement of wearable medical monitoring technologies, bioelectronics has found widespread applications in health detection, medical supervision, and motion detection^[2,3]. The efficient acquisition of high-quality electrophysiological signals has become a research focus^[4].

Currently, commonly used surface bioelectrodes have limitations. Gel electrodes rely on conductive

gel to reduce impedance, but gel drying during long-term monitoring can easily cause skin allergies and increase noise ^[5]. Wet electrodes require skin preparation for long-term use, and sweat can also lead to signal degradation ^[6,7]. Dry electrodes offer convenience and long-term wearability advantages, but they suffer from poor skin conformity, high impedance, and are prone to motion artifacts during movement ^[2,8]. Penetrating electrodes can reduce artifacts, but they lack mechanical stability and comfort ^[7,9].

An ideal skin electrode should possess high conductivity, high stretchability, softness, skin adhesion, and excellent biocompatibility ^[6,10]. Carbon nanotubes (CNTs), with their unique honeycomb lattice structure endowing them with excellent carrier mobility, mechanical flexibility, and chemical stability, have emerged as ideal materials for constructing flexible dry bioelectrodes. In recent years, researchers have successfully prepared various types of flexible CNT EMG electrodes through multiple strategies, achieving significant progress in signal acquisition and application expansion.

This paper systematically reviews the research progress of flexible CNT EMG electrodes in the field of health monitoring, starting from material properties and preparation methods, analyzing signal acquisition performance, discussing application prospects in scenarios such as muscle function assessment, gesture recognition, motion monitoring, and multimodal integrated monitoring, summarizing challenges, and looking forward to development directions.

2. Material Properties and preparation methods of carbon nanotube electromyography electrodes

2.1. Basic electrical and mechanical properties of carbon nanotubes

Carbon nanotubes (CNTs) are formed by rolling up single or multiple layers of graphene, with their honeycomb lattice structure composed of sp^2 hybridized carbon atoms endowing them with exceptional properties. Electrical properties: The conductivity of single-walled carbon nanotubes (SWCNTs) can reach 10^2 – 10^4 S/cm, owing to their unique electronic band structure that allows carrier mobility with minimal scattering limitations ^[1]. Mechanical properties: Their Young's modulus can exceed 1 TPa, enabling them to withstand significant bending and stretching deformations, making them suitable for compounding with flexible substrates to adapt to dynamic skin deformations. Chemical stability and biocompatibility: CNTs are resistant to corrosion or degradation in physiological environments. PDMS/SWCNT/RGO composite electrodes have demonstrated cell counts comparable to control groups in studies on connective tissue cell viability, with some composite materials exhibiting in vitro cell survival rates exceeding 80%, indicating that properly designed CNT-based electrodes possess a safety foundation for long-term skin contact ^[11,12].

2.2. Preparation strategies for flexible electrodes

Based on the assembly methods and processing techniques of carbon nanotubes, the preparation strategies for flexible carbon nanotube electromyography electrodes can be summarized into the following main approaches ^[13].

Direct synthesis method: Growing CNTs directly on flexible substrates can achieve good adhesion, but is limited by the contradiction between growth temperature and substrate thermal stability. The blown aerosol chemical vapor deposition method enables high-yield production of self-supporting dry films in a roll-to-roll manner, with controllable adjustment of thickness, sheet resistance, and light transmittance ^[14].

Transfer method: Patterned transfer of pre-grown CNT films onto flexible substrates allows independent

optimization of growth conditions and substrate selection. Patterning with a resolution of 200 μm can be achieved using laser-cut masks ^[14].

Solution processing method: CNTs are dispersed in a solvent and formed into films through coating, printing, or spraying. MWCNT/PDMS composite electrodes are prepared using a low-cost solution processing method, which is simple and easy to scale up ^[15].

Composite preparation method: CNTs are compounded with a polymer matrix to form composite materials that combine conductivity and flexibility, with performance adjustable by controlling doping concentration and dispersion state ^[13]. Laser irradiation can reduce the resistance of SWCNT/RGO/PDMS composite materials by 13 times to $8 \pm 2 \text{ k}\Omega$ ^[11].

2.3. Typical electrode configurations

Thin-film electrodes: Ultra-thin conformal CNT dry electrodes exhibit high transparency ($> 90\%$) and good conductivity ($60 \text{ }\Omega/\text{sq}$), enabling them to adhere to the skin in the form of temporary tattoos for ECG, EMG, and EEG monitoring ^[14].

Composite electrodes: MWCNT/PDMS composite electrodes demonstrate interface impedance comparable to that of Ag/AgCl electrodes in the frequency range of 20–1 kHz ^[15]. The elastic modulus of PDMS/SWCNT/RGO composite electrodes ($180.0 \pm 5.3 \text{ kPa}$) is highly matched with that of the skin ^[11]. MXene/CNT composite electrodes construct a three-dimensional structure with low impedance and high signal-to-noise ratio ^[16].

Multi-channel array electrodes: A 64-channel three-dimensional Ti_3C_2 MXene/CNT composite electrode array exhibits exceptional flexibility, enabling seamless conformity to skin curvature. Its contact impedance at 100 Hz is 10 times lower than that of Ag/AgCl gel electrodes, providing a hardware foundation for fine gesture recognition ^[16]. **Fabric electrodes:** CNTs combined with textiles form electrodes that combine conductivity and breathability, allowing direct integration into everyday clothing.

Electronic fabrics based on in-situ growth of CNTs on carbonized silk fabrics can monitor pulse waves and joint movements in real time ^[17]. Ultrathin nanosheets assembled from single-walled CNT networks on elastomers exhibit good water vapor permeability, allowing sweat to pass through during exercise ^[18].

3. Signal acquisition performance of flexible carbon nanotube electromyography electrodes

3.1. Skin-electrode interface impedance

Interface impedance directly affects signal fidelity and signal-to-noise ratio (SNR). The three-dimensional MXene/CNT composite electrode exhibits a contact impedance 10 times lower than that of the Ag/AgCl gel electrode at 100 Hz ^[16]. The MWCNT/PDMS composite electrode demonstrates comparable performance to traditional Ag/AgCl electrodes within the frequency range of 20–1 kHz ^[15]. The ultrathin CNT dry electrode achieves a contact impedance of $100 \text{ k}\Omega \cdot \text{cm}^2$ at 10 Hz, comparable to that of commercial wet electrodes ^[14]. Increasing mechanical pressure (0–100 mmHg) reduces impedance across the entire frequency range ^[15]. The PDMS/SWCNT/RGO composite electrode maintains impedance stability during 7 days of simulated sweat exposure ^[11].

3.2. Signal-to-noise ratio and signal quality

The MXene/CNT electrode exhibits lower baseline noise and higher SNR when acquiring surface

electromyography (sEMG) signals from the arm, maintaining a high SNR of approximately 25 dB across different gestures ^[16]. The ultrathin CNT dry electrode achieves an SNR of 67 dB for ECG monitoring, outperforming the 60 dB of Ag/AgCl wet electrodes ^[14]. The ECG signals recorded by the PDMS/SWCNT/RGO composite electrode are comparable in quality to those of Ag/AgCl electrodes, with sharper peak amplitudes ^[11]. Carbon nanomaterial-based dry electrodes combine excellent mechanical flexibility, high conductivity, and wear comfort due to their gel-free design ^[12].

3.3. Signal stability under dynamic conditions

The ultrathin profile and good skin adhesion reduce sensitivity to motion noise, effectively overcoming common motion artifacts associated with dry electrodes ^[2]. Flexible dry electrodes based on CNT/PDMS composites and wrinkled structures effectively control dynamic artifacts ^[5]. CNT-based dry electrodes exhibit high stability against motion artifacts, which typically arise from the inability of thicker rigid dry electrodes to conform to skin deformations ^[8]. The ultrathin electrode composed of single-walled CNT networks allows for the transmission of sweat vapor, reducing sweat accumulation at the interface ^[18].

3.4. Long-term monitoring reliability

The gel-free dry design fundamentally avoids signal attenuation caused by gel drying ^[14]. The PDMS/SWCNT/RGO composite electrode shows no degradation in ECG signals during 7 days of continuous wear, with no observed skin reactions and stable impedance changes recorded over 14 days ^[11]. The MWCNT/PDMS composite dry electrode outperforms commercial Ag/AgCl electrodes in ECG and EMG measurements, making it suitable for ubiquitous health monitoring ^[15]. The mechanical durability of CNT materials supports stable performance under repeated deformation cycles ^[13].

4. Applications of flexible carbon nanotube electromyography electrodes in health monitoring

4.1. Muscle function assessment and rehabilitation monitoring

Surface electromyography signals contain rich information about neuromuscular function and can be used to assess muscle fatigue, coordination, and functional status. Flexible carbon nanotube electromyography electrodes provide a new means for long-term, dynamic assessment of muscle function.

In the field of neurological rehabilitation, continuous sEMG monitoring holds significant clinical value for assessing rehabilitation progress. Carbon nanotube-based dry electrodes, which do not require conductive gel and can be worn for extended periods, are suitable for rehabilitation scenarios requiring continuous monitoring over days to weeks ^[2]. The ultrathin CNT dry electrode can adhere to the skin in the form of a temporary tattoo, enabling simultaneous monitoring of ECG, EMG, and EEG ^[14]. Its low-burden wearability facilitates patients maintaining natural movement states, allowing for the acquisition of more ecologically valid EMG data.

In sports medicine, CNT-based fabric electrodes can be directly integrated into sportswear, enabling real-time tracking of muscle activity during exercise ^[17]. This eliminates the need for additional electrode attachment, allowing athletes to obtain muscle activity data under natural training conditions.

4.2. Gesture recognition and human-computer interaction

Gesture recognition based on sEMG signals has broad application prospects in intelligent prosthesis control,

sign language recognition, and virtual reality interaction ^[10]. Carbon nanotube-based electromyography electrodes, with their high SNR and good dynamic stability, provide high-quality signal input for gesture recognition.

The MXene/CNT electrode achieves an SNR of approximately 25 dB when acquiring sEMG signals for different gestures, with recognition rates exceeding 90% when combined with machine learning ^[16]. The 64-channel three-dimensional MXene/CNT composite electrode array captures high-resolution spatial distribution information of arm muscle activity, providing a data foundation for fine gesture recognition ^[16]. A sign language recognition system supported by mixed ion-electron conductive electrodes achieves a classification accuracy of 96.4% ^[3]. CNT-based dry electrodes combined with machine learning demonstrate high precision and adaptive capabilities in gesture classification ^[4]. The deep integration of carbon nanotube-based electromyography electrodes with artificial intelligence is driving the development of high-precision, low-latency human-computer interaction interfaces.

4.3. Sports training and physical fitness monitoring

Monitoring physiological signals during sports training is crucial for assessing training effectiveness and preventing sports injuries. The wearability and comfort of carbon nanotube-based electrodes give them unique advantages in sports scenarios ^[2].

Fabric-type CNT electrodes can be directly integrated into sportswear, enabling real-time tracking of muscle activity during exercise. This avoids the issues of traditional electrode detachment and displacement, ensuring the continuity and reliability of signal acquisition ^[17]. Single-walled carbon nanotube-SBS nanosheets exhibit excellent resistance to water wiping, which is particularly important for sports scenarios involving frequent sweating and limb friction ^[10,18].

In the field of ergonomics, CNT-based wearable electromyography electrodes can be used to assess muscle load under different work postures, providing data support for workplace ergonomics optimization ^[2]. With the increasing awareness of health management, the demand for real-time monitoring of physiological signals in daily life is growing, and CNT-based electromyography electrodes offer an efficient and durable solution for this purpose ^[6].

4.4. Integrated monitoring of multimodal physiological signals

Single-modality physiological signals are insufficient to comprehensively reflect human health status. The synchronous acquisition and comprehensive analysis of multimodal signals represent an important direction for the next generation of wearable health monitoring systems ^[17]. Carbon nanotube-based flexible platforms demonstrate unique integration potential in this regard.

CNT-based flexible platforms can construct multimodal sensing platforms through flexible substrates and stretchable conductive materials, enabling synchronous acquisition of multidimensional physiological data such as ECG, EMG, body temperature, and sweat metabolites. This allows for comprehensive health status assessment on a single device ^[16].

At the signal processing level, CNT-based flexible amplifiers can achieve a gain of up to 1.5×10^5 (approximately 104 dB) ^[18]. Integrating sensing, amplification, signal processing, and wireless transmission functions into a single flexible platform represents an important development direction for CNT-based wearable health monitoring systems ^[4,10].

5. Challenges and prospects

5.1. Major current challenges

Although flexible carbon nanotube (CNT) electromyography (EMG) electrodes have made encouraging progress in basic research and application exploration, several key challenges remain in transitioning from the laboratory to practical clinical applications ^[1].

Scalable Manufacturing: Mass production of high-performance CNT electrodes at low cost remains a bottleneck for their widespread application. Currently, most research is limited to laboratory-scale preparation, lacking mature industrial production processes ^[13]. How to achieve continuous roll-to-roll, large-area, and low-cost production while maintaining performance is a core engineering issue that needs urgent resolution.

Long-term Stability and Durability: Although some studies have demonstrated continuous monitoring capabilities for 7 days or even 14 days, the long-term performance retention, structural integrity under repeated cleaning and wearing, and practical application scenarios, such as daily use over months or even years, still require systematic validation ^[11].

Lack of Standardized Evaluation Systems: Currently, there is a lack of unified testing standards and specifications for evaluating the performance of CNT-based electrodes. Differences in impedance measurement conditions, signal-to-noise ratio calculation methods, and long-term monitoring evaluation indicators among different studies make it difficult to compare different electrode systems horizontally ^[12]. Establishing industry-recognized standard testing methods is crucial for promoting technological maturity and industrialization.

Long-term Verification of Biosafety: Although existing studies indicate that CNT-based materials exhibit good biocompatibility at the skin contact level, more systematic and in-depth toxicological evaluations are needed for long-term, large-area skin contact ^[11,12]. The potential transdermal absorption pathways and biological effects of carbon nanomaterials still require further clarification ^[6].

Motion Artifact Suppression: Although flexible designs have significantly improved signal stability under dynamic conditions, ensuring signal quality during vigorous exercise and extreme sweating remains a technical challenge ^[8]. Collaborative efforts are needed at multiple levels, including material design, electrode structure optimization, and signal processing algorithms.

5.2. Future development directions

Self-powered Systems: Develop CNT-based self-powered sensors to enable autonomous monitoring without external power sources ^[1]. CNTs, which function as both electrode materials and energy harvesting components, have potential in self-powering technologies such as triboelectric generation and thermoelectric conversion due to their excellent electrical conductivity and mechanical properties ^[13]. Self-powered designs will significantly enhance the portability and continuous operation capabilities of wearable devices.

Integration with Artificial Intelligence: Combine machine learning for intelligent signal recognition and disease early warning ^[16]. The integration of high-density sEMG signals with deep learning algorithms is driving continuous improvements in gesture recognition accuracy ^[3,16]. In the future, artificial intelligence will not only be used for post-processing analysis of signals but can also be embedded in electrode-side signal preprocessing and feature extraction to enable edge intelligence ^[4].

Multifunctional Integration: Integrate sensing, amplification, signal processing, and wireless transmission functions into a single flexible platform ^[10]. Full flexibility and full integration are the

development goals of wearable health monitoring systems^[12]. CNTs can serve as both sensing materials and components for constructing flexible amplifiers and transistors, providing a material foundation for system-level integration^[18].

Smart Textiles: Seamlessly integrate CNT electrodes into daily clothing to enable unobtrusive health monitoring^[17]. Transitioning from standalone patch electrodes to fabric electrodes integrated into clothing is an important path to improving user compliance^[2]. Future smart clothing will combine daily wearing comfort with continuous physiological signal monitoring capabilities^[10].

Clinical Translation: Promote the technological translation from laboratory research to practical medical applications. This requires interdisciplinary collaboration, integrating expertise from material science, electronic engineering, clinical medicine, and data analysis^[6]. At the same time, large-scale clinical trials are needed to verify the effectiveness and safety of electrodes in real medical scenarios, obtain regulatory approval, and ultimately achieve a technological leap from hospital monitoring to home health management^[2].

6. Conclusion

Flexible carbon nanotube EMG electrodes exhibit significant application value in the field of health monitoring due to the unique electrical, mechanical, and chemical properties of CNT materials. In terms of signal acquisition performance, CNT-based dry electrodes have achieved contact impedance and signal-to-noise ratios comparable to or even better than commercial Ag/AgCl wet electrodes, while overcoming the inherent limitations of wet electrodes, such as gel drying and skin irritation during long-term use. In terms of application scenarios, these electrodes show broad prospects in muscle function assessment and rehabilitation monitoring, gesture recognition and human-computer interaction, sports training and physical fitness monitoring, and integrated monitoring of multimodal physiological signals.

However, to transition from laboratory research to large-scale clinical applications, flexible CNT EMG electrodes still need continuous efforts in scalable manufacturing processes, long-term stability verification, the establishment of standardized evaluation systems, and in-depth research on biosafety. Looking ahead, with the collaborative development of cross-disciplinary fields such as self-powering technologies, artificial intelligence algorithms, and smart textiles, flexible CNT EMG electrodes are expected to become the core sensing elements of the next generation of wearable health monitoring systems, driving a substantial leap in non-invasive health monitoring technology from hospital monitoring to home health management and providing critical technical support for personalized, continuous, and intelligent health management.

Funding

Research Project of the Chongqing Municipal Education Commission on Science and Technology—Research on the Preparation and Performance Optimization of Carbon Nanotube Adhesive Electromyography Electrodes in a Comprehensive Health Monitoring System (Project No.: KJ2025056144525339)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Kang B, Ha T, 2018, Wearable Carbon Nanotube Based Dry-Electrodes for Electrophysiological Sensors. *Japanese Journal of Applied Physics*, 57(5S): 05GD02.]
- [2] Zhang L, Kumar K, He H, et al., 2020, Fully Organic Compliant Dry Electrodes Self-Adhesive to Skin for Long-Term Motion-Robust Epidermal Biopotential Monitoring. *Nature Communications*, 11(1): 4683.
- [3] Yu J, Tian F, Wang W, et al., 2023, Design of Highly Conductive, Intrinsically Stretchable, and 3D Printable PEDOT:PSS Hydrogels via PSS-Chain Engineering for Bioelectronics. *Chemistry of Materials*, 35(15): 5936–5944.
- [4] Kireev D, Kampfe J, Hall A, et al., 2022, Graphene Electronic Tattoos 2.0 with Enhanced Performance, Breathability and Robustness. *npj 2D Materials and Applications*, 6(1): 46.
- [5] Lo L, Zhao J, Aono K, et al., 2022, Stretchable Sponge Electrodes for Long-Term and Motion-Artifact-Tolerant Recording of High-Quality Electrophysiologic Signals. *ACS Nano*, 16(8): 11792–11801.
- [6] Wang Y, Haick H, Guo S, et al., 2022, Skin Bioelectronics Towards Long-Term, Continuous Health Monitoring. *Chemical Society Reviews*, 51(9): 3759–3793.
- [7] Liu B, Yang Z, Zheng Y, et al., 2021, Preparation and Performance Study of Metal Microneedle Array Electrodes for Physiological Electrical Signal Measurement. *Journal of Mechanical Engineering*, 57(11): 61–68.
- [8] Lee E, Kim M, Lee C, 2019, Skin-Mountable Biosensors and Therapeutics: A Review. *Annual Review of Biomedical Engineering*, 21(1): 299–323.
- [9] Satti A, Kim J, Yi E, et al., 2021, Microneedle Array Electrode-Based Wearable EMG System for Detection of Driver Drowsiness Through Steering Wheel Grip. *Sensors*, 21(15): 5091.
- [10] Pan S, Zhang F, Cai P, et al., 2020, Mechanically Interlocked Hydrogel–Elastomer Hybrids for On-Skin Electronics. *Advanced Functional Materials*, 30(29): 1909540.
- [11] Kuksin A, Morozova S, Vasilevskaya O, et al., 2025, Dry Electrodes based on Carbon Nanotubes-Graphene Hybrid Nanostructures for Long-Term ECG Monitoring. *Diamond and Related Materials*, 159: 112895.
- [12] Shahab M, Rahim M, Mutahir S, et al., 2025, Carbon-based Nanocomposites Powering the Future of Wearable and Bioelectronics: A Review. *Journal of Alloys and Compounds*, 1043: 184296.
- [13] Wang X, Jin C, Bai Z, 2025, Review of Carbon Nanotube-Based Flexible Wearable Strain Sensors: From Materials to Applications for the Human Body. *Sensors and Actuators A: Physical*, 382: 117071.
- [14] Wen Y, Fox D, Cai Y, et al., 2024, Ultra-Thin Conformal CNT Dry Electrodes for Surface Electrophysiology Monitoring. 2024 MRS Fall Meeting.
- [15] Sharma P, Baloda S, Verma D et al., 2024, Multiwall Carbon Nanotube/Polydimethylsiloxane Composites-Based Dry Electrodes for Bio-Signal Detection. *IEEE Journal on Flexible Electronics*, 3(3): 108–114.
- [16] Cheng L, Guo A, Li J, et al., 2024, Three-Dimensional MXene/Carbon Nanotube Composite Electrodes in Flexible 64-Channel Arrays for Noninvasive Electromyography Signal Acquisition. *Science China Materials*, 67(9): 2977–2984.
- [17] Wang M, Liu L, Yang H, et al., 2026, Integrated Fabrication and Applications of Multifunctional Electronic Fabrics Based on Carbon Nanotubes/Carbonized Silk Fabrics. *Chemical Engineering Journal*, 525: 175376.
- [18] Zhang H, Yuan Y, Hu J, et al., 2025, Advances of Carbon Nanotube Based Flexible Amplifiers for Skin-Mounted Physiological Signal Monitoring. *Advanced Electronic Materials*, 11(10): 2400991.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

LSTM-Based Operating-State Forecasting and Risk Warning for Small-Scale Battery Energy Storage Systems Using Multi-Source Aging Data

Qian Wang^{1*}, Xing Wan², Lihua Luo³

¹School of Artificial Intelligence, Leshan Vocational and Technical College, Leshan, China

²School of Intelligent Manufacturing, Leshan Vocational and Technical College, Leshan, China

³Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia

*Corresponding author: Qian Wang, wqian8603@gmail.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Senseless computing resources are needed to achieve reliable forecasting in small-scale battery energy storage. In this research, the authors present an LSTM-based framework to normalize NASA and CALCE aging data, separate the data for battery-level prediction, predict 5-cycle-ahead state of health (SOH), and transform residual and health signals to warnings. In this work, LSTM was tested against the NASA and CALCE datasets using a common seven-model protocol and obtained competitive cross-cell performance (NASA: MAE 0.01219, RMSE 0.01619; CALCE: MAE 0.01149, RMSE 0.02331). Controlled anomaly tests detected all severe abrupt drops, and low-SOH risk was separated from model unexplained deviations by replay on CALCE CS2_38. Results of a rerun with an independent CPU were consistent with the original rankings of models and the statistical interpretation. The result is that the framework offers a simple and repeatable foundation for battery-state monitoring, but validation of the field-fault is still required.

Keywords: Battery energy storage system; Long short-term memory; State of health; Time-series forecasting; Risk warning

Online publication: Sept 3, 2026

1. Introduction

Battery energy storage systems (BESSs) support renewable-energy integration, yet many small-scale systems lack regular capacity tests and advanced sensing. Routine cycling records can therefore provide early evidence of degradation through state-of-health (SOH) forecasting. SOH estimation methods include experimental, model-based, and data-driven approaches ^[1]. LSTM networks are designed to capture long-term dependencies, and recent battery studies have combined LSTM with engineered health indicators, hybrid feature extraction, and transfer learning ^[2-5]. However, reported comparisons are often weakened by

inconsistent datasets, warning rules that conflate low SOH with model error, and cycle-level leakage across train-test splits.

In this study, the new recurrent unit is not developed, but a compact evidence chain is built instead. The software standardizes NASA and CALCE data under explicit quality checks, separates batteries for training, validation, and testing, compares seven models under training, validation, and testing with the same windows, horizons, and repeats, links forecasts to residual and/or health-based warnings, and verifies run-level reproducibility on an independent CPU computer. All models were implemented in PyTorch^[6].

2. Materials and methods

This section describes the data sources, preprocessing procedure, forecasting model, and risk-warning method used in this study. It first introduces the multi-source battery aging data and cycle-level feature construction, followed by the LSTM forecasting framework, baseline models, evaluation metrics, and residual-based warning rules.

2.1. Data and preprocessing

The NASA dataset comprised B0005, B0006, B0007, and B0018 (636 processed discharge cycles), and the CALCE CS2 set comprised CS2_35–CS2_38 (3,915 processed cycles)^[7,8]. The CALCE CS2 dataset has been widely adopted for battery prognostics and SOH estimation studies^[9–11]. The cycle-level data were mapped to a common schema containing capacity, voltage, current, time, and available temperature fields.

Processing included field and unit checks, discharge-cycle aggregation, duplicate removal, continuity checks, missingness screening, and physical-range audits. SOH was calculated as $SOH_t = C_t / C_0$, where C_0 is the first valid capacity. Cell-level splitting prevented adjacent observations from the same cell from appearing in both training and test sets.

2.2. Forecasting and warning framework

The input consists of the cycle-level feature vectors from the previous W cycles, and the prediction target is the SOH at cycle $t+H$. The main experiments used $W = 12$ and $H = 5$. The LSTM estimates SOH change and then generates the future SOH value using Equation (1). The model was trained using mean squared error and the Adam optimizer. Early stopping was applied according to the validation loss, and the checkpoint with the lowest validation loss was retained. The baselines were persistence, SVR, random forest, vanilla RNN, GRU, and TCN.

$$\begin{aligned}\widehat{SOH}_{t+H} &= SOH_t + \widehat{\Delta SOH}_{t+H} \\ e_{t+H} &= \widehat{SOH}_{t+H} - SOH_{t+H}\end{aligned}\tag{1}$$

The mean squared error, Adam, validation set early stopping, and the best checkpoint were employed for training. The validation set was used to calibrate the residual thresholds. The forecast residual was defined as $e_{t+H} = \widehat{SOH}_{t+H} - SOH_{t+H}$. Moderate and strict thresholds were applied to the absolute residual, $|e_{t+H}|$ to define the attention and warning states, respectively. A persistent warning was triggered when the strict threshold was exceeded repeatedly. An additional rule, $SOH \leq 0.80$, was used to identify low battery health independently of unexplained model deviations. Thus, low SOH and abnormal forecast residuals were retained as two distinct

risk sources. See Figure 1.

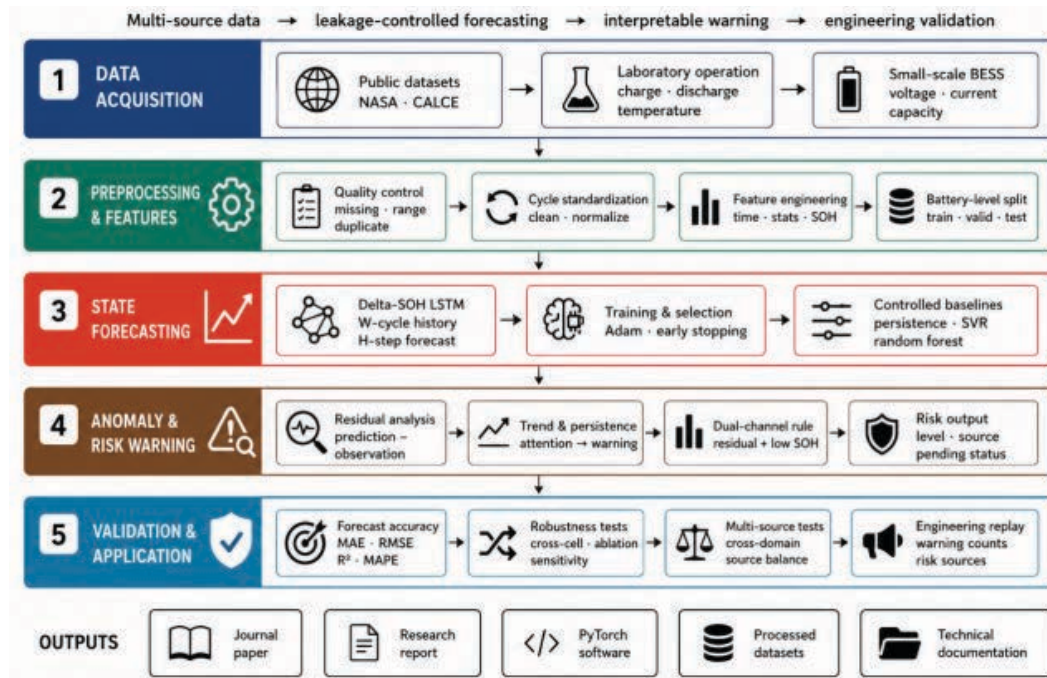


Figure 1. Workflow for multi-source preprocessing, leakage-controlled forecasting, warning calibration, and validation.

3. Experimental design

NASA held-out training was on B0005 / B0006, validation was B0007, and test was B0018. Cross-cell robustness was based on 4 folds and 5 seeds per fold. Ten seeds of CS2_35/36 were used for training, while a different ten seeds of CS2_37 were used for validation, and a different ten seeds of CS2_38 were used for testing for all CALCE testing. NASA-only and joint-source training involved the same CALCE test cell. The metrics used were MAE, RMSE, R², and MAPE. Holm correction at $\alpha = 0.05$ was used for paired tests.

Since the datasets do not have known operational errors, abrupt decreases, progressive degradation, and transients were only introduced to evaluate algorithmic response. There was a total of 210 runs in the full seven-model benchmark. The data used here were processed identically; the battery split was the same, and the seeds and settings were the same; an independent CPU rerun of the data resulted in 202 runs within 10⁶, and eight TCN runs had small differences that did not affect any ranking or significance conclusion. No sampling-level predictions were kept, and hence only run-level reproducibility is claimed.

4. Results and discussion

This section presents and discusses forecasting and risk-warning results. It first evaluates the predictive performance of the LSTM against the baseline models and then examines the effects of different data-source configurations, cross-domain generalization, and residual-based warning performance.

4.1. Forecasting performance

LSTM showed the best performance on all metrics in the fixed NASA holdout (MAE 0.01489; RMSE

0.02002; R^2 0.9185), with a 10.9% reduction in RMSE compared to the persistence method. When analyzing both data sets within the same repeated protocol (**Table 1**), no model was found to dominate both. SVR achieved minimum CALCE MAE, GRU achieved minimum CALCE RMSE, and NASA MAE and random forest achieved minimum NASA RMSE. LSTM was always in the lead, and it trained faster than the other deep sequence models, having 29,729 parameters and a state dictionary of 119.6 KiB. So, the evidence suggests that LSTM is a viable, small core forecasting system; it's not best all the way. See **Table 1**.

Table 1. Principal unified-protocol results (mean $\pm$ SD where repeated)

Scenario	Model	MAE	RMSE	R^2
CALCE	LSTM	0.01149 $\pm$ 0.00016	0.02331 $\pm$ 0.00012	0.9832
CALCE	GRU	0.01142 $\pm$ 0.00023	0.02324 $\pm$ 0.00004	0.9833
CALCE	RNN	0.01165 $\pm$ 0.00021	0.02336 $\pm$ 0.00013	0.9831
CALCE	TCN	0.01153 $\pm$ 0.00020	0.02329 $\pm$ 0.00010	0.9832
CALCE	SVR	0.01125	0.02328	0.9833
NASA	LSTM	0.01219 $\pm$ 0.00413	0.01619 $\pm$ 0.00464	0.9614
NASA	GRU	0.01189 $\pm$ 0.00391	0.01597 $\pm$ 0.00432	0.9624
NASA	RNN	0.01250 $\pm$ 0.00420	0.01644 $\pm$ 0.00453	0.9594
NASA	TCN	0.01223 $\pm$ 0.00408	0.01616 $\pm$ 0.00448	0.9615
NASA	Random forest	0.01238	0.01579	0.9640

It was more important to have domain shift than to have an architecture choice. On the CALCE test cell, CALCE-only training achieved MAE 0.01149 and RMSE 0.02331, compared with 0.01758 and 0.03258 for NASA-only training (Holm-adjusted $p = 0.0234$). Joint training showed very close in-domain performance (MAE 0.01174; RMSE 0.02317); unfortunately, not all metrics improved. However, source diversity does not necessarily transfer automatically from pooled data and will need explicit validation.4.2.

4.2. Warning response and reply

The strict detection was 100% at 6% severity, and higher for abrupt deviations than for progressive and transient deviations, but the attention channel enhanced sensitivity (**Figure 2**). In the 1,010-cycle CS2_38 replay, forecasting achieved MAE 0.01151, RMSE 0.02325, and R^2 0.98329. 371 critical labels were activated by $SOH \leq 0.80$. 18 strict residual exceedances – 15 were below the low SOH, and three were above. This decomposition allows a battery with potentially degraded performance but one having a known characteristic to be distinguished from an unexpected model error. The anomaly results quantify only controlled sensitivity, and not real-fault precision or recall. See **Table 2**.

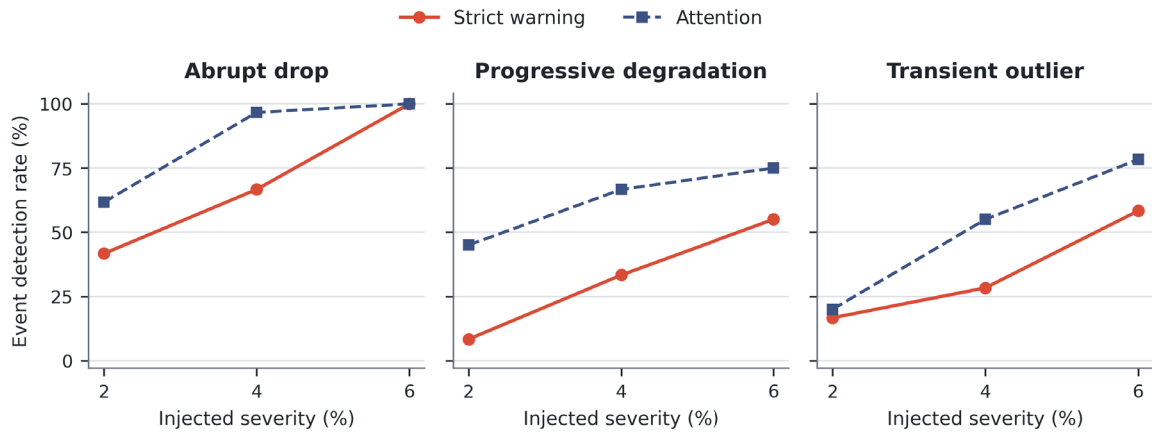


Figure 2. Strict-warning and attention-level event detection by simulated anomaly type and severity.

Table 2. Warning response and CS2_38 replay summary

Condition	Strict	Attention/share	Interpretation
Abrupt drop, 6%	100.0	100.0	Controlled event
Progressive, 6%	55.0	75.0	Controlled event
Transient, 6%	58.3	78.3	Controlled event
CS2_38 low SOH	371 cycles	36.73%	Health rule
CS2_38 strict residual	18 cycles	1.78%	Model deviation

Its major drawbacks are lab aging data, simulated and not proven fault, replaying offline single cells, and lack of saved sample-level prediction for independent metric recalculation. Field deployment should include a record of predictions, incorporate BMS data and maintenance records, recalibrate thresholds, and document expert-confirmed event recall, false-alert rates, and warning lead time.

4.3. Practical implications

The evidence comes together and suggests three implications. For one, when the data boundary of a compact neural model is not explicit, it is helpful. The cross-domain experiments reveal that the variations in cell design, cycling protocol, measuring instruments, as well as features can be as large as the performance disparities between LSTM, GRU, TCN, and conventional regressors. A deployment pipeline should thus be able to determine where each new record comes from, be able to detect feature drift, and initiate recalibration if the target system is outside the validation domain. Second, the semantics of warning should be maintainable. Split reporting of SOH and residual exceedance will enable the operator to distinguish between normal aging and an unexpected trajectory, and attention and strict thresholds will give different levels of sensitivity in one joint alarm.

Thirdly, reproducibility should be considered as an operational control. The independent CPU rerun validated that the central ranking and statistical interpretation were independent of any one computer, but there were minor differences in TCN. Each forecast should be stored to a field implementation with its model version, scaler, threshold, input window, and decision about the event. This would enable reconstruction, post-event review, and controlled updates to models at the sample level. In this respect, these practices are suitable for small storage installations and provide a more valuable monitoring component for the proposed framework than an accuracy demonstration on its own.

5. Conclusion

A small LSTM framework was integrated with multi-source data quality, leakage-controlled SOH forecasting, fair model comparison, and interpretable warning rules. Results at NASA and CALCE indicate competitive forecasting, high sensitivity to extreme sharp turnaround errors, and good separation between low health and unexplained residual risk. The main conclusions were maintained in the rerun by an independent CPU. However, operational claims must wait for prospective validation with prospective BMS and fault records.

Data and code

NASA and CALCE public archives ^[7,8]. The source code will be made publicly available in a permanent repository upon acceptance.

Author contributions

Q.W. conducted the experiments and drafted the manuscript. X.W. conceived and designed the study. L.H.L. analyzed the data and contributed to manuscript preparation. All authors reviewed and approved the final manuscript.

Funding

Leshan Engineering Technology Research Center for Energy Storage Materials and Systems through the project “Research on Operation State Perception and Risk Early Warning Methods for Small-Scale Energy Storage Systems” (Project No.: CNY202607)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Yang K, Zhang L, Zhang Z, et al., 2023, Battery State of Health Estimate Strategies: From Data Analysis to End-Cloud Collaborative Framework. *Batteries*, 9(7): 351.
- [2] Hochreiter S, Schmidhuber J, 1997, Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.
- [3] Zhao S, Luo L, Jiang S, et al., 2023, Lithium-Ion Battery State-of-Health Estimation Method Using Isobaric Energy Analysis and PSO-LSTM. *Journal of Electrical and Computer Engineering*, 2023: 5566965.
- [4] Feng S, Song M, Lin Y, et al., 2024, Convolutional Neural Network-Long Short-Term Memory-Based State of Health Estimation for Li-Ion Batteries Under Multiple Working Conditions. *Energy Technology*, 12(2): 2301039.
- [5] Lin M, Hu D, Meng J, et al., 2025, Transfer Learning-Based Lithium-Ion Battery State of Health Estimation with Electrochemical Impedance Spectroscopy. *IEEE Transactions on Transportation Electrification*, 11(3): 7910–7920.
- [6] Paszke A, Gross S, Massa F, et al., 2019, PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.
- [7] NASA 2026, Li-Ion Battery Aging Datasets, visited on 29 Jul 2026, <https://data.nasa.gov/dataset/li-ion-battery->

aging-datasets

- [8] Center for Advanced Life Cycle Engineering, n.d., CALCE Battery Data: CS2 Battery, University of Maryland, visited on 29 Jul 2026, <https://calce.umd.edu/battery-data>.
- [9] He W, Williard N, Osterman M, et al., 2011, Prognostics of Lithium-Ion Batteries Based on Dempster-Shafer Theory and the Bayesian Monte Carlo Method. *Journal of Power Sources*, 196(23): 10314–10321.
- [10] Xing Y, Ma E, Tsui K, et al., 2013, An Ensemble Model for Predicting the Remaining Useful Performance of Lithium-Ion Batteries. *Microelectronics Reliability*, 53(6): 811–820.
- [11] Williard N, He W, Osterman M, et al., 2013, Comparative Analysis of Features for Determining State of Health in Lithium-Ion Batteries. *International Journal of Prognostics and Health Management*, 4(1): 1–7.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Electrochemical Oxidation Dispersion of Graphene and Improvement of Its Compatibility with Coating Matrices

Qian Zhang*, Kaijie Yuan, Manjiao Deng, Hailin Wu

Grahope New Material Technology Inc., Shenzhen 518063, Guangdong, China

*Corresponding author: Qian Zhang, zqsky2008@163.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: The research employs a mild, controllable, and efficient electrochemical oxidation process using high-quality monolayer graphene powder prepared by a liquid-phase thermochemical route as the starting material. By optimizing key parameters including applied voltage and electrolyte concentration, a small number of hydrophilic oxygen-containing functional groups were selectively introduced at the edges of graphene, yielding a stably dispersed aqueous solution. The optimal conditions were determined as: graphene sheet anode, copper cathode, 0.4 mol/L NaOH electrolyte, voltage 8 V, and treatment time 30 min. Raman spectroscopy revealed an ID/IG ratio of 0.42. Thermogravimetric analysis and X-ray photoelectron spectroscopy confirmed that the oxygen content of electrochemically oxidized graphene (EGO) is low and that oxygen species are predominantly located at the edges; after reduction, the oxygen signals completely vanished, whereas residual oxygen remained in reduced GO. The as-prepared EGO aqueous dispersion achieved a concentration of 0.66 mg/mL and remained stable over six months of storage. TEM and AFM verified that the product retains a monolayer structure with an intact lattice. This approach preserves the intrinsic structure of graphene to the greatest extent while ensuring satisfactory dispersibility, thus offering a feasible pathway for fabricating high-performance graphene/siloxane ceramic resins and heavy-duty anti-corrosion coatings.

Keywords: Graphene; Electrochemical oxidation; Dispersion stability; Edge oxidation; Coating compatibility

Online publication: Sept 3, 2026

1. Introduction

Graphene, a two-dimensional carbon nanomaterial, possesses ultra-high specific surface area (2630 m²/g), outstanding mechanical, electrical, thermal conductivity, and chemical stability^[1]. It shows great potential in heavy-duty anti-corrosion coatings, composites, energy storage, etc^[2,3]. Especially in anti-corrosion coatings, its lamellar structure blocks permeation of corrosive media (water, oxygen, chloride ions), and its conductivity suppresses electrochemical corrosion, making it an ideal functional filler^[4-7]. However, two

major bottlenecks hinder practical use: (i) strong van der Waals forces (2.3 eV/nm^2) and large surface area cause agglomeration, preventing uniform dispersion; (ii) inert surface leads to poor interfacial compatibility with polar coating matrices (e.g., epoxy resins), and direct mixing causes phase separation^[8,9].

To overcome these, various strategies have been explored^[10,11]. Physical methods (high-speed shearing, ultrasonication) are simple but unstable and limited in concentration. Chemical modification introduces conjugated groups (phenyl, pyrenyl) to enhance π - π interaction. For instance, phenylbearing silane monomers stack onto graphene, then crosslink with silane coupling agents to improve matrix compatibility, but this requires stable concentrated aqueous dispersions.

Currently, two methods prepare aqueous dispersions: (1) direct ultrasonic or solventassisted exfoliation, simple but poor stability, low concentration ($< 0.1 \text{ mg/mL}$), and reagglomeration; (2) Hummers method or modified, using strong oxidants to oxidize graphite to GO, then disperse in water, but this severely disrupts sp^2 conjugation, introduces many inplane defects, degrading electrical and mechanical properties irreparably^[12]. Thus, a mild, controllable, efficient, and environmentally friendly method is urgently needed to ensure aqueous dispersibility while preserving pristine structure.

Recently, electrochemical oxidation has attracted attention due to mild conditions, controllable process, and selective modification^[12–14]. Under an electric field, reactive oxygen species preferentially attack graphene edges, introducing limited oxygen groups (carboxyl, hydroxyl) to impart hydrophilicity without damaging the inplane lattice. In this study, highquality monolayer graphene powder prepared by a liquidphase thermochemical route was electrochemically oxidized. Effects of voltage and electrolyte concentration on structure, oxygen content, and dispersion stability were systematically investigated. Through a series of characterizations, the effectiveness and superiority of this strategy are demonstrated, providing a theoretical basis and technical support for graphene application in coatings.

2. Experiments

2.1. Materials and reagents

(1) Graphene powder

Self-prepared by a liquid-phase thermochemical method (graphitization at 3000°C), confirmed by SEM, TEM, AFM, and particle size analysis to be large-size (ca. $7 \mu\text{m}$), monolayer, and highly crystalline.

(2) N-methyl-2-pyrrolidone (NMP) and polyvinylidene fluoride (PVDF)

Analytical grade, commercially available.

(3) Sodium hydroxide (NaOH)

Analytical grade, for electrolyte preparation.

(4) Copper foil (purity $\geq 99.9\%$)

Used as cathode.

(5) Dialysis membrane (molecular weight cutoff: 1000 Da)

For product purification.

2.2. Preparation of EGO

The preparation involved five steps:

(1) Slurry preparation

Graphene powder and PVDF were mixed at a mass ratio of 10:1, and an appropriate amount of NMP

was added. The mixture was magnetically stirred at room temperature for 18 h to obtain a homogeneous slurry.

(2) Coating and drying

The slurry was coated onto a conductive substrate (e.g., titanium foil) by dip coating, followed by vacuum drying at 70–100 °C to obtain graphene sheet electrodes.

(3) Electrochemical oxidation treatment

The graphene sheet served as the anode, a copper foil as the cathode, and NaOH solution as the electrolyte. The treatment was conducted at a constant voltage for 30 min. Using the single-variable method, the effects of voltage (4, 6, 8, 10, 12 V, with fixed NaOH concentration of 0.4 mol/L) and electrolyte concentration (0.1, 0.2, 0.4, 0.6, 0.8 mol/L, with fixed voltage of 8 V) were investigated separately.

(4) Dialysis and concentration

The resulting solution was transferred into a dialysis bag and dialyzed against deionized water for 24 h to remove residual electrolytes and by-products, followed by concentration.

(5) Drying

Part of the concentrated solution was freeze-dried to obtain solid EGO powder for solid-state characterizations; the remaining liquid was kept for concentration and stability tests.

2.3. Characterization methods

The preparation involved five steps

(1) Raman spectroscopy

Performed on a confocal Raman spectrometer (excitation wavelength 532 nm) to analyze the positions and intensity ratios of D, G, and 2D bands for defect density evaluation.

(2) Thermogravimetric analysis (TGA)

Conducted under a nitrogen atmosphere from room temperature to 1000 °C at a heating rate of 5 °C/min to determine the weight loss profile and estimate the content of oxygen-containing functional groups.

(3) Elemental analysis

Carried out using X-ray photoelectron spectroscopy (XPS) and energy-dispersive spectroscopy (EDS) to determine carbon/oxygen atomic percentages.

(4) Fourier-transform infrared spectroscopy (FTIR)

KBr pellet method, scanning range 400–4000 cm⁻¹.

(5) UV–visible spectroscopy (UV–vis)

Aqueous solutions, wavelength range 200–800 nm.

(6) Scanning electron microscopy (SEM) and transmission electron microscopy (TEM)

For morphology, lateral size, lattice observation; selected area electron diffraction (SAED) to verify crystal symmetry.

(7) Atomic force microscopy (AFM)

Tapping mode to measure thickness and evaluate lateral dimensions.

(8) Concentration determination

A known volume of EGO dispersion was freeze-dried, and the dry mass was weighed to calculate the mass concentration.

(9) Stability test

All samples were stored under identical ambient conditions for six months, and visual observation of sedimentation or stratification was recorded.

3. Results and discussion

3.1. Morphological characterization of the prepared graphene powder

The morphology and structure of the liquid-phase-derived graphene powder were characterized. TEM (**Figure 1**) showed large (micron-sized) transparent sheets, confirming their extreme thinness. HR-TEM (**Figure 2a, 2b**) displayed a large-area (20 nm $\times$ 20 nm) defect-free hexagonal honeycomb lattice; the edge regions exhibited ultra-thin features, and the FFT pattern showed hexagonal symmetry, confirming single crystallinity. AFM (**Figure 3**) gave a thickness of about 0.6 nm, consistent with the theoretical thickness of monolayer graphene (ca. 0.34 nm, considering adsorbed layers). Particle size distribution (**Figure 4**) centered at approximately 7 μm , ranging from 1 to 30 μm . These results confirmed that the starting material is high-quality, large-size, monolayer graphene, providing an ideal substrate for subsequent electrochemical oxidation.

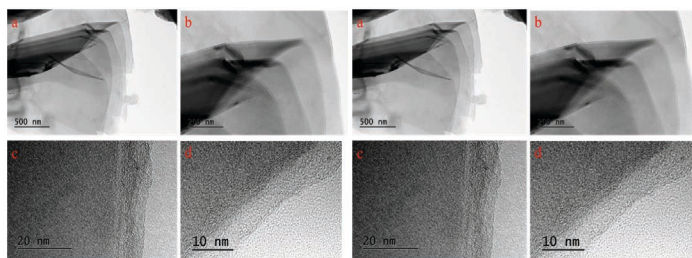


Figure 1. TEM image of a 3000°C graphitized graphene.

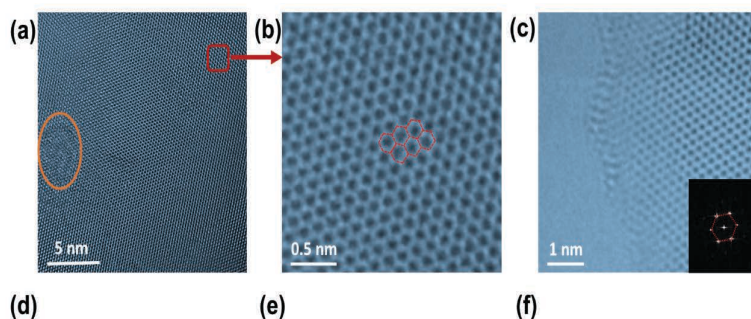


Figure 2. HR-TEM image of a 3000 °C graphitized graphene.

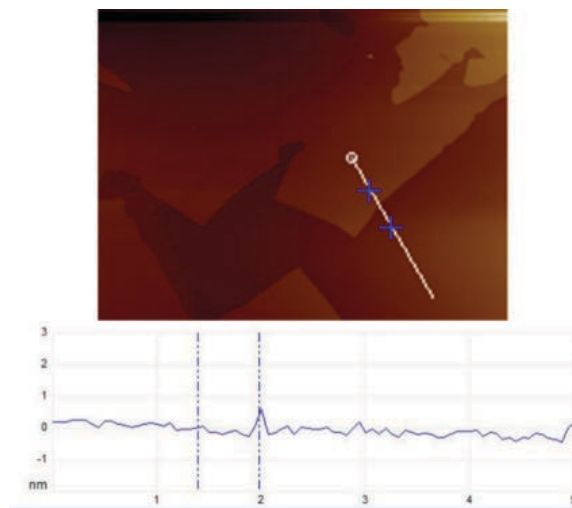


Figure 3. AFM image of a 3000 °C graphitized graphene.

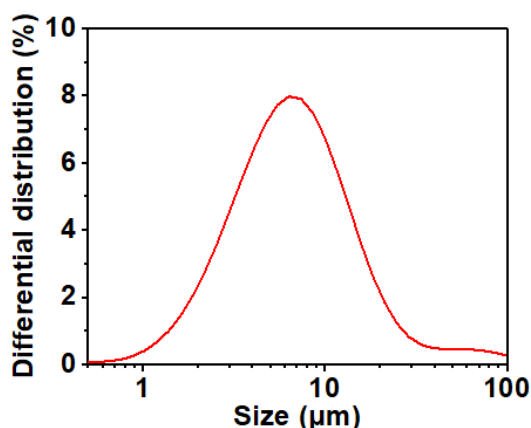


Figure 4. Particle size distribution of graphene powder.

3.2. Effect of voltage on EGO preparation

With a fixed NaOH concentration of 0.4 mol/L and treatment time of 30 min, the effect of applied voltage from 4 to 12 V was examined.

(1) Structural integrity (Raman)

As shown in **Figure 5** and **Table 1**, the D, G, and 2D band positions remained essentially unchanged (around 1350, 1580, and 2680 cm^{-1}) for all samples, indicating no significant chemical doping. However, with increasing voltage, the D-band intensity continuously increased, indicating more edge defects due to the introduction of hydrophilic groups. The G-band intensity first increased and then decreased, reaching a maximum at 8 V. The calculated ID/IG ratio reached its minimum (0.42) at 8 V, suggesting the fewest relative defects; the IG/I2D ratio was also smallest (0.67) at 8 V, indicating that the 2D band remained sharp and the lattice order was high. Therefore, 8 V was identified as the optimal voltage.

(2) Oxygen content (TGA and elemental analysis)

TGA (**Figure 6**) showed three-stage weight loss for all samples: < 200 °C (desorption of adsorbed water), 200–400 °C (decomposition of edge oxygen-containing groups), and > 400 °C (decomposition of more stable in-plane oxygen groups). With increasing voltage, the total weight loss increased, especially

above 8 V where the third-stage loss became pronounced, suggesting that oxygen groups penetrated the basal plane. Elemental analysis (**Figure 7**) further confirmed that as voltage rose from 4 V to 12 V, carbon content decreased from 98.89% to 64.52% and oxygen increased from 1.11% to 35.38%, with a sharp inflection at 8 V, indicating that 8 V is the threshold between selective edge oxidation and excessive bulk oxidation.

(3) Concentration and stability

Concentration measurements (**Figure 8**) showed that from 4 V to 8 V, the concentration jumped from 0.018 to 0.66 mg/mL, a dramatic increase; beyond 8 V, the increment became marginal (0.73 and 0.75 mg/mL). In terms of stability, samples at 4–8 V remained stable after six months, while the 10 V sample was “relatively stable” and the 12 V sample was unstable. Thus, 8 V yields the best compromise between good dispersibility and minimal oxidation.

Table 1. The ratio of peak intensity to peak height of EGO obtained under different voltages

	4V	6V	8V	10V	12V
D-peak intensity	188	523	868	1018	1427
G-peak intensity	270	490	2022	952	1269
2D-peak intensity	96	584	2996	139	188
I_D/T_G	0.69	1.06	0.42	1.06	1.12
I_G/T_{2D}	2.81	0.83	0.67	6.84	6.75

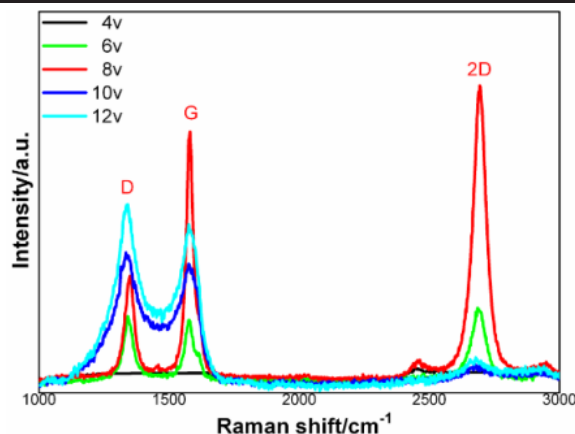


Figure 5. Raman spectra of EGO under different voltages.

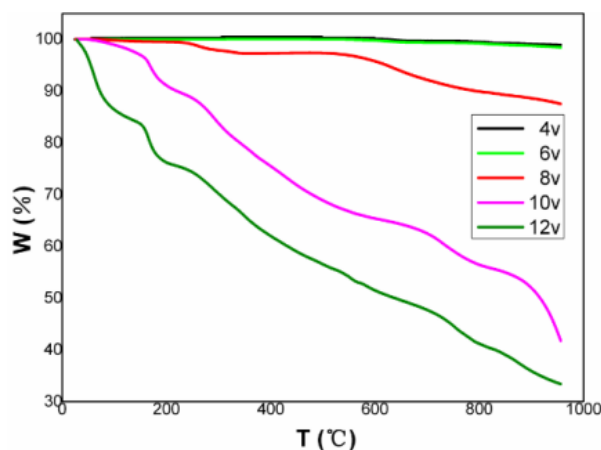


Figure 6. TGA of EGO under different voltages.

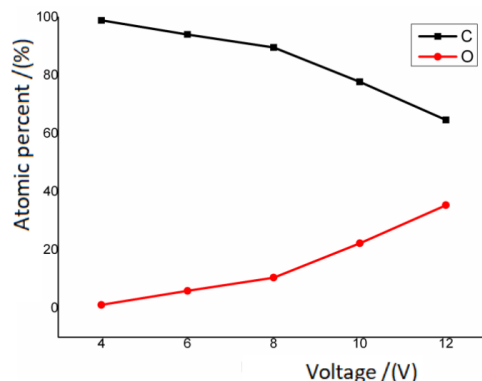


Figure 7. Carbon and oxygen element analysis of EGO under different voltages.

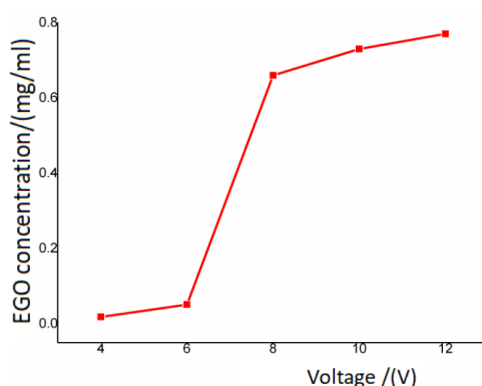


Figure 8. Concentration of EGO under different voltages.

3.3. Effect of electrolyte concentration on EGO preparation

With a fixed voltage of 8 V, NaOH concentrations from 0.1 to 0.8 mol/L were investigated.

(1) Structure (Raman)

Figure 9 and **Table 2** indicate that as concentration increased, the D-band intensity gradually rose; the ID/IG ratio at 0.4 mol/L was 0.73, slightly higher than 0.19 at 0.1 mol/L, but the latter showed a high IG/I2D of 2.37 (indicating significant broadening of the 2D band). Overall, the 2D band remained sharp at 0.4 mol/L, and the degree of defects was acceptable. In general, the influence of electrolyte concentration on structure was weaker than that of voltage.

(2) Oxygen content (TGA and elemental analysis)

TGA (**Figure 10**) also exhibited three-stage weight loss, but the loss increased more gently with concentration, and no further loss occurred above 600 °C. Elemental analysis (**Figure 11**) showed oxygen content increasing from 7.03% at 0.1 mol/L to 18.63% at 0.8 mol/L, without abrupt changes, indicating a gradual effect, with overall oxidation levels much lower than those at high voltage (35.38% at 12 V).

(3) Concentration and stability

Concentration data (**Figure 12**) indicated that at 0.4 mol/L the concentration reached 0.67 mg/mL, after which it nearly saturated (0.72–0.74 mg/mL). Samples at 0.4 mol/L and below were stable; 0.6 mol/L was “relatively stable”, and 0.8 mol/L was unstable. Considering all factors, 0.4 mol/L was chosen as the optimal electrolyte concentration.

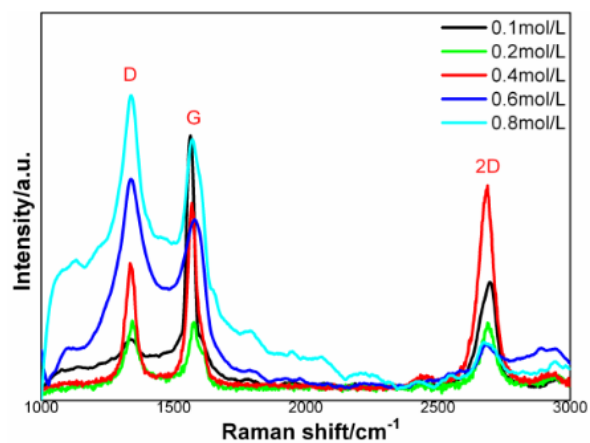


Figure 9. Raman spectra of EGO under different electrolyte concentrations.

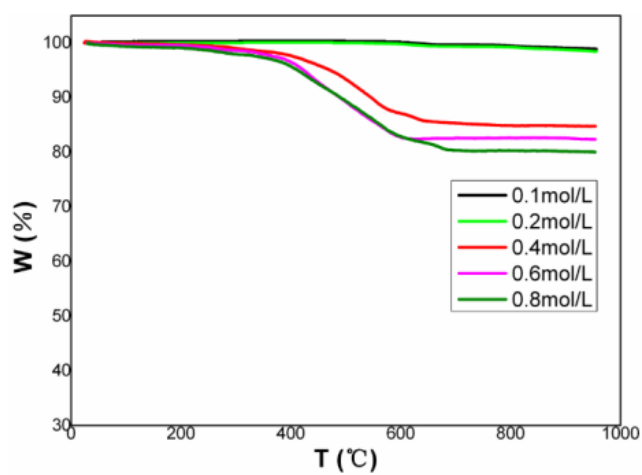


Figure 10. TGA of EGO under different electrolyte concentrations.

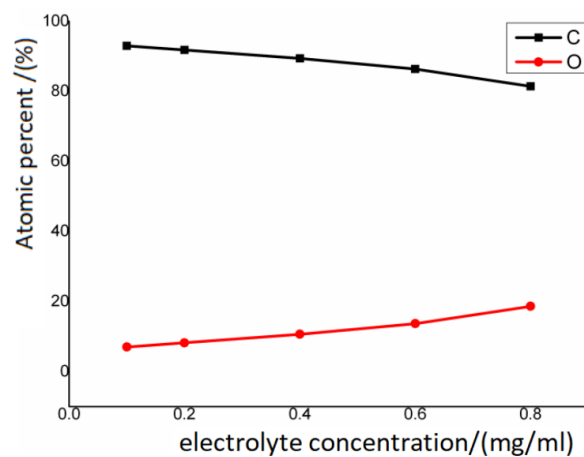


Figure 11. Carbon and oxygen element analysis of EGO under different electrolyte concentrations.

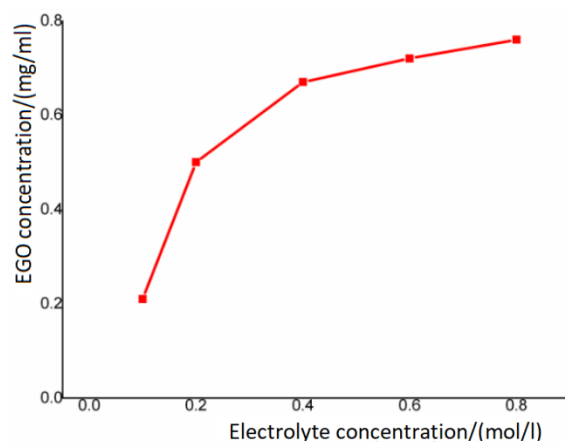


Figure 12. Concentration of EGO under different electrolyte concentrations.

Table 2. The ratio of peak intensity to peak height of EGO obtained under different electrolyte concentrations

	0.1 mol/L	0.2 mol/L	0.4 mol/L	0.6 mol/L	0.8 mol/L
D-peak intensity	386	506	986	1661	2322
G-peak intensity	1993	512	1446	1344	1988
2D-peak intensity	839	521	1817	349	366
I_D/T_G	0.19	0.98	0.73	1.23	1.16
I_G/T_{2D}	2.37	0.98	0.79	3.85	5.43

3.4. Study on EGO structure under optimal conditions

Under the optimal conditions (8 V, 0.4 mol/L NaOH), EGO was prepared, and the structural morphology was characterized.

(1) TEM (Figure 13)

EGO appeared as thin, loose sheets with enlarged interlayer spacing due to edge hydrophilic groups. TEM confirmed that EGO remained transparent and thin, with clear lattice fringes; SAED exhibited a six-fold symmetric diffraction pattern, confirming monolayer structure and intact crystallinity.

(2) AFM (Figure 14)

The thickness of EGO was 0.7–1.2 nm, falling within the typical range for monolayer oxidized graphene (considering group and water adsorption), confirming the monolayer nature.

(3) XPS comparison (Figure 15)

Before reduction, the C/O atomic ratio of EGO was significantly higher than that of GO, indicating lower oxidation. After reduction, the O 1s signal completely vanished for rEGO, whereas a residual O signal remained for rGO, demonstrating that the in-plane oxygen groups of GO are difficult to remove fully, while the edge groups of EGO are readily reducible. This further verifies the minimal damage and good reversibility of the electrochemical oxidation approach.

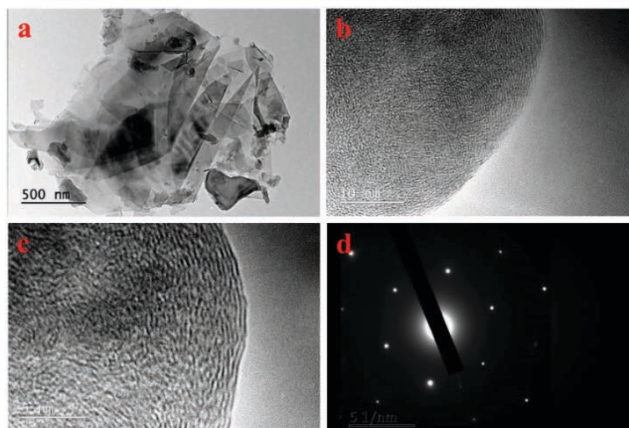


Figure 13. TEM image of EGO and its selected area electron diffraction pattern.

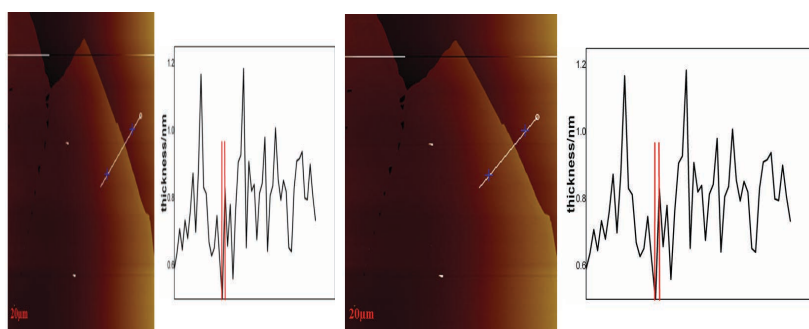


Figure 14. AFM image of EGO.

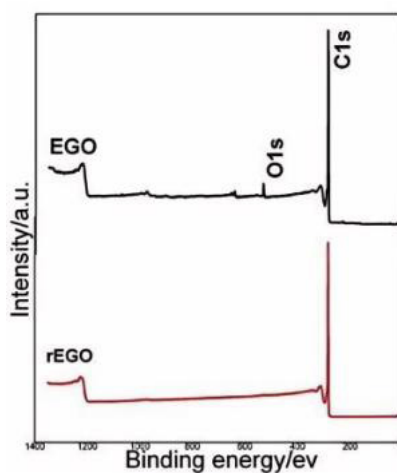


Figure 15. XPS spectra of EGO and rEGO.

4. Discussion

Collectively, the mechanism of electrochemical oxidation can be rationalized as follows: under the applied electric field, OH^- and reactive oxygen species in the aqueous solution preferentially attack the carbon atoms at the edges of graphene (due to their higher reactivity), generating carboxyl, hydroxyl, and other hydrophilic groups, while the in-plane hexagonal rings, being more energetically stable and kinetically hindered, are

less susceptible to oxidation. Moderate oxidation (8 V, 0.4 mol/L) introduces sufficient groups to provide hydration-layer electrostatic repulsion (increased Zeta potential) for stable dispersion, while avoiding in-plane defects. When voltage or electrolyte concentration exceeds the optimum, the oxidation intensifies, and groups extend into the basal plane, leading to lattice deterioration and reduced stability. Compared with the Hummers method, the electrochemical route avoids strong acids and oxidants, operates under mild conditions, and allows precise control of oxidation degree via electrical parameters, thus representing a green and controllable strategy.

The EGO dispersion prepared by this method achieves a concentration of 0.66 mg/mL, far exceeding that of ultrasonication-based methods (typically < 0.1 mg/mL), with excellent stability. More importantly, the preserved intact lattice ensures that the electrical conductivity and mechanical properties of graphene are retained, which is beneficial for subsequent grafting of phenyl-containing silane molecules via π - π stacking to improve compatibility with coating matrices, thereby providing high-quality feedstock for the fabrication of high-performance graphene/siloxane ceramic resin anticorrosion coatings.

5. Conclusion

- (1) Using large-size (ca. 7 μ m), monolayer, highly crystalline graphene powder prepared by a liquid-phase thermochemical route as the starting material, an electrochemical oxidation method was employed to selectively introduce hydrophilic oxygen-containing functional groups (hydroxyl, carboxyl) at the edges of graphene, successfully yielding a stably dispersed aqueous graphene dispersion and overcoming the obstacle of graphene agglomeration in water.
- (2) Through systematic optimization, the optimal preparation conditions were determined as: graphene sheet anode, copper cathode, 0.4 mol/L NaOH electrolyte, voltage 8 V, and treatment time 30 min. The resulting EGO dispersion reached a concentration of 0.66 mg/mL and remained stable without sedimentation for six months.
- (3) Multi-dimensional characterizations (Raman, TGA, elemental analysis, XPS) demonstrated that the oxidation degree of EGO is much lower, with an ID/IG ratio of only 0.42, oxygen species predominantly located at the edges, and an intact in-plane lattice. EGO preserves the intrinsic structure of graphene to the greatest extent while ensuring satisfactory dispersibility.
- (4) The electrochemical oxidation method is mild, controllable, and pollution-free. It offers a new route for the scalable application of graphene in waterborne coatings and composite materials.

Acknowledgments

Industrial Foundation Strengthening Project Shenzhen Key Industry R&D Plan

Funding

Industrial Foundation Strengthening Project Shenzhen Key Industry R&D Plan (Project No.: JSGG20200925145800001)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Geim A, Novoselov K, 2007, The Rise of Graphene. *Nature Materials*, 6(3): 183–191.
- [2] Guo X, Qing F, Li X, 2021, Applications of Graphene in Anti-Corrosion of Metal Surface. *Acta Physica Sinica*, 70(9): 098102.
- [3] Zhang Y, Sun J, Xiao X, et al., 2022, Graphene-Like Two-Dimensional Nanosheets-Based Anticorrosive Coatings: A Review. *Journal of Materials Science & Technology*, 129: 139–162.
- [4] Wang P, Cai D, 2020, Preparation of Graphene-Modified Anticorrosion Coating and Study on Its Corrosion Resistance Mechanism. *International Journal of Photoenergy*, 2020: 1–9.
- [5] Daradmare S, Raj S, Bhattacharyya A, et al., 2018, Factors Affecting Barrier Performance of Composite Anti-Corrosion Coatings Prepared by Using Electrochemically Exfoliated Few-Layer Graphene as Filler. *Composites Part B*, 155: 1–10.
- [6] Li T, Zhan H, Lan H, et al., 2023, Graphene Oxide Modification and Its Application in Anticorrosive Epoxy Coatings. *Journal of Chemical Engineering of Chinese Universities*, 37(4): 525–536.
- [7] Wang B, Cunnning B, Ruo R, et al., 2016, Graphene Coatings as Barrier Layers to Prevent the Water-Induced Corrosion of Silicate Glass. *ACS Nano*, 10(11): 9794–9800.
- [8] Yue X, Zhang X, Qiao Y, et al., 2025, A Review on Properties of Graphene Modified Thermosetting Resin Matrix Composites. *Materials Reports*, 39(9): 240401771–9.
- [9] Govindaraj P, Sokolova A, Salim N, et al., 2021, Distribution States of Graphene in Polymer Nanocomposites: A Review. *Composites Part B: Engineering*, 226: 109353.
- [10] Georgakilas V, Tiwari J, Kemp K, et al., 2016, Noncovalent Functionalization of Graphene and Graphene Oxide for Energy Materials, Biosensing, Catalytic, and Biomedical Applications. *Chemical Reviews*, 116(9): 5464–5519.
- [11] Qamar S, Ramzan N, Aleem W, 2024, Graphene Dispersion, Functionalization Techniques and Applications: A Review. *Synthetic Metals*, 307: 117697.
- [12] Pei S, Wei Q, Huang K, et al., 2018, Green Synthesis of Graphene Oxide by Seconds Timescale Water Electrolytic Oxidation. *Nature Communications*, 9: 145.
- [13] Yang S, Lohe M, Mullen K, et al., 2016, New-Generation Graphene From Electrochemical Approaches: Production and Applications. *Advanced Materials*, 28(29): 6213–6221.
- [14] Hao S, Hou S, Liu Y, et al., 2025, Enhanced Anticorrosion of Waterborne Epoxy Coatings by Electrochemical Exfoliated Graphene Oxide. *Progress in Organic Coatings*, 202: 109147.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Prediction of Grain Yield in Anhui Province Based on Multi-Source Remote Sensing Data

Introduction

Mengxin Ji, Haifeng Jiang*

School of Business, Anhui University of Technology, Ma'anshan 243032, Anhui, China

**Author to whom correspondence should be addressed.*

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Grain output forecasting is an important means to ensure national food security, and accurate output forecasts can provide scientific support for government decision-making. This paper takes Anhui Province as the research area, builds a Stacking ensemble learning model based on multi-source remote sensing data, and predicts grain output in Anhui. The study uses annual structured features (arable land area, sown area, electricity consumption, fertilizer usage, etc.) and monthly time series features (NDVI integral, NDVI mean, sunshine hours, etc.) to build a Stacking ensemble learning prediction model, integrating XGBoost, MLP, and RandomForest base models. XGBoost uses annual features to capture structured information at the annual scale, MLP uses monthly features to capture temporal variation patterns within the growing season, and RandomForest uses fusion features to capture interactions between features. By adaptively learning the weights of each base model through the Ridge regression meta-learner, effective integration of multi-source information is achieved. Experimental results show that the stacking ensemble model performs well under the retain-one cross-validation method: the determination coefficient reaches 0.9575, RMSE is 140.90 kg/ha, and average absolute percentage error MAPE is 4.08%. Compared to the XGBoost benchmark model that uses only annual features, the stacking model improves by 0.0039, and RMSE decreases by 6.21 kg/ha, validating the advantages of multi-source data fusion. Feature importance analysis showed that sown area and arable land area were the most critical factors affecting yield forecasting, together accounting for 60.4% of the characteristic importance. Meta-learner weight analysis shows that XGBoost is the main contributor (weight 1.0224), while MLP and RandomForest serve as auxiliary contributors (weights 0.0309 and 0.0230, respectively), providing dynamic growth season information that annual characteristics cannot capture. The research results of this paper provide a new method for grain output forecasting in Anhui Province, validating the application value of multi-source remote sensing data fusion in yield forecasting, which can offer scientific support for government decision-making and promote sustainable agricultural development.

Keywords: Grain production forecasting; Multi-source remote sensing data; Stacking integrated learning; XGBoost; MLP

Online publication: Sept 3, 2026

1. Introduction

Grain has always been the foundation of national survival and civilizational progress, and it is also a critical resource for maintaining social stability and driving economic growth. The stability and security of grain production are directly related to the high-quality development of the national economy and the comprehensive advancement of society. As one of the major grain-producing regions in China, the southern part of the country is characterized by complex and variable climatic conditions, considerable differences in soil types, a high degree of intensive agricultural production, and significant sensitivity to climatic disturbances. In this context, Anhui Province, as a major agricultural province, plays an important role in safeguarding national food security, regulating regional supply and demand, and stabilizing the national grain market. Fluctuations in grain yield therefore have profound implications for the national food security landscape and the sustainable development of agriculture.

Domestic and international scholars have conducted extensive research in the field of grain production and forecasting^[1], and most studies have found that, compared with single models^[2], ensemble models can improve predictive performance by leveraging the complementary advantages of different models^[3]. Jia Mengqi et al. combined ARIMA, GRNN, and LSTM models to forecast grain yield in Baoding City, Hebei Province, and found that the combination of ARIMA and GRNN achieved the best predictive performance^[4]. Shen Y et al. investigated the integration of LSTM with the random forest (RF) model and demonstrated that this hybrid model could better capture both time-series characteristics and nonlinear features^[5]. Wang X et al. further constructed a dual-branch deep learning model, using LSTM to process meteorological and remote sensing data on one branch and CNN to simulate soil characteristics on the other, confirming that the model accurately integrates different data sources while maintaining sufficient prediction accuracy^[6]. Zhao Yayue et al. proposed a combined forecasting scheme based on support vector regression (SVR) and a CNN-LSTM-attention mechanism for grain yield in Hebei Province^[7]. Ma Dianjing et al. used the Stacking ensemble algorithm to predict grain yield in southern China, providing an important reference for this study^[8]. Although the above studies have achieved favorable results in specific scenarios, as data become increasingly complex, how to leverage the advantages of multiple models further to optimize prediction remains a challenge that needs to be addressed in future research.

Subsequent scholars incorporated multi-source remote sensing data into their studies, steering grain yield prediction in a new direction^[9]. Basso et al. added remotely sensed spatial measurements to traditional indicators and achieved a prediction error of only 10%, demonstrating the suitability of remote sensing data for grain yield forecasting and providing an effective reference for subsequent researchers^[10]. Since then, the integration of remote sensing data and environmental features has been widely adopted in ensemble models, especially in processing soil and meteorological data^[11]. Panov et al. found that combining Sentinel-1B radar satellite-derived remote sensing parameters with ground observation data could improve the accuracy of cereal yield predictions^[12]. Building on this, Zhang Lixin used a GCN-DSTAMNet ensemble model with multi-source remote sensing data to predict grain yield, achieving promising results and providing important inspiration for the present study^[13].

Employing scientific models for predictive analysis can not only provide data support for policy-making but also promote the transformation of agricultural production toward sustainability and efficiency, thereby contributing to rural economic development and social stability, and fostering synergies among food security, ecological security, and economic growth^[14]. However, with the continued intensification of risks such as climate change, the global food security situation is becoming increasingly complex and urgent, and such

uncertainty poses a serious threat to the agricultural sector^[15]. Under these circumstances, using multi-source remote sensing data to predict future grain yield in Anhui Province can both promptly identify and correct unreasonable issues in local grain production and provide a reference for grain production decisions in other provinces and at the national level, thereby ensuring the safety and stability of the food supply.

Most existing studies have adopted feature modeling based solely on a single temporal scale, without exploiting the complementary information embedded in multi-scale data. To address this gap, this paper proposes a dual-scale feature fusion strategy that combines annual statistical features with monthly remote sensing features: annual data reflect the basic conditions of agricultural production, while monthly data capture the dynamic changes in crop growth, and the two types of data complement each other at the model level. In addition, current studies tend to use either a single machine learning model or a simple weighted averaging ensemble approach, which often fails to realize the advantages of different models fully. To this end, this study constructs a three-layer Stacking ensemble learning framework—the base model layer comprises three heterogeneous models, namely XGBoost, MLP, and RF, each handling different types of features; the meta-learner layer employs Ridge regression to adaptively learn the optimal weight combination, thereby integrating the strengths of all models. Grain yield prediction often faces the challenge of limited sample sizes, and most existing studies lack targeted feature engineering and evaluation strategies, which can easily lead to the curse of dimensionality and data leakage. In this paper, we extract growing-season features with clear agronomic significance to reduce dimensionality, adopt a combination of leave-one-out cross-validation and time-series split to prevent leakage of future information, and then screen the core features through analysis of variance and correlation analysis. This approach enhances the model's generalization ability and computational efficiency, enabling a more accurate analysis of future grain yield trends and providing a reference for research on grain yield prediction methods in Anhui Province.

2. Data sources and data preprocessing

2.1. Data sources

This study selected grain yield data of Anhui Province from 1989 to 2023, along with a set of relevant indicator data, which were compiled from the China Statistical Yearbook and the Anhui Statistical Yearbook. Meanwhile, using the Google Earth Engine (GEE) platform, we obtained the Normalized Difference Vegetation Index (NDVI) data from the MODIS Terra satellite and meteorological data within the boundaries of Anhui Province for the same time period.

2.2. Data preprocessing

First, the data were cleaned by removing invalid records and those with missing key fields, and the remote sensing data and yield data were temporally aligned to ensure consistency in their time ranges. Second, from the monthly remote sensing data, we extracted 16 statistical features for the growing season (April–September), including NDVI integral, NDVI peak, cumulative sunshine duration, etc., converting the monthly timeseries data into annuallevel features to align the temporal granularity with the annual data. Finally, since the features differ substantially in their scales (e.g., agricultural power generation is on the order of 10^5 , while agricultural electricity consumption is on the order of 10^2 , with a difference exceeding 1000fold), we applied Zscore standardization to all features. This ensured that each feature carried equal

weight in model training and prevented scale disparities from affecting model performance.

2.3. Feature selection

A twostep procedure was adopted to screen the seventeen annual indicators, aiming to eliminate multicollinearity and select the most predictive features.

2.3.1. Step 1: Collinearity removal

Pearson correlation coefficients were calculated among all features, with a correlation threshold set at 0.95. When the correlation between two features exceeded this threshold, the feature with the lower correlation was removed. A total of 21 highly correlated feature pairs were identified, leading to the removal of 8 redundant features and retaining 9 features.

2.3.2. Step 2: Random forest feature importance validation

A Random Forest model was used to evaluate the importance of each feature. The results showed that irrigated area and sown area were the two most important features, with importance values of 0.2810 and 0.2652, respectively. This is consistent with agricultural production knowledge—irrigation conditions and sown area are the most critical factors affecting grain yield^[16].

2.3.3. Final selection results

After the twostep screening, 8 annual key features were determined as model inputs, namely: irrigated area, sown area, agricultural electricity consumption, agricultural power generation, damaged area, pure nutrient equivalent of chemical fertilizer, affected area, and soil erosion area.

3. Model construction

3.1. XGBoost model

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on Gradient Boosting Decision Tree (GBDT), proposed by Chen and Guestrin in 2016. It iteratively trains multiple decision trees to minimize the loss function. Compared with traditional GBDT, XGBoost incorporates a regularization term into the loss function to control tree complexity and prevent overfitting, while also introducing secondorder derivative information, enabling faster convergence and higher accuracy.

The core idea of XGBoost is to train multiple decision trees sequentially, where each tree fits the residuals of the previous model, and the final prediction is obtained by the weighted sum of the predictions from all trees. This “stepwise approximation” strategy allows the model to continuously correct previous prediction errors, thereby achieving high predictive accuracy^[17].

The objective function of XGBoost consists of a loss function and a regularization term:

$$L(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k)$$

where $l(y_i, \hat{y}_i^{(t)})$ is the loss function of the t -th round, and $\Omega(f_k)$ is the regularization term of the k -th tree, which is used to control the complexity of the model:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

where T is the number of leaf nodes in the tree, w_j^2 is the weight of the j -th leaf node, γ controls the complexity of leaf nodes, and λ controls the degree of regularization of the weights. By adjusting these parameters, a balance can be achieved between the model's fitting ability and its generalization capability.

3.2. MLP neural network model

MLP (MultiLayer Perceptron) is a feedforward neural network consisting of an input layer, one or more hidden layers, and an output layer. Each neuron in this network is connected to neurons in the preceding layer via weights, and the input signals, after being processed by a nonlinear activation function, are passed to the next layer. MLP is capable of learning complex nonlinear relationships among features and is also applicable to data with temporal characteristics.

The forward propagation process of MLP is as follows:

$$z_1 = W_1 x + b_1$$

$$a_1 = \sigma(z_1)$$

$$z_2 = W_2 a_1 + b_2$$

$$a_2 = \sigma(z_2)$$

$$\hat{y} = W_3 a_2 + b_3$$

Among them, W_i and b_i are the weight matrix and bias vector of the i -th layer, respectively, and $\sigma(\cdot)$ denotes the activation function (ReLU is used in this paper). By stacking multiple layers of nonlinear transformations, the MLP is able to learn deep representations of the input features, thereby capturing complex nonlinear patterns.

The MLP is trained using the backpropagation algorithm, which employs gradient descent to minimize the mean squared error between the predicted values and the ground truth. During the training process, the error signal is propagated backward layer by layer from the output layer to the input layer, and the weight and bias parameters of each layer are updated accordingly, driving the model to approach the optimal solution gradually.

3.3. Random forest model

Random Forest, proposed by Breiman in 2001, is an ensemble learning algorithm based on decision trees. It trains a large number of decision trees in a parallel manner, where each tree is built using random sampling of both samples and features. The final prediction is produced by voting (for classification) or averaging (for regression). This "collective decision" mechanism reduces the variance of the model and thus improves its generalization ability.

The prediction process of Random Forest is as follows:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K f_k(x)$$

Among them, K represents the number of decision trees, and $f_k(x)$ denotes the predicted value of

the k -th tree. By averaging the prediction results from multiple trees, the random forest can counteract the fluctuations caused by the randomness of a single tree, thereby yielding more stable predictions.

3.4. Stacking ensemble framework

Stacking (stacked generalization) is a hierarchical ensemble learning method proposed by Wolpert in 1992, which employs a metalearner to learn from the prediction results of base models, thereby achieving adaptive fusion of different models. Unlike simple weighted averaging or voting strategies, Stacking automatically learns the optimal weights of each base model by training a metalearner, making better use of the complementary information across different models.

The fundamental idea of Stacking is to transform the model fusion problem into a new learning problem: using the prediction outputs of base models as input features and the original labels as the target, the metalearner is trained to learn how to combine these predictions optimally. This strategy adaptively adjusts the weights of each base model according to the data characteristics, without requiring manual specification.

3.5. Metalearner and ensemble strategy

In this study, Ridge regression is adopted as the metalearner, which learns the weights of the base models by minimizing the regularized squared error:

$$\min_w \sum_{i=1}^n \left(y_i - \sum_{m=1}^M w_m \hat{y}_{i,m} \right)^2 + \alpha \sum_{m=1}^M w_m^2$$

where w_m denotes the weight of the m base model, α is the regularization coefficient (set to 1.0 in this study), and M is the number of base models ($M=3$ in this study).

L2 regularization is introduced in Ridge regression to constrain the magnitude of the weights, preventing any single base model from being assigned an excessively large weight, thereby ensuring the stability of the ensemble model. Moreover, Ridge regression has a closedform solution, which allows for relatively fast training. See **Figure 1**.

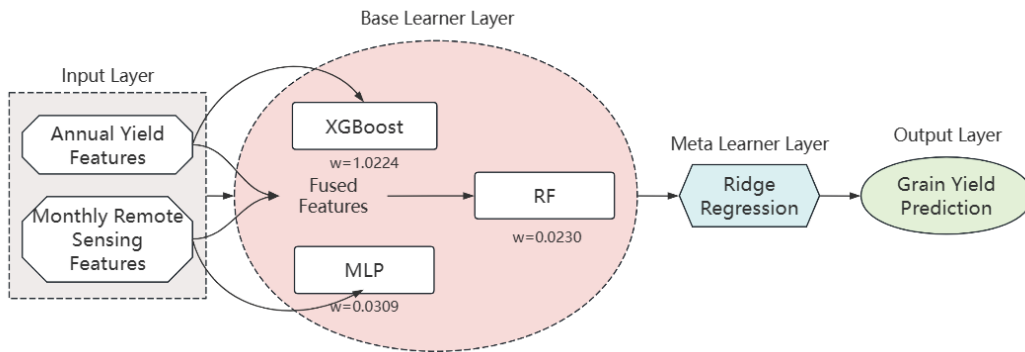


Figure 1. Model structure.

4. Model prediction

4.1. Experimental design

4.1.1. Data partitioning and validation strategy

In this study, leaveoneout crossvalidation (LOOCV) was adopted to evaluate model performance. Specifically, data from one year were selected as the test set, while the data from the remaining 34 years

were used as the training set; this procedure was repeated 35 times, covering all years from 1989 to 2023. This approach makes full use of the limited sample and avoids the fluctuations in results that may arise from random partitioning, thereby ensuring more reliable evaluation outcomes.

4.1.2. Evaluation metrics

Three metrics were selected to assess model performance:

(1) Coefficient of determination (R^2)

Reflects the proportion of variance explained by the model; the closer the value is to 1, the better the goodness of fit.

(2) Root mean square error (RMSE)

Represents the standard deviation of prediction errors, expressed in kg/ha.

(3) Mean absolute percentage error (MAPE)

Reflects the average level of relative error, presented as a percentage.

4.1.3. Baseline model settings

To examine the effectiveness of the Stacking ensemble model, three groups of baseline models were established:

(1) XGBoost baseline model: trained using only annual structured features.

(2) Direct feature fusion model: annual and monthly features were directly concatenated and then fed into XGBoost for training.

(3) Stacking ensemble model: integrates the prediction results from three base models, namely XGBoost, MLP, and RandomForest.

4.2. Model performance comparison

4.2.1. Overall performance evaluation

It can be seen from Table 1 that the Stacking ensemble model outperforms the other models in terms of both R^2 and RMSE:

- (1) The R^2 reaches 0.9575, which is an improvement of 0.0039 over the XGBoost baseline model;
- (2) The RMSE is 140.90 kg/ha, which is 6.21 lower than that of the XGBoost baseline model;
- (3) The MAPE is 4.08%, which is comparable to that of the XGBoost baseline model. This indicates that the Stacking ensemble strategy can effectively integrate multi-source information and improve prediction accuracy.

Table 1. Performance comparison of different models

Model	R^2	RMSE	MAPE
XGBoost baseline	0.9536	147.11	4.07%
Direct feature fusion model	0.9296	181.23	4.51%
Stacking ensemble model	0.9575	140.9	4.08%

4.2.2. Comparison between actual and predicted values

Figure 2 presents scatter plots of actual vs. predicted values for the Stacking ensemble model and the XGBoost baseline model. In the figure, the diagonal line represents perfect prediction, and the closer the

scatter points are to the diagonal, the higher the prediction accuracy.

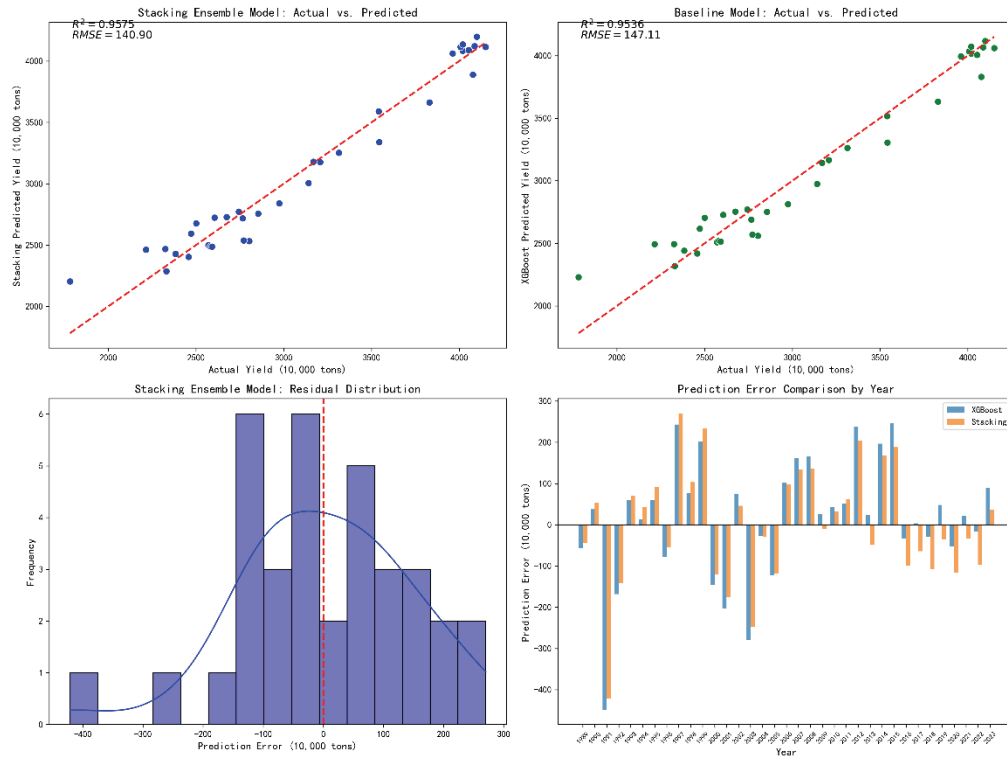


Figure 2. Comparison of model prediction results.

As can be seen from **Figure 2**:

- (1) The upper left subplot shows the prediction results of the Stacking ensemble model, with scatter points concentrated near the diagonal line, $R^2 = 0.9575$, and $RMSE = 1.4090$ million tons, indicating relatively high prediction accuracy;
- (2) The upper right subplot shows the XGBoost baseline model, with scatter points relatively more dispersed, $R^2 = 0.9536$, and $RMSE = 1.4711$ million tons, indicating slightly lower accuracy than the Stacking model;
- (3) The lower left subplot shows the residual distribution; the residuals of the Stacking model approximately follow a normal distribution with a mean close to zero, indicating little model bias;
- (4) The lower right subplot shows the error comparison; in most years, the absolute prediction error of the Stacking model is smaller than that of the XGBoost baseline model, especially in years with larger errors such as 1991 and 2003, where the Stacking model demonstrates a more pronounced effect in reducing errors.

4.2.3. Comparison of time series prediction

Figure 3 presents a comparison of the prediction curves generated by the three models:

- (1) In terms of overall trend, all three curves are able to track the variations in actual yield, particularly during the period of steady growth after 2012;
- (2) The Stacking model demonstrates a clear advantage, with its prediction curve more closely matching the

- actual yield curve, and it also captures yield fluctuations more accurately in years with large variations, such as 1991 and 2003;
- (3) The direct feature fusion model performs relatively poorly, with larger prediction errors and notably overestimated values in lowyield years, indicating that simply combining features does not improve model performance.

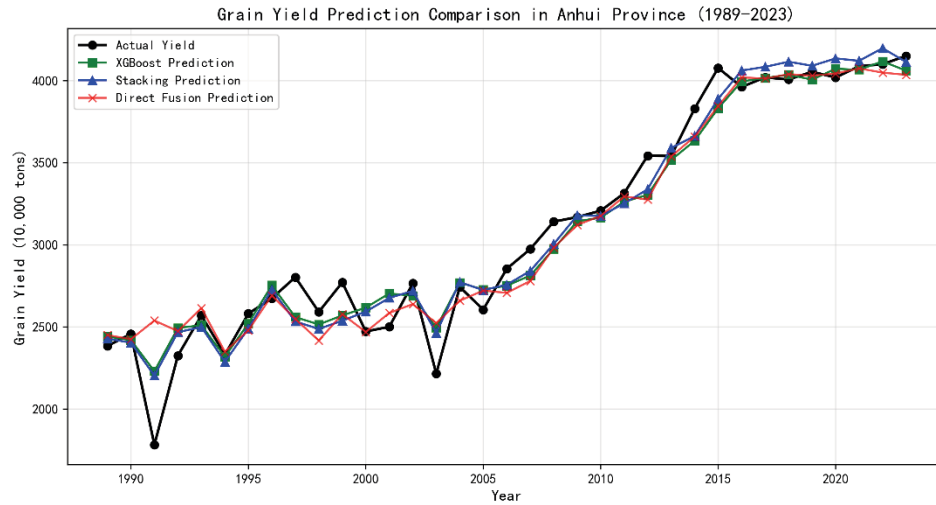


Figure 3. Comparison of timeseries forecasting of grain yield in Anhui Province (1989–2023).

4.2.4. P

Performance improvement analysis

Table 2. Performance improvement of the Stacking ensemble model over the baseline models

Indicator	Baseline model	Stacking model	Improvement
R^2	0.9536	0.9575	0.41%
RMSE	147.11	140.90	−4.22%
MAPE	4.07%	4.08%	0.25%

Although the improvements in R^2 and RMSE are not substantial, in the context of grain yield prediction with a limited sample size, even marginal performance gains still hold practical value. The RMSE decreased by 6.21, corresponding to an improvement in prediction accuracy of 4.22%. As Anhui Province is a major grainproducing region in China, the accuracy requirements for yield forecasting are inherently high, and these gains are therefore valuable in realworld applications.

4.2.5. Comparison with direct feature fusion

The direct feature fusion model performed even worse than the baseline model, with R^2 decreasing by 0.0240 and RMSE increasing by 34.12. This indicates that simply concatenating annual features and monthly features does not improve predictive performance; rather, it may lead to overfitting due to increased feature dimensionality and mutual interference among features. In contrast, the Stacking ensemble strategy, which automatically learns the weights of each base model through the metalearner, is better able to exploit the complementary information across different feature types.

5. Results and discussion

5.1. Value and significance of multisource data fusion

Grain yield prediction is a highly complex task involving multiple interacting factors from the natural, social, and economic spheres. Relying on a single data source is insufficient to comprehensively cover these factors, making multisource data fusion a key approach to improving prediction accuracy. In this study, by employing the Stacking ensemble learning framework to integrate annual structural features with monthly timeseries features, we validated the effectiveness of multisource data fusion in grain yield prediction.

In terms of model performance, the Stacking ensemble model that fused annual and monthly data achieved an R^2 of 0.9575, which is an improvement of 0.0039 over the XGBoost baseline model that used only annual features, and the RMSE decreased by 6.21. Although the improvement is modest, it remains practically meaningful for smallsample regression problems such as grain yield prediction. An RMSE reduction of 6.21 corresponds to a 4.22% improvement in prediction accuracy. For Anhui Province, as a major grainproducing region, this level of accuracy enhancement can provide more reliable decisionmaking support for government authorities.

With regard to feature contributions, annual features and monthly features play distinct roles in yield prediction. Among the annual features, sown area and cultivated area are the most critical predictors, together accounting for 60.4% of feature importance, indicating that the fundamental input factors of agricultural production exert a decisive influence on yield. Although the weights of monthly features such as NDVI and sunshine duration are relatively small (approximately 5%), they provide growthseason dynamic information that annual features cannot capture, and in certain years they notably improve prediction accuracy. This demonstrates that multisource data fusion not only enhances overall prediction accuracy but also captures subtle variations that are difficult to reflect using a single data source.

Regarding the ensemble strategy, the Stacking ensemble model effectively fuses multisource information by adaptively learning the weights of each base model through the metalearner. XGBoost, as the primary contributor (weight of 1.0224), captures annualscale structural information; MLP and RandomForest, as secondary contributors (weights of 0.0309 and 0.0230, respectively), capture timeseries variation patterns within the growing season and interactions among features. The weights of all three base models are positive, indicating that their predictions are complementary, thereby verifying the effectiveness of the Stacking ensemble strategy.

The significance of multisource data fusion in grain yield prediction lies not only in improving prediction accuracy but also in providing technical support for the development of intelligent agriculture. By integrating multisource data—including remote sensing, meteorology, soil, and crop information—and building intelligent agricultural decisionmaking systems, it is possible to achieve precise, automated, and intelligent agricultural production. This holds strategic importance for addressing climate change challenges, safeguarding national food security, and promoting sustainable agricultural development.

5.2. Practical applications and policy implications of the model

The grain yield prediction model constructed in this study has practical application value and can provide reference for government decisionmaking and sustainable agricultural development.

At the level of government decisionmaking, accurate yield predictions enable authorities to grasp the grain production situation in a timely manner and formulate appropriate grain policies. If the predictions indicate a potential decline in yield, the government can intervene proactively—for example, by increasing

agricultural subsidies, promoting highyield technologies, or strengthening disaster prevention and control—to stabilize grain production. The model itself can also identify key factors affecting yield, providing a reference for policy formulation. For instance, since sown area and cultivated area are core influencing factors, this suggests that governments should prioritize infrastructure construction such as farmland water conservancy and land consolidation, while ensuring that the quantity of cultivated land is maintained and its quality is improved.

For farmers, the feature importance results output by the model can help them understand which factors have the greatest impact on yield, allowing them to adjust their agricultural input structure accordingly and improve resource use efficiency. Fertilizer is one of the important factors affecting yield; based on the analytical results, farmers can apply fertilizer more rationally, promote precision fertilization techniques, and reduce waste and production costs.

In the field of precision agriculture, the application of remote sensing data plays a critical supporting role. By monitoring crop growth dynamics in real time, problems can be detected and addressed promptly, enhancing the precision of agricultural management. The NDVI feature reflects realtime crop growth status; when its value falls below the normal threshold for the corresponding period, farmers can implement timely field management measures such as irrigation and fertilization to help restore crop growth.

5.3. Research limitations and future directions

This study has certain limitations that need to be addressed in future work.

In terms of data, the limited sample size is a major shortcoming. This study used only 35 years of data from 1989 to 2023, and the relatively small sample size constrains the model's generalization ability to some extent. NDVI data for certain years also suffered from missing values or quality issues, which affected the accuracy of feature extraction. Future work could collect longer timeseries data to expand the sample and strengthen data preprocessing procedures to improve overall data quality.

With regard to the model, the relatively high complexity is another issue. The Stacking ensemble model, comprising three base models and one metalearner, has a deep architecture that entails increasing computational costs. The interpretability of the model is also somewhat limited, as its internal decisionmaking process remains difficult to understand intuitively. Future research could explore more concise ensemble strategies to reduce complexity and introduce interpretability tools such as SHAP values and LIME to deconstruct the decision logic and enhance model transparency quantitatively.

In terms of application, prediction errors of the model are notably larger in years with natural disasters, primarily because such extreme events account for a very small proportion of the training samples. Future work could supplement the dataset with historical disasteryear data to construct prediction modules tailored to extreme scenarios, thereby strengthening the model's ability to handle anomalous situations. In addition, this model was trained and validated exclusively on data from Anhui Province; direct transfer to other provinces or regions may not be directly applicable, and parameter recalibration and regional validation would be required.

In summary, grain yield prediction is a continuously evolving research field. With increasing data volumes, advances in modeling techniques, and changing application requirements, future research will move toward higher accuracy, stronger interpretability, and broader applicability, thereby making greater contributions to ensuring national food security.

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Gao X, Dong Y, Xu W, et al., 2025, Analysis and Prediction of Spatiotemporal Changes in Grain Production in Central Asia Based on ARIMA Model. *Journal of University of Chinese Academy of Sciences*, 42(4): 472–486. (in Chinese)
- [2] Zhang W, Sun D, Wang Y, et al., 2019, Prediction of Grain Yield in Liaoning Province Based on Support Vector Machine. *Economic Mathematics*, 36(1): 96–99. (in Chinese)
- [3] Dong M, Miao L, 2025, Research on Grain Yield Prediction Based on ARIMA-X-RFR Model. *Journal of Jiamusi University (Natural Science Edition)*, 43(11): 21–24. (in Chinese)
- [4] Jia M, Cai Z, Hu J, et al., 2021, Research on Grain Yield Prediction Model Based on Machine Learning. *Journal of Hebei Agricultural University*, 44(3): 103–108. (in Chinese)
- [5] Shen Y, Mercatoris B, Cao Z, et al., 2022, Improving Wheat Yield Prediction Accuracy Using LSTM-RF Framework Based on UAV Thermal Infrared and Multispectral Imagery. *Agriculture*, 12(6): 892.
- [6] Wang X, Huang J, Feng Q, et al., 2020, Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. *Remote Sensing*, 12(11): 1744.
- [7] Zhao Y, Fan L, Qin Z, et al., 2025, Analysis of Influencing Factors and Combined Prediction of Grain Yield Based on SVR and CNN-LSTM-ATTENTION Model. *Journal of the Chinese Cereals and Oils Association*: 1–12. (in Chinese)
- [8] Ma D, Zhao J, Yan W, et al., 2025, Grain Yield Prediction in Southern China Based on Stacking Ensemble Algorithm. *Hubei Agricultural Sciences*, 64(5): 155–159 + 184. (in Chinese)
- [9] Zou Y, 2024, Research on Maize Yield Prediction in China Based on Machine Learning, thesis, Nanjing University of Information Science and Technology (in Chinese)
- [10] Basso B, Ritchie J, Pierce F, et al., 2011, Spatial Validation of Crop Models for Precision Agriculture. *Agricultural Systems*, 68: 97–112.
- [11] Liu J, He X, Wang P, et al., 2019, Early Prediction of Winter Wheat Yield Using Long Time Series Meteorological Data Combined with Random Forest Method. *Transactions of the Chinese Society of Agricultural Engineering*, 35(6): 166–174. (in Chinese)
- [12] Panov D, Sakharova E, 2022, Using Radar Data for Grain Crops Yield Forecasting in the Novosibirsk Region. *Russian Meteorology and Hydrology*, 47(6): 473–478.
- [13] Zhang L, 2025, Research on Grain Yield Prediction Model Based on Deep Learning, thesis, Changchun University of Technology (in Chinese)
- [14] Jeong J, Resop J, Mueller N, et al., 2016, Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, 11(6): 56–57.
- [15] Brown J, Hochman Z, Holzworth D, et al., 2018, Seasonal Climate Forecasts Provide More Definitive and Accurate Crop Yield Predictions. *Agricultural and Forest Meteorology*, 260: 247–254.
- [16] Chen C, Lin Z, Sun X, 2007, Dynamic Prediction and Suggestions for Grain Yield in Shandong Province. *Journal of Anhui Agricultural Sciences*, 35(28): 8773–8775. (in Chinese)
- [17] Tian Q, Li J, Ding Y, et al., 2025, Machine Learning-Driven Grain Yield Prediction and Influencing Factor Analysis in Heilongjiang Province. *Grain and Feed Science and Technology*, 2025(2): 13–15. (in Chinese)

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Analysis of the Problems Existing in Electronic Product Testing and Corresponding Countermeasures

Haibo Liang

Fanguang Inspection & Testing Co., Ltd., Guangzhou 511450, Guangdong, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Electronic product testing serves as the core link to guarantee product quality, safe use, and stable performance. However, the traditional testing mode has defects in testing procedures, technical equipment, personnel capabilities, and management systems, which cannot satisfy practical production demands. Enterprises need to carry out optimization measures for every testing link to improve testing accuracy and the detection rate of hidden faults, so as to guarantee product quality better. Based on the practical problems in electronic product testing, this paper puts forward feasible strategies including constructing standardized testing procedures, accelerating the upgrading of technical equipment, strengthening professional talent training, and perfecting closed-loop management and data mechanisms, aiming to provide references for the standardized development of electronic product testing work.

Keywords: Electronic products; Testing; Problems; Countermeasures

Online publication: Sept 3, 2026

1. Introduction

Product testing runs through the whole chain of electronic products, covering research and development, production and factory delivery, and largely determines product quality and service safety. In the process of pursuing market competitiveness, enterprises must attach great importance to the optimization and upgrading of electronic product testing technologies, make up for deficiencies in all testing links, and fundamentally improve the reliability and accuracy of testing results. Nevertheless, there is a gap between actual electronic product testing and practical demands in real operation, which fails to adapt to the research and production scenarios of high-end electronic products. At present, optimizing electronic product testing strategies guided by practical problems has become an important research topic for enterprises.

2. The important significance of electronic product testing

2.1. Strictly control product quality

Electronic product testing acts as an important checkpoint for factory quality control. From the sales perspective, a complete and scientific electronic product testing system can systematically screen quality defects such as substandard performance, functional failures, insufficient stability, and potential safety hazards before products enter the market, thus preventing unqualified products from flowing to consumers. From the production perspective, accurate and comprehensive testing results can locate process defects and fluctuations in component quality, and guide enterprises to optimize product design and production procedures ^[1].

2.2. Support iterative product research and development

Relevant testing data can reflect the strengths and weaknesses of existing products in performance, power consumption, compatibility, and other dimensions, and clarify the direction of product iteration and upgrading. For instance, testing results of new product prototypes can intuitively reveal logical loopholes and scenario adaptation problems in research and design ^[2]. If enterprises discover defects at the prototype stage, they can avoid massive waste of early R&D investment caused by core flaws found after product finalization. Besides, testing results of mass production can reflect unreasonable design schemes, and provide data support for minor revisions and functional upgrades, helping enterprises continuously optimize products and enhance user experience ^[3].

2.3. Comply with industrial supervision requirements

At present, all industries have issued clear supervision standards for quality, safety, and performance indicators. Manufacturers are required to complete compliance tests on specified items and issue corresponding test reports before products go on sale to obtain market access qualifications. In addition, exported electronic products need to meet special technical specifications and testing requirements formulated by destination countries; otherwise, they cannot pass customs inspection. Therefore, standardized product testing is a necessary prerequisite for electronic products to enter the market. Enterprises must attach importance to electronic product testing to avoid compliance risks and ensure stable production and operation.

3. Main problems in electronic product testing

3.1. Incomplete and non-standard testing procedures

To shorten production cycles and cut costs, some enterprises simplify testing processes by removing or shortening key links such as aging tests and extreme environment long-term stability tests, making it impossible to verify product reliability under extreme conditions fully. Besides, inconsistent implementation of testing standards also leads to non-standard procedures ^[4]. Some enterprises fail to formulate unified testing processes for different product types, leading to inconsistent operating standards among batches and inspectors. Such randomness in procedures results in undetected hidden faults, and products frequently crash, lag, or suffer battery attenuation after delivery, which seriously damages user experience and brand reputation.

3.2. Backward testing technologies and equipment

With the integration of smart chips, Internet of Things, and artificial intelligence into electronic products, products become increasingly complex with integrated and intelligent functions. While bringing better user experience, this trend raises higher standards for testing precision, efficiency, and comprehensiveness. In fact, many enterprises still rely on manual testing without timely procurement of automatic and intelligent testing equipment. Backward testing technologies and equipment can only complete basic function verification, failing to cover high-precision performance tests, scenario simulation tests, and multi-device compatibility tests. Thus, testing results cannot fully and objectively reflect product quality ^[5]. Especially for cutting-edge products such as 5G communication modules, AI chips, and flexible screens, hidden performance defects are harder to detect, and traditional testing means can hardly meet practical demands.

3.3. Uneven professional competence of testing teams

Electronic product testing requires inspectors to master circuit principles, software systems, testing technologies, and operation methods of various professional equipment. However, many testing staff in enterprises lack systematic professional training. They only grasp basic operations and have insufficient understanding of testing key points and fault mechanisms of new electronic products ^[6]. This leads to irregular testing of complex electronic products and failure to identify potential performance defects accurately. Moreover, most enterprises do not attach importance to competency development of testing teams or establish regular training systems, which restricts inspectors' learning of new technologies, equipment, and standards.

3.4. Imperfect testing management system

The defects of the testing management system are reflected in three aspects. First, test data records are non-standard and incomplete, which hinders review and analysis and makes it hard to find omissions in testing links. Second, unclear division of testing responsibilities means specific persons cannot be traced after testing mistakes, so quality control requirements cannot be effectively implemented. Third, there is no full-life-cycle traceability mechanism for testing, resulting in isolated test data from R&D, trial production, and mass production. It hinders data collection and overall analysis, making it difficult to find common quality problems in different production stages and failing to provide sufficient data support for design optimization and process adjustment. Without effective solutions to these problems, a closed-loop mechanism of "testing-feedback-correction-optimization" cannot be formed, resulting in massive waste of data resources.

4. Optimization countermeasures for problems in electronic product testing

4.1. Construct standardized testing procedures

Standardized testing procedures serve as the foundation for resolving electronic product testing challenges, improving the production quality of electronic goods, and consolidating enterprises' high-quality development. Enterprises must face up to deficiencies in their existing testing workflows and establish standardized testing procedures covering the entire product lifecycle. For instance, during the product research and development phase, specific operational steps and judgment criteria for functional verification, performance benchmarking, and compatibility testing shall be defined in line with product functional designs and performance parameter requirements, so as to accurately identify product defects at the design

stage and offer guidance for product design optimization ^[7]. Upon entering the pilot and mass production phases, enterprises shall adopt a hierarchical testing control model, specifying requirements for basic functional inspections, sampling aging tests, and extreme environment simulation tests for each batch of products to guarantee consistent testing procedures and standards across all batches. In accordance with the characteristics of each product category, testing contents and operational specifications shall be stipulated to compile standardized technical manuals, mandating all testing personnel to conduct tests strictly in compliance with the manuals. Quality control requirements for each testing phase shall be clarified, and testing review mechanisms shall be implemented at every node to prevent issues such as shortened testing durations and omission of critical testing steps. This set of testing procedures defines key testing points and operational standards for distinct production stages, minimizes testing oversights from a procedural perspective, and further enhances the reliability of test results ^[8].

4.2. Accelerate the upgrading of testing technologies and equipment

Enterprises should increase investment in electronic product testing, timely introduce new testing technologies and equipment, and advance the intelligent transformation of testing methods. First of all, enterprises need to introduce high-precision testing instruments and intelligent testing systems tailored to their product testing requirements, and promptly replace outdated equipment that has exceeded its service life and fails to meet precision standards, so as to ensure testing facilities are compatible with integrated and intelligent electronic products. For example, enterprises can deploy scenario-based simulation testing systems, leveraging their technical strengths to improve the comprehensiveness and accuracy of test results for high-end electronic products and guarantee production quality. Such systems can simulate a wide range of complex application scenarios for electronic products and conduct special testing items including multi-device linkage compatibility tests and extreme environmental adaptability tests, enabling more accurate identification of latent performance defects in products. Secondly, enterprises shall accelerate the application and promotion of automated testing technologies, and develop automated testing scripts for routine basic testing procedures to realize automated batch testing of products ^[9,10]. On one hand, this eliminates errors caused by manual operation and improves the consistency and precision of test results; on the other hand, it boosts testing efficiency, shortens testing cycles, and aligns product testing technologies with the demands of mass production. Finally, enterprises shall attach importance to independent innovation of testing technologies, encourage technical staff to develop customized testing schemes based on the testing demands of core products, and strengthen the compatibility between testing technologies and their own product systems ^[11].

4.3. Promote training and team building for testing talents

To ensure the smooth progress of all electronic product testing work, enterprises should attach great importance to the training of testing personnel and team building, and continuously increase efforts in training and recruiting testing staff. For enterprises, training testing personnel constitutes internal self-sustained talent development. Enterprises shall establish regular training mechanisms based on the capability deficiencies of existing testing staff. The training content mainly covers testing standards for new electronic products, fault analysis methods, and operating specifications for new testing equipment; training methods include offline centralized lectures, online special courses, and on-the-job practical training. In practice, enterprises may select customized training content and approaches for testing personnel at different

competency levels to help them timely update their knowledge systems, master cutting-edge testing skills, and build a competency framework matching current product testing requirements. Meanwhile, enterprises shall actively recruit high-level testing technical professionals to form specialized testing teams and elevate the overall technical capacity of testing teams. Such external talent introduction can help enterprises rapidly recruit professional specialists for high-end electronic product testing ^[12]. By hiring high-level testing technical talents, enterprises can directly upgrade the team's technical capabilities, and these professionals can drive the competency improvement of internal testing staff through knowledge sharing and mentoring.

4.4. Improve testing management and data closed-loop mechanisms

Perfect testing management and data closed-loop mechanisms lay the foundation for traceable test data and locatable problems. Enterprises shall formulate complete testing management systems, clarify powers and responsibilities of all testing links, and require staff to standardize testing and record data and abnormal conditions ^[13]. The analyzed data can identify defects in testing and production processes and provide a reliable basis for optimizing testing procedures and product design. On this basis, build a full-life-cycle test data management platform to break data barriers among R&D, trial production, and mass production, and realize real-time sharing and centralized storage of test data ^[14]. The platform forms a complete chain of “data collection-data summary and analysis-problem positioning-feedback-implementation of optimization-verification”, supporting continuous iteration of testing technologies. Enterprises can improve closed-loop management of testing to guarantee reliable and comprehensive test results, feed data back to R&D departments, and realize a circulation of “testing-analysis-improvement-verification”, thus lifting testing efficiency and product quality ^[15].

5. Conclusion

Product testing is a vital link to guarantee electronic product quality, and its technical level and working efficiency determine enterprise competitiveness and consumer rights. Enterprises shall address defects including non-standard procedures, backward technologies, insufficient talent development and weak management, and perfect testing systems from four dimensions: institutional construction, technical upgrading, talent cultivation and management optimization. Coordinated progress of these dimensions can standardize and intellectualize testing work, solve various problems in testing links, and bring important benefits to enterprises and the whole industry. Enterprises shall keep track of technical trends in electronic testing, and dynamically optimize testing management systems according to product iteration directions, so as to guarantee product quality and sustain the lifeline of corporate and industrial development.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Tian F, Chen Y, Wang C, 2025, Analysis of the Impact of 5G Communication Technology on Power Consumption and Energy Efficiency of Electronic Products. *Daily Electrical Appliances*, 67(11): 75–79 + 84.

- [2] Yin X, Gu Y, 2025, Research on Electronic Product Fault Diagnosis Technology Based on Artificial Intelligence. *Computer Knowledge and Technology*, 21(32): 33–35.
- [3] Li W, Guo M, Li L, et al., 2025, Open Circuit Failure Analysis of Electronic Products Based on Time Domain Reflectometry. *Printed Circuit Information*, 33(11): 34–38.
- [4] Li S, 2025, Design and Implementation of Automatic Electromagnetic Interference Test System for Electronic Products. *Electronic Production*, 33(18): 90–93.
- [5] Lin J, 2025, Research on Enterprise Electronic Product Testing Management System. *Product Reliability Report*, 2025(8): 147–149.
- [6] Yang Q, 2025, Research on the Influence of Hydrazine Hydrate on Halogen Analysis and Testing of Electronic and Electrical Products. *Laboratory Testing*, 3(13): 233–235.
- [7] Cheng J, Dai S, Gong Q, 2025, Architecture Design and Integrated Application of Intelligent Production Line for Aviation Electronic Assembly Testing. *Avionics Technology*, 56(2): 65–73.
- [8] Li S, Wang W, Li Q, 2025, Research on Batch Screening Method and Test System Design for Circuit Boards. *Electronic Production*, 33(5): 16–19.
- [9] Jiang W, 2025, Application of Electromagnetic Compatibility Testing Technology in the Development of Electronic and Electrical Products. *Modern Industrial Economy and Informationization*, 15(2): 116–119.
- [10] Wang J, 2024, Test Analysis of Protective Conductor Current and Contact Current of Switching Power Supplies. *Safety & EMC*, 2024(6): 72–76 + 83.
- [11] Du S, Wang Y, Wang S, 2024, Research on Key Technologies and Instrument Application of High-Frequency Electronic Product Performance Testing. *Instrumentation Users*, 31(11): 66–67 + 70.
- [12] Peng L, 2024, Optimization Strategies for Electronic Component Reliability Evaluation and Quality Control. *Product Reliability Report*, 2024(10): 41–42.
- [13] Xu X, 2023, Discussion on Design Ideas of Testing Methods for Internet of Things Electronic and Electrical Products. *Home Appliance Maintenance*, 2023(10): 26–28.
- [14] Li X, Li B, 2023, Design and Implementation of Handling Structure for Electronic Product Testing. *Modern Manufacturing Technology and Equipment*, 59(5): 98–101.
- [15] Wang Z, 2023, Discussion on Application of Non-Vector Testing Technology in Aviation Electronic Product Testing. *Southern Agricultural Machinery*, 54(6): 143–145.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on Medium- and Long-Term Wind Power Forecasting Technology Based on Ensemble Forecasting

Li Liu, Xiaolin Xu, Qian Liang, Jiayu Chen, Sen Duan

State Power Investment Corporation Energy Science and Technology Research Institute, Shanghai 201109, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: With the continuous growth of grid-connected wind power capacity and the full participation of new energy in power market transactions, medium and long-term power forecasting has become a key supporting technology to guarantee the safe and economic operation of power systems. The forecasting ability of deterministic numerical weather prediction (NWP) decays sharply for lead times longer than 4 days. Ensemble forecasting provides a probabilistic information foundation to break this bottleneck, yet its application at wind farm scale is confronted with critical problems including coarse spatial resolution, prominent systematic bias, and low utilization rate of multi-member information. Taking Yuanqu Wind Farm in Shanxi Province as the research object, this paper systematically carries out research on multi-source ensemble forecast data adaptation processing, hybrid forecasting modeling driven by random forest and physical models, as well as probabilistic forecasting methods focusing on medium- and long-term wind power forecasting based on ensemble forecasts. In terms of data adaptation processing, four global ensemble forecasting systems including ECMWF-ENS, UKMO-EGRR, CMA-BABJ and NCEP-GEFS are compared and evaluated for their applicability in complex mountain wind farms. The results show that ECMWF-ENS has the lowest forecasting error and optimal seasonal stability. On this basis, a complete processing workflow consisting of spatial downscaling, quantile mapping bias correction, and ensemble spread consistency reconstruction is established, which reduces the root mean square error (RMSE) of wind speed forecasting by 15–25%. In terms of forecasting modeling, a hybrid driving framework combining random forest wind speed correction and equivalent power curve physical conversion is proposed. The random forest is adopted to learn local nonlinear residuals, while physical models provide hard constraints for energy conversion. The model is trained with data from January 2024 to February 2025 and verified in the independent test period from January to April 2026. The power generation accuracy of ensemble mean point forecasting reaches 83.26%, with an RMSE of 42.73 MW, meeting the requirements of power industry standards. This study constructs a complete technical chain from raw ensemble forecast data to wind farm-level probabilistic power forecasting, and verifies the effectiveness and engineering practicability of the hybrid driving method for medium and long lead times.

Keywords: Medium and long-term wind power forecasting; Ensemble forecasting; Random forest; Physical-data hybrid driving

Online publication: Sept 4, 2026

1. Introduction

Driven by global energy transition and carbon neutrality goals, wind power and photovoltaic power are accelerating to become the mainstay of power supply^[1]. By the end of 2025, China's installed wind power capacity reached 640 million kilowatts, and the cumulative installed proportion of wind and photovoltaic power hit 47.3%, exceeding thermal power for the first time in history. The strong randomness, intermittency, and fluctuation of high-proportion new energy transform power systems from the traditional "load-following generation" mode to the collaborative interaction mode of "source-grid-load-storage", putting forward higher requirements for the lead time and accuracy of power forecasting^[2,3]. Meanwhile, the Circular of the National Development and Reform Commission on Deepening the Market-oriented Reform of New Energy On-grid Electricity Prices and Promoting High-quality Development of New Energy (Document No. FGJG (2025) 136) clearly promotes the full access of new energy to the power market, and power transactions urgently require the support of high-precision meteorology and power forecasting technologies.

Medium and long-term power forecasting (4–30 days in advance) serves as core information supporting power balance, maintenance scheduling, cross-provincial and cross-regional transactions, and reserve optimization^[4–6]. At present, forecasting technologies relying on deterministic numerical weather prediction are only effective within 0–3 days. After the 4th day, the nonlinear chaotic effect of the atmosphere leads to a sharp rise in errors of single deterministic forecasts^[7–10]. By constructing multiple groups of initial value and physical perturbation members to describe the probability density distribution of forecasting states, ensemble forecasting acts as a fundamental approach to break the lead time bottleneck. Taking the ECMWF global ensemble forecasting system as an example, it provides 51-member forecasts for 0–15 days and subseasonal forecasts extended to 46 days, offering abundant probabilistic information for wind and solar power forecasting for 4–30 days^[11,12].

Nevertheless, ensemble forecasting has relatively coarse spatial resolution (about 18–36 km), resulting in scale mismatch with actual meteorological exposure of wind farms; each member suffers from systematic bias and inaccurate ensemble spread^[13]. Therefore, research on multi-timescale wind power forecasting based on ensemble forecasting is of great significance to break the bottleneck of medium and long-term new power cognition for power grids and improve the safety and economy of high-proportion new energy systems.

At present, ECMWF Ensemble Forecast (ENS), ECMWF Subseasonal Forecast, and the US NCEP Global Ensemble Forecast System (GEFS) are widely applied in renewable energy forecasting businesses of many countries. Relevant studies generally prove that ensemble mean forecasting can effectively smooth random errors caused by internal atmospheric variability on the 4–10 day scale and outperform single deterministic forecasts. However, the RMSE of ensemble mean forecasting still increases significantly for lead times over 10 days. Domestic scholars have carried out preliminary adaptive research on ensemble forecasting in medium- and long-term wind/photovoltaic forecasting targeting China's complex terrain and climate characteristics, yet obvious deficiencies exist in refined preprocessing at wind farm scale and deep coupling with power models^[14–17].

2. Technical route

Aiming at the above problems and challenges, this project centers on the main line of "data foundation–forecasting core–decision support" and sets three progressive research contents:

First, characteristic analysis and adaptive processing of multi-source ensemble forecast data. This step systematically analyzes the uncertainty features of typical ensemble forecast systems, investigates spatial downscaling methods at the wind farm station scale, and develops quantile mapping bias correction as well as ensemble spread consistency reconstruction technologies. Second, construction of wind power prediction models based on ensemble forecasts. Targeting two time scales of 4–15 days and 15–30 days, a hybrid driving model combining random forest-corrected wind speed and physical power curves is established. Third, system integration and engineering verification. The above methods are integrated into a station-level prediction system, whose prediction accuracy and practical adaptability are verified using actual operational data.

This study takes the Yuanqu Wind Farm in Shanxi Province as a typical research object. The prediction horizons are set as 10 days and 20 days with a 15-minute temporal resolution, plus a 30-day horizon adopting daily average power, so as to meet the demands of medium and long-term power trading. The training and validation period spans from January 2024 to December 2025, and the independent testing period covers January to April 2026, with actual grid-connected power from SCADA data adopted as the verification benchmark.

3. Characteristic analysis and adaptive processing of multi-source ensemble forecast data

The performance of medium and long-term wind power forecasting highly depends on the quality and adaptability of input meteorological forecast data. Starting with characteristic analysis of multi-source ensemble forecast data, this chapter compares and verifies the superiority of ECMWF as the core data source. Targeting the demands of wind power generation, it systematically carries out meteorological element extraction and quality control, spatial downscaling and site adaptation, multi-timescale feature construction, as well as bias correction and consistency reconstruction of ensemble members, laying a solid data foundation for subsequent forecasting modeling.

3.1. Characteristic analysis of multi-source ensemble forecast data

Based on wind speed forecast verification data of Yuanqu Station in Shanxi Province from January to November 2024, comprehensive analysis is conducted on the data quality of four global ensemble forecasting systems, including ECMWF-ENS (ECMF), UKMO-EGRR (EGRR), CMA-BABJ (BABJ), and NCEP-GEFS (KWBC). Judging from monthly overall statistical indicators RMSE and MAE, ECMWF always achieves the optimal performance, with monthly average RMSE fluctuating between 0.90 m/s (April) and 1.24 m/s (March), and corresponding MAE ranging from 0.78 m/s (October) to 0.90 m/s (March), featuring the lowest error and favorable seasonal stability. EGRR ranks second, with monthly RMSE from 1.09 m/s (April) to 1.40 m/s (March) and MAE from 0.83 m/s (October) to 1.05 m/s, also showing high forecasting reliability. BABJ has significantly higher errors than the former two, with monthly RMSE between 1.32 m/s (October) and 1.74 m/s (March) and MAE from 1.04 m/s (October) to 1.47 m/s, accompanied by drastic seasonal fluctuations, which indicates systematic deficiencies in physical processes or initial field processing of the model under complex terrain. KWBC delivers the worst performance in all months, with the minimum monthly RMSE of 1.46 m/s (April) and maximum of 2.31 m/s (January), and MAE ranging from 1.46 m/s to 1.79 m/s. Its errors rise sharply in high-winter periods, revealing severe inadaptability to mountain wind farm forecasting.

From the perspective of time series variation, errors of all systems are higher in winter (January, March), decrease in spring (April, May), stay low in summer (July, August), and rise slightly in autumn (September–November). This rule is related to strong northwest winds and atmospheric stability changes in the Yuanqu area, and also reflects the greater forecasting difficulty of numerical models for high wind speed events. It is noteworthy that ECMF and EGRR maintain a low error growth slope in all months throughout the year, while errors of BABJ and KWBC accelerate with the extension of lead time, especially KWBC. Based on 11-month data evaluation, ECMF and EGRR can be adopted as core input sources for wind power forecasting systems; BABJ requires bias correction before application, and KWBC cannot be directly used for wind speed forecasting and must undergo intensive correction or downscaling processing.

3.2. Bias correction and consistency reconstruction of ensemble members

According to wind speed forecast verification results of Yuanqu Station from January to November 2024, different global ensemble forecasting systems exhibit prominent systematic bias and error growth characteristics. Among all systems, ECMWF-ENS (ECMF) has the minimum RMSE and MAE across all lead times and months with optimal seasonal stability, so subsequent wind power forecasting research mainly relies on ECMF data. Nevertheless, raw ECMF outputs are global grid fields with a spatial resolution of about 9 km, which cannot directly characterize local wind features of mountainous terrain where Yuanqu Station is located. Even after spatial interpolation, accumulated systematic biases still exist with the extension of lead times. To improve the quantitative accuracy and probabilistic reliability of ECMF ensemble forecasting for local wind power prediction, spatial downscaling and site matching, systematic bias correction and ensemble spread consistency reconstruction must be implemented in sequence. This section follows the technical route of “from space to numerical value, from mean value to distribution”.

ECMF ensemble forecasting contains multiple perturbation members, and its theoretical spread should reflect forecasting uncertainty. Verification shows that the ensemble spread of ECMF at Yuanqu Station is often smaller than actual errors (under-dispersion), especially for short lead times. This leads to excessively narrow confidence intervals of probabilistic forecasting and reduces the reliability of ensemble forecasting. Therefore, a method combining variance inflation and member reordering is adopted for consistency reconstruction. It is expected that after spatial downscaling and statistical correction, the RMSE of ECMWF wind speed forecasting can be reduced by 15–25%, and the reliability of probabilistic forecasting will be significantly improved. Corrected ECMF ensemble members will be taken as inputs for subsequent wind power forecasting models to enhance the accuracy of wind power curve conversion and the confidence level of probabilistic forecasting.

In summary, spatial downscaling, site matching, bias correction, and consistency reconstruction implemented on ECMF can acquire more accurate and reliable wind speed probabilistic forecasts, providing high-quality inputs for the wind power forecasting model in the next chapter.

4. Physical-data hybrid driven wind power forecasting model based on ensemble forecasting

4.1. Overview

Aiming at the problems of fast error accumulation of single physical models and poor physical consistency of pure data models, a hybrid driving architecture integrating random forest wind speed correction and physical

power curve conversion is proposed. The random forest learns local nonlinear residuals of wind speed forecasts, while physical models provide hard constraints for energy conversion, achieving complementary advantages of the two methods.

4.2. Physical modeling and baseline construction of wind power

4.2.1. Physical model of wind turbine power curve

The theoretical relationship between wind turbine output power and hub-height wind speed is determined by aerodynamics. For wind speeds below rated speed, power is approximately cubic to wind speed:

$$P_{ideal}(v) = \begin{cases} 0 & v < v_{cut-in} \\ \frac{1}{2} \rho A C_p(\lambda, \beta) v^3 & v_{cut-in} \leq v < v_{rated} \\ P_{rated} & v_{rated} \leq v < v_{cut-out} \\ 0 & v \geq v_{cut-out} \end{cases}$$

Where: ρ stands for air density (kg/m^3), A is the swept area of the rotor (m^2), C_p is the wind power utilization coefficient (theoretical Betz limit 0.593, actual value 0.4–0.5), λ is the tip speed ratio, and β is the pitch angle. In actual operation, measured power curves deviate from theoretical curves due to turbulence, wake effect, blade fouling, and control strategies. Therefore, equivalent power curves based on wind farm SCADA data are established as physical baselines.

4.2.2. Equivalent power curve fitting based on binning regression

Wind turbine theoretical power is determined by aerodynamics, while actual curves deviate due to turbulence, wake, and blade contamination. Based on SCADA data of Yuanqu Wind Farm (single unit capacity 2.5 MW, hub height 90 m), equal-width binning (0.5 m/s interval) is adopted to construct equivalent power curves. Abnormal operating conditions are eliminated, and the median power of each wind speed bin is calculated. Local weighted scatterplot smoothing (LOWESS) is applied to smooth bin medians to eliminate random noise and obtain the final physical baseline, as shown by the red line in **Figure 1**.

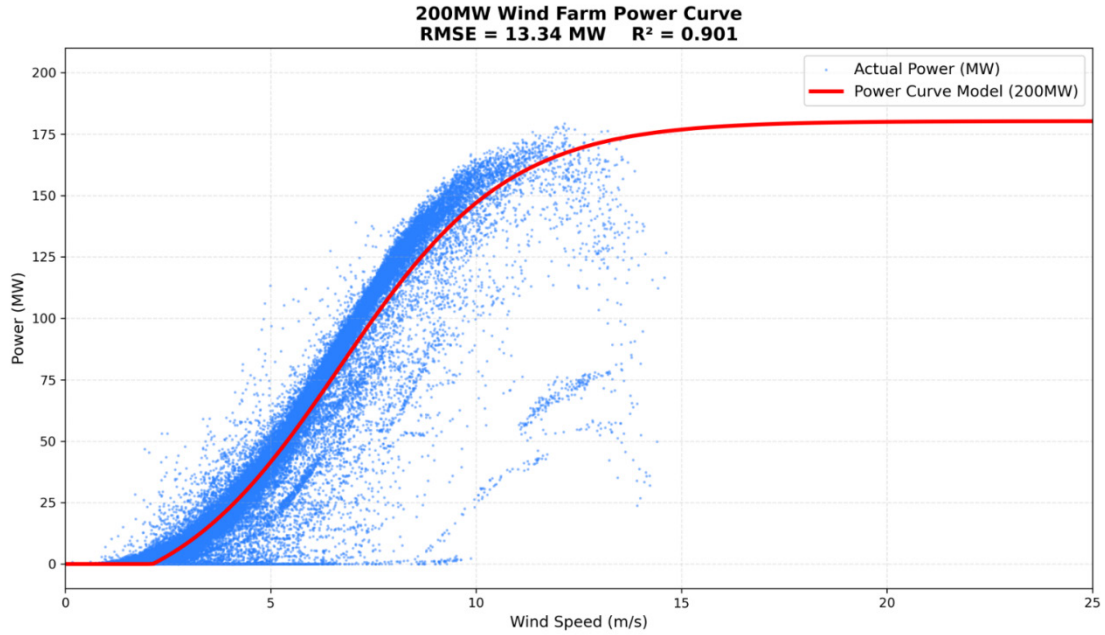


Figure 1. Equivalent power curve of Yuanqu wind farm.

4.3. Wind speed correction based on random forest

4.3.1. Characteristic analysis of wind speed forecasting errors

The bias correction in Chapter 2 mainly eliminates overall systematic shifts, while ECMF outputs still contain non-stationary residuals varying with meteorological conditions. Analysis of historical data from January to November 2024 shows that wind speed forecasting error $e = V_{\text{ECMF}} - V_{\text{obs}}$ has significant correlation with the following factors:

First, wind speed magnitude: ECMWF systematically overestimates wind speed in low-wind-speed areas (< 4 m/s) with a mean bias of $+0.6$ m/s, while it underestimates wind speed in high-wind-speed areas (> 12 m/s) with a mean bias of -0.8 m/s. Second, atmospheric stability: the forecast error of near-surface wind speed increases significantly under stable nocturnal boundary layer conditions. Third, ensemble spread: when the spread among ensemble members is large (> 1.5 m/s), the forecast features high uncertainty and a larger absolute error. Fourth, topographic complexity: wind speed errors show opposite signs in valley and ridge locations, with underestimation in valleys and overestimation on ridges.

Simple overall bias correction cannot eliminate local nonlinear errors, so data-driven methods are introduced to learn complex error mapping relationships.

4.3.2. Design of random forest regression model

Random forest (RF) is an ensemble learning algorithm that reduces variance by constructing multiple decision trees and averaging prediction results. Compared with deep neural networks, RF is robust to high-dimensional features and missing values, capable of outputting feature importance, less prone to overfitting, and fast in training.

(1) Input feature design

An independent RF model is built for each ensemble member. m (51 parallel training models). Input features $X_{m,t}$ fall into four categories (see **Table 1**):

Category	Specific features	Physical significance
Forecast wind speed	$V_{m,t}, V_{\text{mean},t}, V_{\text{max},t}, V_{\text{min},t}$	Information of single member and ensemble statistics
Ensemble uncertainty	Ensemble standard deviation, inter-quantile range ($Q_{90}-Q_{10}$)	Indicator of forecasting credibility
Atmospheric environment	Wind direction, air density, geopotential height, time coding (month/day/hour)	Characterize local meteorological background.
Historical errors	e_{t-1}, e_{t-2} (residuals of previous two lead times)	Autocorrelation of errors

(2) Model training

Data from January to October 2024 serve as the training set, and November data as the validation set. The target variable is measured wind speed from a 90 m anemometer tower. RF hyperparameters are optimized via grid search: 300 decision trees, maximum depth 20, minimum sample count for node splitting = 5, feature subset size = $\lceil \log_2(15)+1 \rceil$. The loss function adopts MSE. Out-of-bag (OOB) samples are used to monitor overfitting during training.

(3) Correction

For lead time t and member m , RF outputs corrected wind speed $v_{m,t} = f_{\text{RF}}^{(m)}(X_{m,t})$. The corrected wind speed ensemble $V_{m,t}$ retains the physical correlation structure of original members without reordering, with lower bias and more accurate spread.

4.3.3. Evaluation of correction effect

After random forest correction, the RMSE of wind speed forecast drops from 2.93 m/s to 0.80 m/s, and R^2 rises from -0.58 to 0.88. Data points distribute closely along the 1:1 reference line, with both systematic and random errors effectively suppressed, which proves the RF model can capture nonlinear deviation relationships between model-forecasted and measured wind speeds. (**Figure 2**)

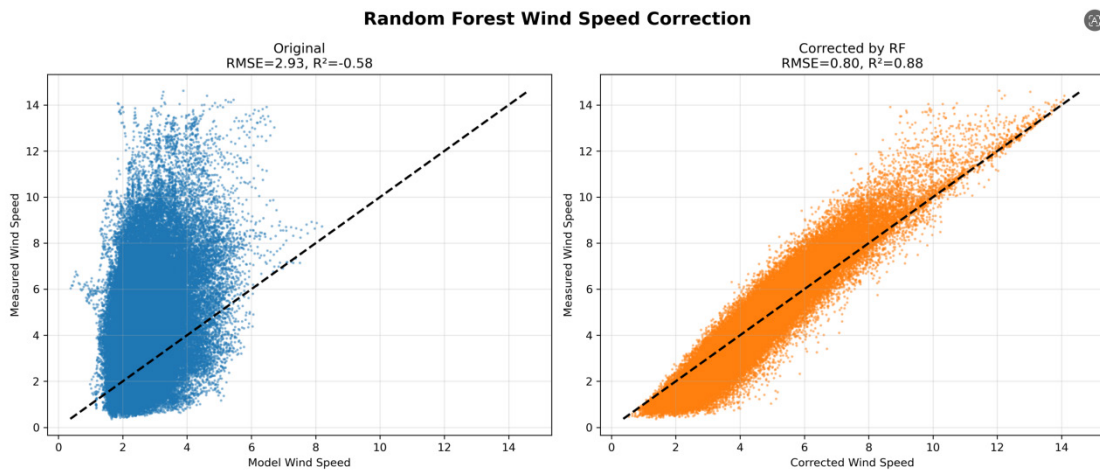


Figure 2. Comparison of wind speed before and after correction at Yuanqu wind farm.

4.4. Physical-data hybrid forecasting framework

4.4.1. Hybrid model architecture

Combining physical prior power values with RF-corrected wind speed, a hierarchical progressive hybrid

forecasting framework is proposed:

Where $\alpha_t \in [0,1]$ is the mixing weight dynamically adjusted with lead time. t ; Φ_{RF} represents the secondary power correction model taking corrected wind speed and auxiliary features as inputs. Based on error analysis of the validation period, α_t adopts inverse variance weighting:

$\sigma_{phy,t}^{-2}$ and $\sigma_{RF,t}^{-2}$ are historical RMSE reciprocals of the physical model and RF model at lead time t , respectively. Higher weight is assigned to the physical model when it has smaller errors at a certain lead time.

4.4.2. Multi-timescale rolling forecasting strategy

A segmented modeling + rolling update strategy is adopted for the 4–30 day horizon:

Short extended period (4–10 days): Forecasting errors grow slowly, dominated by the physical model ($\alpha_t=0.7$), with RF capturing small-scale fluctuations;

Medium period (11–20 days): Errors increase significantly, physical and data models share equal weight ($\alpha_t=0.5$);

Long period (21–30 days): Physical model errors saturate, dominated by data = 0.2- ing ($\alpha_t=0.2-0.3$), RF learns low-frequency climate signals.

ECMF latest forecasts are updated daily, and wind speed correction and power prediction are recalculated. The forecasting bias of RF power correction is adaptively adjusted online by comparing previous predictions with measured values. The effectiveness of this strategy relies on the ensemble spread consistency reconstruction in Chapter 2 to guarantee stable relative reliability of members across different lead times.

5. System integration and engineering verification

5.1. Overview

The previous chapters complete the method system, including data preprocessing, ensemble forecast bias correction, RF-physical hybrid power forecasting, and uncertainty quantification. This chapter integrates the above methods into a wind farm-level forecasting system and conducts engineering verification based on actual operating data of Yuanqu Wind Farm from January to April 2026, focusing on evaluating forecasting accuracy and business adaptability under practical operating environments.

5.2. System integration and verification data description

The wind power forecasting system constructed in this study adopts a modular hierarchical architecture. It automatically pulls the latest ECMWF ensemble forecast data (51 members covering the next 30 days) every day, and sequentially implements spatial interpolation, bias correction, random forest wind speed correction, and physical power curve conversion, finally outputting ensemble mean point forecasts, 90% probability intervals, and clustered typical multi-scenario trajectories. The whole workflow takes about 1.5 hours daily to meet business lead time requirements. The test period covers 60 days from January 6 to March 6, 2026, spanning winter to early spring, covering cold wave strong winds, clear calm wind, cloudy weak wind, and frontal transit weather, which can comprehensively test model performance under various meteorological conditions. The verification benchmark is synchronous actual grid-connected power of Yuanqu Wind Farm SCADA system, with installed capacity 200 MW and 15-min time resolution.

5.3. Verification of wind speed correction and evaluation of power forecasting accuracy

Based on test data from February 1 to April 8, 2026 (about 67 days), the accuracy of RF-corrected ECMWF wind speed is evaluated. **Figure 3** shows the comparison between corrected wind speed and measured values. In the scatter diagram, the horizontal axis represents forecasted wind speed after correction, and the vertical axis represents measured wind speed; ideal points are distributed on the 1:1 diagonal line. The time series diagram covers the whole test period from early February to early April. Quantitatively, the RMSE of corrected wind speed is 2.28 m/s $R^2=-0.07$. An RMSE of 2.28 m/s is acceptable for forecasting complex mountain wind farms, basically consistent with the error range of 2.0–2.5 m/s of typical commercial systems. Negative R^2 indicates that the average value of corrected wind speed is inferior to directly taking the measured mean as the prediction, mainly because forecasting is systematically about 1.5 m/s lower during cold wave strong wind periods in early February, leading to low overall goodness of fit. Time series curves show that corrected wind speed is smoother than measured values, with limited ability to capture high-frequency fluctuations.

Overall, random forest correction effectively reduces systematic biases of raw ECMWF forecasts, yet there is still room for improvement in capturing extreme wind speed events and short-term fluctuations.

Figure 4 shows the comparison between ensemble mean deterministic forecast and measured power. The power generation accuracy of ensemble mean point forecasting reaches 83.26% with an RMSE of 42.73 MW (normalized 21.4%), complying with the standard NB/T 10215-2019 ($\geq 80\%$). By power segments: the medium power zone (40–120 MW) achieves the optimal accuracy of 86%; the low power zone is overestimated (78%); the high power zone is underestimated (81%); the accuracy near full load (>180 MW) is 75%.

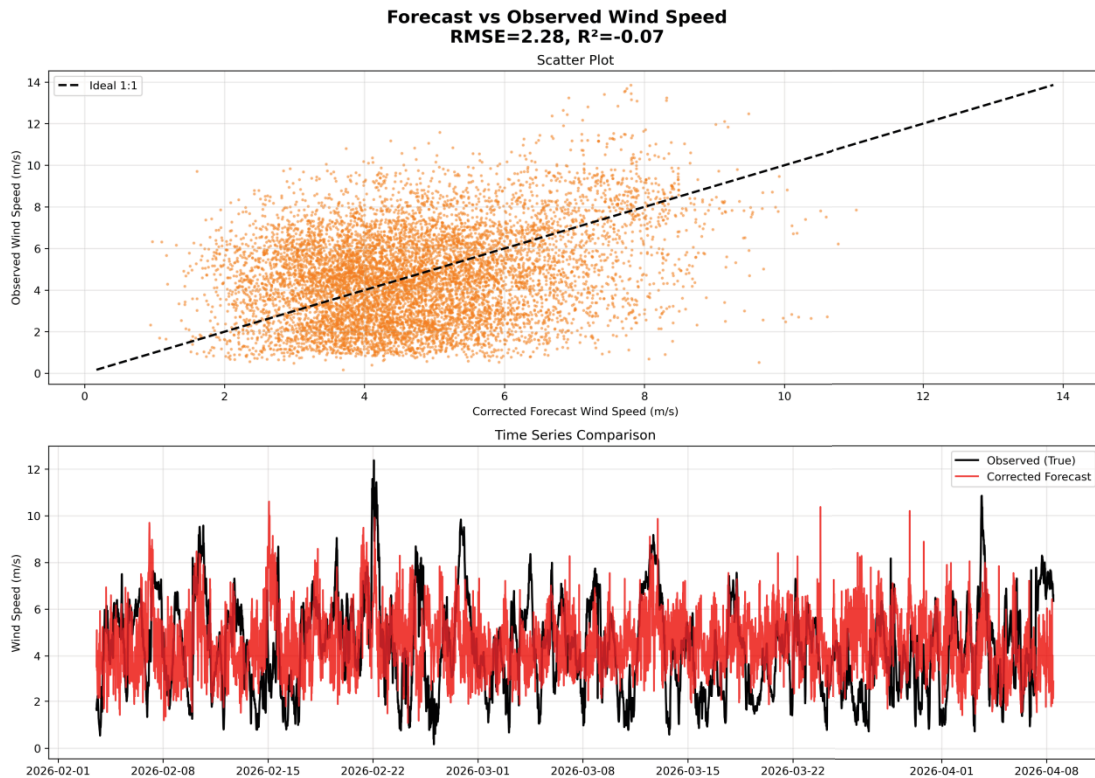


Figure 3. Wind speed comparison before and after correction at Yuanqu wind farm.

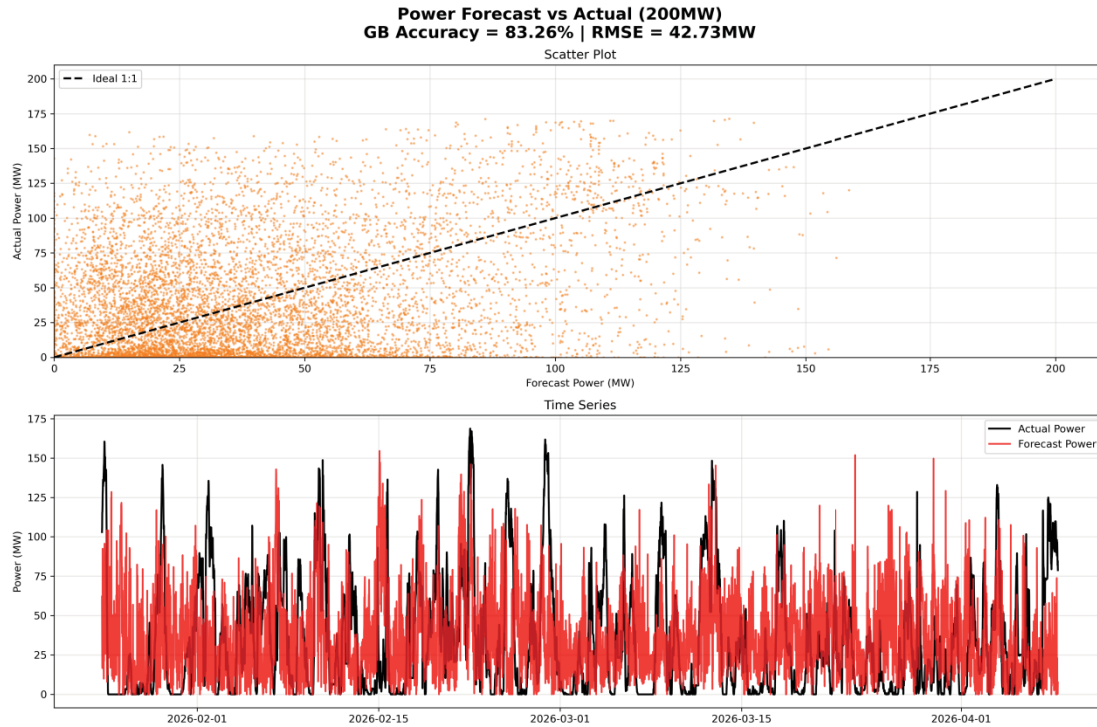


Figure 4. Time series of wind power forecasting at Yuanqu wind farm.

5.4. Limitations and improvement directions

Although the power generation accuracy of 83.26% is achieved at Yuanqu Wind Farm and meets industry standards, several limitations remain. The forecasting capacity for extreme wind events is insufficient (power underestimation up to 30% during cold waves), RF outputs tend to be smooth, the coverage rate for 21–30 day lead times is only 79%, and verification is limited to a single wind farm. Follow-up improvements include oversampling of extreme events, integration of radar and satellite data, online incremental learning and multi-wind farm extended verification.

6. Conclusions and prospects

6.1. Main conclusions

Centering on multi-timescale wind power forecasting and uncertainty quantification based on ensemble forecasting and taking Yuanqu Wind Farm as the research object, this paper carries out systematic research on three dimensions, including multi-source ensemble forecast data adaptive processing, random forest-physical hybrid power forecasting, and ensemble member probabilistic multi-scenario generation. Model training and independent verification are completed based on ECMWF ensemble forecast data and wind farm measured data from January 2024 to February 2025, with main conclusions as follows:

- (1) Adaptive processing of ensemble forecast data significantly improves wind speed forecasting accuracy. Systematic comparison of four global ensemble forecasting systems (ECMWF-ENS, UKMO-EGRR, CMA-BABJ, NCEP-GEFS) verifies that ECMWF has the minimum forecasting error (monthly RMSE 0.90–1.24 m/s) and optimal seasonal stability under complex mountain terrain. A complete processing workflow including spatial downscaling, quantile mapping, systematic bias correction, and ensemble

spread consistency reconstruction is established for ECMWF data. After adaptive processing, the RMSE of ECMWF wind speed forecasts is reduced by 15–25% across all lead times, and the matching degree between ensemble spread and actual errors is greatly improved, laying a high-quality data foundation for subsequent power forecasting.

- (2) Random forest-physical hybrid driving method realizes high-precision medium and long-term power forecasting. A hybrid strategy of correcting wind speed via random forest followed by equivalent power curve conversion is proposed to address the problem that single physical models accumulate errors and pure data models lack physical consistency. In wind speed correction, RF integrates ensemble statistics, atmospheric environment, and historical error features, reducing ECMWF wind speed RMSE to 2.28 m/s ($R^2=0.07$) and effectively compensating systematic biases of extreme wind speeds. In power forecasting, equivalent power curves based on SCADA data serve as physical baselines, and dynamic weight fusion is adopted for physical and data outputs. Test results show that the power generation accuracy of ensemble mean point forecasting reaches 83.26% with an RMSE of 42.73 MW (normalized to 21.4%), satisfying power industry standards. Lead time analysis reveals the highest accuracy for 4–10 days, followed by 11–20 days, and acceptable performance for 21–30 days. Segmented power analysis shows the best performance in the medium power range (40–120 MW), with overestimation in low-power zones and underestimation in high-power zones.
- (3) System integration proves the engineering deployment value of the proposed method. The modular forecasting system automatically completes the whole workflow of data pulling, preprocessing, wind speed correction, power conversion, and scenario generation within about 1.5 hours daily, satisfying business time constraints. Taking the independent test period from January to March 2026 as an example, its supporting capacity for 10-day/20-day 15-min interval ten-day transactions and 30-day daily average monthly total transactions is verified. Taking cold wave strong wind events as examples, high-power warnings are issued 3 days in advance and low wind risks are identified beforehand, providing operable decision-making information for medium- and long-term contract electricity adjustment and reserve capacity verification. Despite the 83.26% accuracy reaching industry standards, there is still room for improvement in forecasting precision for extreme wind events and long lead times of 21–30 days.

6.2. Research prospects

This study achieves an accuracy of 83.26% at Yuanqu Wind Farm and verifies the effectiveness of the proposed method, yet multiple directions deserve further exploration:

- (1) Extension to subseasonal and seasonal forecasting horizons (30–60 days). Current research focuses on 4–30 day lead times relying on ECMWF medium-range ensemble products. Next, ECMWF Subseasonal-to-Seasonal (S2S) forecasts and large-scale climate mode indices (ENSO, AO, NAO) will be introduced, integrating low-frequency climate signals into forecasting models to break the predictability bottleneck beyond 30 days. Preliminary ideas include adding climate indices as RF features or building low-frequency component correction modules in hybrid models.
- (2) Multi-wind farm and regional collaborative forecasting. Current research targets a single wind farm, while power grid dispatching and electricity trading pay more attention to the overall output of regional wind and photovoltaic clusters. Subsequent research will extend the method to regional networks containing multiple wind and photovoltaic stations, and study spatial correlation between wind farm outputs and its influence on regional forecasting uncertainty. Technically, graph neural networks (GNN) can be adopted to model topological relations, or Copula functions can be used to describe the multivariate joint distribution of outputs.

Disclosure statement

The authors declare no conflict of interest.

References

- [1] National Development and Reform Commission, 2025, Circular on Deepening the Market-Oriented Reform of New Energy On-Grid Electricity Prices and Promoting High-Quality Development of New Energy: FGJG (2025) 36.
- [2] National Energy Administration, 2026, 2025 National Electric Power Industry Statistical Data. Beijing.
- [3] Shu Y, Chen G, He J, et al., 2022, Research on Framework of New Power System Dominated by New Energy Sources. *Proceedings of the Chinese Society for Electrical Engineering*, 42(12): 4233–4245.
- [4] Jin L, Wang Y, Chang J, 2023, Ultra-Short-Term Wind Power Forecasting Based on Parallel Machine Learning. *Journal of Hydroelectric Engineering*, 42(1): 40–51.
- [5] Yuan C, Tang Y, Mei R, et al., 2021, A PSO-LSTM Model of Offshore Wind Power Forecast Considering the Variation of Wind Speed in Second-Level Time Scale. *Mathematical Problems in Engineering*, 2021: 1–9.
- [6] Li M, 2021, Wind Power Forecasting Based on Wind Farm Meteorological Data, thesis, Inner Mongolia University of Science and Technology.
- [7] Mu M, Wang Q, 2016, Research on Predictability in Numerical Weather Prediction. *Meteorology*, 42(7): 781–790.
- [8] Bao M, Ni Y, 2003, Research on Error Growth in Medium-Range Numerical Weather Prediction. *Chinese Journal of Atmospheric Sciences*, 27(5): 889–900.
- [9] Wang Z, Zhang Y, et al., 2014, Analysis on Application of ECMWF Ensemble Forecast Products in Medium-Range Weather Forecasting. *Meteorology*, 40(11): 1361–1368.
- [10] Liu Y, Zhao D, Song L, 2020, Research on Wind Speed Correction Method Based on ECMWF Ensemble Forecast. *Meteorology*, 46(8): 1069–1078.
- [11] Wang B, Wang N, Liu C, 2014, Wind Power Forecasting Technology Based on Numerical Weather Prediction. *Power System Technology*, 38(10): 2741–2747.
- [12] Jin R, Ma J, Li Y, 2015, Research on Application of ECMWF Ensemble Forecast for Medium-Range Weather Forecasting in China. *Meteorology*, 41(5): 573–582.
- [13] Du J, Chen J, 2021, Review and Prospect of Ensemble Forecasting Methods. *Meteorology*, 47(12): 1433–1447.
- [14] Feng S, Wang B, Wang Z, et al., 2021, Research Progress and Prospect of Wind Power Forecasting Technology in China. *Power Generation Technology*, 42(1): 80–87.
- [15] Ma J, Wang B, Liu C, 2019, Research on Photovoltaic Power Forecasting Method Based on Ensemble Forecasting. *Renewable Energy Resources*, 37(11): 1601–1607.
- [16] Zhan W, Che J, Wang B, et al., 2024, Regional Multi-Output Photovoltaic Forecasting Method Based on Gridded Numerical Weather Prediction. *Electric Power*, 57(3): 144–151.
- [17] Zhu R, Zhang H, Li W, 2025, Application of Numerical Weather Prediction Technology in Renewable Energy Generation. *Hydropower and New Energy*, 39(9): 69–72.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on Equipment for the Preparation of PAN Photocatalytic Nanofiber Membranes

Zhifeng Yang

Shenyang Polytechnic College, Shenyang 110045, Liaoning, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Efficient degradation of organic dye wastewater is a core urgent problem to be solved in the field of water pollution control. Conventional water treatment processes suffer from low degradation efficiency and severe secondary pollution, and cannot meet the discharge standards of industrial wastewater with complex compositions and fluctuating concentrations. This is the core reason for the rapid popularization and practical application of green photocatalysis technology. Visible light photocatalysis is driven by clean solar energy, featuring environmental friendliness, mild reaction conditions, and no demand for high-temperature and high-pressure auxiliary equipment, which makes it highly suitable for large-scale treatment of industrial dye wastewater. Polyacrylonitrile (PAN) nanofiber membranes possess unique merits including low cost, high flexibility, stable structure, and abundant pores. Through multi-step precise chemical modification to support heteropolyacid active components, recyclable visible-light-responsive photocatalytic membrane materials with high degradation efficiency and excellent stability can be fabricated.

Keywords: PAN photocatalysis; Nanofiber membrane; Preparation; Equipment research

Online publication: Sept 4, 2026

1. Equipment system for electrospinning of virgin PAN nanofiber membranes

The electrospinning equipment system for virgin PAN nanofiber membranes is developed to produce polyacrylonitrile substrate nanofiber membranes with uniform structure, stable performance, and low defect rate, which provides standardized, defect-free base materials for subsequent multi-step refined chemical modification and experimental research on visible light photocatalytic performance.

1.1. Equipment functional orientation

The complete electrospinning device consists of three independent units: pretreatment, main molding unit, and environmental control unit, which have clear divisions of labor and operate cooperatively throughout the whole process to form an integrated molding system. The pretreatment unit can effectively break up agglomerated powder particles, remove raw material impurities and bubbles, and prepare uniform, stable, non-delaminated spinning solutions. The main machine cooperates with the environmental control unit to

jointly guarantee the overall molding quality and microstructural uniformity of nanofiber membranes. The equipment supports customized spinning parameters exclusive to polyacrylonitrile materials and can store mature process data for long-term use, which effectively eliminates errors caused by manual debugging and ensures consistency and repeatability of experiments across batches. The single molding duration can be precisely regulated, and the finished product size perfectly matches the feeding specifications of subsequent modification equipment, which minimizes mechanical loss during substrate cutting and transfer and improves raw material utilization ^[1].

1.2. Core equipment structure and operating parameters

The spinning solution pretreatment equipment is composed of an ultrasonic cleaner, constant-temperature stirrer, and sealed liquid storage device. All structural parts in contact with materials are made of corrosion-resistant materials to adapt to long-term operation with strong polar organic solvents. Stable dissolution and dispersion processes eliminate molding defects such as fiber beads and broken filaments at the source, and fully guarantee the comprehensive quality and long-term storage stability of spinning solutions. The TL-Pro fully enclosed electrospinning main machine serves as the core molding equipment. Its high-precision high-voltage electric field, micro quantitative liquid supply, and dynamic roller collecting system can precisely match the exclusive spinning process of high-viscosity PAN spinning solutions. The equipment is equipped with intelligent parameter monitoring and automatic overload protection functions to monitor voltage, flow rate, and rotational speed in real time, which ensures the molding precision and uniform diameter of nanofibers and greatly improves the overall safety of experimental operations. The supporting intelligent temperature and humidity control unit can precisely regulate the microenvironment inside the spinning chamber to maintain optimal conditions for steady, uniform volatilization of organic solvents and avoid structural defects caused by excessively fast or slow solvent evaporation. The integrated negative pressure exhaust and waste gas adsorption device stabilizes fiber-forming conditions and prevents uneven membrane thickness induced by airflow fluctuation, while eliminating safety and environmental hazards brought by organic solvent volatilization ^[2].

1.3. Adaptability analysis between equipment and PAN spinning process

The low-speed uniform output mode of micro-injection pumps is highly compatible with the refined molding demands of high-viscosity PAN spinning solutions, avoiding pulse output defects of conventional pumps. The continuous and stable fluid output fully stretches and refines fiber structures and sufficiently exposes cyano active groups on membrane surfaces, laying a solid structural foundation for subsequent cyclization and quaternization modification reactions. The stepless adjustable roller speed of the equipment generates moderate stretching orientation on deposited nanofibers and optimizes the internal molecular arrangement of fibers ^[3]. A reasonable rotational speed range balances fiber flexibility and overall structural integrity, and completely eliminates structural defects such as loose fiber stacking, fragile membranes, and overstretched broken filaments. Constant temperature and humidity conditions precisely control solvent volatilization rates to prevent fiber adhesion and pore closure, facilitating the preparation of high-quality fiber membranes with high porosity, large specific surface area, and loose structures. The abundant three-dimensional pore channels inside the membranes can boost mass transfer efficiency and reaction uniformity during subsequent functional group modification and ion exchange, reducing incomplete local reactions. The fully enclosed and

corrosion-resistant structure of the equipment adapts to continuous experiments with organic solvents for a long time without pipeline corrosion, material leakage, or solvent volatilization. The standardized process parameters stored in the system minimize batch experimental errors and greatly enhance the stability and repeatability of substrate preparation.

2. Complete equipment system for chemical modification of PAN fiber membranes

Virgin polyacrylonitrile fiber membranes only possess physical screening and supporting structures without any photocatalytic activity or pollutant degradation capacity. Multi-step refined chemical modification is mandatory to realize functional upgrading of materials. The complete set of modification equipment covers full procedures, including high-temperature reflux reaction, multi-solvent cleaning, constant-temperature ion exchange, and vacuum drying, and ensures stable, standardized, and efficient progress of modification reactions.

2.1. Equipment functional orientation

The four modification equipment units perform their respective duties and connect in sequence to form a complete functional modification production line, which sequentially completes core modification reactions such as cyclization and quaternization of fiber surface functional groups. Standardized equipment operating conditions largely reduce residual impurities and significantly improve the loading uniformity and interfacial bonding stability of active components. All material-contact parts of the modification equipment are made of anti-corrosion and solvent-resistant materials to fully adapt to complex organic systems and alternating acid-base working conditions. The exclusive flexible membrane clamping tool matched with the equipment effectively prevents stretching and breakage of fiber membranes during soaking and transfer, greatly improving the overall success rate and intact rate of experiments ^[4].

2.2. Core equipment structure and operating parameters

The double-layer glass reactor acts as the core modification reaction equipment. Its high-precision closed constant-temperature system perfectly meets the demands of long-term reflux reactions lasting several hours to more than twenty hours, with far superior temperature stability compared with conventional heating equipment. The matched condensation reflux and low-speed stirring structure maintains stable solvent concentration and temperature of the reaction system, fully guaranteeing sufficient modification reactions of fiber surface functional groups. The special membrane cleaning workstation abandons the extensive stirring and soaking mode of traditional equipment and adopts multi-angle spraying to treat soft, flexible fiber membranes in a targeted manner. The standardized multi-solvent alternating cleaning procedure set by the equipment removes residual organic reagents and free impurities layer by layer, ensuring the authenticity and accuracy of subsequent photocatalytic performance test data ^[5]. The constant-temperature magnetic stirring water bath provides mild, low-temperature liquid-phase environments for anion exchange reactions and avoids decomposition and deactivation of active components caused by high temperatures. The fully sealed oxygen-isolated design of the equipment prevents oxidation of heteropolyacid active components and remarkably improves their loading uniformity and bonding strength inside fiber membranes. The vacuum low-temperature drying oven rapidly removes residual organic solvents and water inside membrane pores

under low-pressure vacuum conditions to shorten the drying cycle. The mild low-temperature vacuum working condition will not damage fiber skeletons or surface-active functional groups, and completely preserves the microstructures and visible-light photocatalytic performance of modified membranes ^[6].

2.3. Adaptability analysis between equipment and modification processes

The circulating heat conduction temperature control mode of double-layer glass reactors features tiny temperature drift and superior temperature stability to conventional water baths, which is highly compatible with long-term continuous organic modification reactions. The sustained stable high-temperature environment fully activates inert cyano groups on fiber surfaces and greatly raises the grafting efficiency and overall uniformity of cyclic active functional groups. Modified fiber membranes undergo an obvious decline in mechanical properties after multi-step chemical reactions, featuring reduced flexibility and loose structures, and are prone to breakage during cleaning. The special clamping tools matched with the equipment can fully protect membrane structures while thoroughly removing free impurities on membrane surfaces, effectively avoiding inflated deviations in photocatalytic experimental data ^[7]. The mild low-speed stirring condition of the equipment prevents shedding of functional groups on membrane surfaces and ensures smooth and ordered solid-liquid ion exchange reactions. Constant temperature accelerates anion diffusion and migration, enabling uniform loading of heteropolyacid active components inside the three-dimensional fiber skeletons and improving loading stability. Polyionic liquid functional groups have poor high-temperature resistance; traditional high-temperature drying easily causes decomposition and shedding of active components. The vacuum low-temperature drying process avoids this technical problem and stably retains intact microstructures and visible-light catalytic activity of modified membranes.

3. Complete equipment system for characterization and performance testing of photocatalytic membranes

The quality of modified photocatalytic fiber membranes cannot be intuitively judged only by preparation process parameters. Systematic and comprehensive testing is required to verify their microstructural characteristics and practical catalytic performance. The complete set of testing equipment covers full workflows of structural characterization, performance detection, and sample pretreatment, and provides reliable data support for material performance analysis and process optimization. The complete testing equipment system serves as a core link connecting material preparation and performance verification. It breaks the barrier between equipment processes, material structures, and catalytic performance, accurately identifies process defects and performance deficiencies generated during preparation, and provides an all-round, high-precision experimental basis for iteration of modification and optimization of equipment parameters. It is an important foundation for guaranteeing the scientificity and repeatability of overall experiments.

3.1. Equipment functional orientation

Multi-dimensional precision characterization equipment comprehensively analyzes the micro-morphology, chemical composition, crystal structure, and photoelectric properties of fiber membranes, and establishes an internal correlation between material microstructures and macroscopic catalytic performance. The special visible light testing system can quantitatively measure the dye degradation capacity and long-term cycling

stability of materials, and objectively evaluate their practical industrial application value. The complete testing equipment is highly compatible with the material characteristics and testing requirements of PAN-based modified photocatalytic membranes, featuring strong equipment compatibility, high detection precision, and excellent repeatability. The platform supports simultaneous parallel testing of multiple groups of samples, which can intuitively compare performance differences of materials prepared via different modification processes and different active component loading routes ^[8].

3.2. Core equipment structure and operating parameters

Various precision characterization instruments have clear divisions of labor and complement each other. They conduct all-round quality detection from multiple dimensions, including micro-morphology and surface functional groups. The comprehensive detection combining multiple testing methods can accurately judge the completion degree of modification reactions and the actual loading quality and uniformity of active components. Mutual verification among different characterization devices avoids one-sided results caused by single detection means, fully analyzes the influence mechanism of equipment processes on material microstructures, and precisely locates tiny differences in fiber pore structures, functional group distribution, and active component loading, providing sufficient micro-data support for analyzing equipment adaptability and material performance rules. The visible light photocatalytic testing system accurately screens pure visible light bands via special filters, eliminating ultraviolet interference and fully conforming to laboratory research standards and practical water treatment scenarios ^[9]. The fully light-shielded, constant-temperature, and automatic continuous sampling design integrated into the equipment minimizes external environmental interference and guarantees precise and reliable degradation experimental data. The supporting sample pretreatment equipment standardizes basic operations such as sample preparation, unifying preparation standards for the whole experimental flow. Unified sample preparation standards eliminate random errors induced by manual operations and fully ensure the scientificity and comparability of multiple parallel experimental data.

3.3. Adaptability analysis between equipment and testing processes

The multi-dimensional integrated characterization system comprehensively verifies the practical effects of modification processes and accurately screens defects in material microstructures and unstable loading of active components. The equipment is highly adapted to the testing characteristics of flexible thin-film samples; it simplifies sample preparation without damaging the original microstructures and active systems of samples. The multi-layer optical shielding and band filtering design of the testing system completely eliminates experimental interference from ambient stray light and ultraviolet rays, ensuring single and stable light sources. The stable automatic sampling mode guarantees continuous progress of photocatalytic reactions and precisely fits the kinetic laws of dye degradation to truly restore the intrinsic photocatalytic performance of materials. The three types of testing, cleaning, and drying equipment form a closed-loop testing system to efficiently and continuously complete long-term cycling stability tests and shorten experimental cycles ^[10].

Fiber membranes can be directly clamped and recovered, which thoroughly solves the technical pain points of easy loss and difficult recovery of active components of traditional powder catalysts, making test results more consistent with practical industrial application scenarios. The complete closed-loop testing equipment system fits the operating logic of industrial water treatment equipment and can simulate

continuous working states of actual water treatment processes. It effectively verifies the structural stability and performance retention rate of photocatalytic membranes during long-term service, accurately predicts the industrial service life and application feasibility of materials, and provides authentic experimental reference for subsequent large-scale equipment transformation and industrialization of technologies. High-precision spectrophotometers are specially adapted to micro-detection of low-concentration organic dye solutions with extremely low detection errors and excellent linear fitting, delivering highly credible test data. The instrument can capture tiny performance differences of materials under different process conditions and provide solid data support for subsequent process optimization, performance improvement, and breakthroughs in degradation efficiency.

4. Conclusion

The three major equipment systems of electrospinning molding, chemical modification, and performance characterization form a full-process closed-loop process for preparing PAN photocatalytic nanofiber membranes, realizing integration of preparation and detection. The equipment parameters of each system match each other with complementary functions and can stably produce recyclable photocatalytic functional membranes with intact structures and excellent performance. This customized special equipment system targets many industry pain points of traditional photocatalytic materials, such as difficult preparation, high recovery cost, and easy secondary pollution, and possesses remarkable advantages in production cost and service performance. The modular, upgradable, and automatically iterative characteristics of the equipment can adapt to the demands of pilot and industrialization in the future, providing solid equipment support and systematic technical reference for the industrial application of visible light photocatalytic water treatment technology.

Funding

Basic Scientific Research Project of Liaoning Provincial Department of Education (2023): Research on Polyionic Liquid Nanofiber Membranes Prepared via Chemical Modification of PAN and Their Photocatalytic Properties (Project No.: JYTMS20231863)

Disclosure statement

The author declares no conflict of interest.

References

- [1] Zhang H, Wang L, Zhang X, et al., 2025, Fabrication of g-C₃N₄/PAN+TPU Composite Nanofiber Membrane and Its Performance for Treating Printing and Dyeing Organic Pollutants. *Journal of Molecular Catalysis*, 39(2): 129–140.
- [2] Jia Y, Jiang C, Guo Z, et al., 2017, Research Progress on Advanced Treatment and Reuse of Printing and Dyeing Wastewater. *Journal of Textile Research*, 38(8): 172–180.
- [3] Liu Y, Li X, Wang D, 2021, Fabrication and Photocatalytic Performance of Molybdenum Disulfide/PAN

- Composite Nanofiber Membrane. *Dyeing & Finishing*, 47(8): 62–66.
- [4] Wang F, Zhao Y, Chen C, 2023, Modification of PAN-Based Composite Nanofiber Membrane and Its Visible Light Catalytic Degradation of Dye Wastewater. *Technology of Water Treatment*, 49(5): 78–83.
- [5] Li M, Zhang Y, Sun X, 2022, Fabrication of PAN-Based Photocatalytic Membrane via Electrospinning and Its Cyclic Stability. *Journal of Functional Materials*, 53(11): 11132–11137.
- [6] Wang X, Zhao W, Li Q, 2022, Fabrication and Visible Light Catalytic Performance of Phosphotungstic Acid Modified PAN Nanofiber Membrane. *Chemical Industry and Engineering Progress*, 41(7): 3210–3218.
- [9] Yang S, Li H, Wang Z, 2024, Photocatalytic Mechanism of Heteropolyacid Modified PAN Membrane. *Acta Scientiae Circumstantiae*, 44(2): 156–163.
- [10] Zhao X, Chen Y, Ma W, 2023, Fabrication of Visible-Light-Responsive PAN Composite Membrane and Its Application in Printing and Dyeing Wastewater Treatment. *Industrial Water Treatment*, 43(8): 94–99.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on Teaching Resource Generation for the Theory-Practice Integrated Curriculum in Higher Vocational Education from the Perspective of Human-Machine Collaboration

Liming Feng, Ning Pan*

City College of Huizhou, Huizhou 516025, Guangdong, China

*Corresponding author: Ning Pan, panning@tm.hzc.edu.cn

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: In recent years, there have been several issues with the development of theory-practice integrated teaching resources in higher vocational education: slow updates, a lack of individualization, and a long cycle. Generative AI offers a new approach to human-machine collaborative development of resources. Based on the case of the “Firewall Technology” course, an all-encompassing chain intelligent resource generation framework has been established in this paper to cover the conversion of knowledge points into project tasks, the generation of theory-practice integrated assignments, the creation of practical training manuals, intelligent generation of test item banks, and support for students’ self-assessment and peer assessment. Using a prompt engineering approach that integrates role empowerment, task decomposition, format constraints, and iterative optimization, we developed a reusable prompt template library. Based on the above practice, the human-machine collaboration model can reduce the initial draft generation time of a single resource from several hours to a few minutes, achieve a quality level of good to excellent according to peer review, and increase the efficiency of personalized exercise generation by more than 80%. The model for human-machine collaboration in resource generation for the theoretical-practical integrated courses at the higher vocational education level put forward in this paper is effective and transferable.

Keywords: Human-machine collaboration; Theory-practice integration; Intelligent generation of teaching resources; Prompt engineering; Large language model

Online publication: Sept 9, 2026

1. Introduction

The main way to train technical and skilled personnel at the higher vocational level is through theory-practice integrated courses, and the quality of the supporting teaching materials will directly affect teaching results. The three main problems that have arisen in the development of traditional resource models over the

years are: First, the development cycle is long; it takes several hours or even days for teachers to create high-quality theory-practice integrated assignments and training materials. Second, the speed of content updates is slow; static resources cannot keep up with changes in the industry, especially in rapidly changing areas such as cybersecurity. Third, there is a lack of individualized support; standardized resources do not meet the various learning needs of all students^[1-3].

With the development of generative AI, new types of human-machine collaboration and development have appeared. Large language models can understand the instructions given in natural language, generate organized text, and have multi-turn conversations. Although some previous studies have applied artificial intelligence in the areas of instructional design and test item generation, most of them have only reached the level of “tool use” and have not established a general method for covering the whole resource development process^[4-6]. The difference is more pronounced in the area of theory-practice integrated courses, which are typical features of vocational education. Wang Yali and others have investigated the full-cycle construction of AI-enabled course resources; Cao Rui and others have developed a platform for generating personalized test items; Chen Lu has studied the effect of prompt engineering on the creation of course syllabi and exercises; Ou Zhigang and others have tested the usability of AIGC teaching materials.

“Firewall technology” is taken as the case study in this paper to demonstrate the whole framework of intelligent resource development, covering all links from “knowledge points → project task transformation → theory-practice integrated assignment generation → training manual generation → intelligent test bank creation → support for student self-assessment and peer assessment”. Prompt Engineering based on “role empowerment - task decomposition - format constraints - iterative optimisation” is employed to build a practical human-machine collaboration model for the development of vocational education resources here.

2. Theoretical foundation and overall framework

2.1. Theoretical foundation of human-machine collaborative education

The foundation of the development of human-machine collaborative teaching resources is the Distributed Cognition Theory. Cognitive activities are distributed among people, tools and environments according to this theory, and technological tools can extend the cognitive abilities of people. Teachers have disciplinary knowledge, pedagogical experience, and the ability to judge the quality of teaching resources; AI is good at rapid generation, text processing, and pattern recognition. The collaboration between the two forms a complementary, mutually reinforcing closed-loop process of “teacher design—AI generation—teacher review—iterative optimization”^[4,5].

This paper will also present the Human-Machine Collaborative Cognitive Load Theory and the Socio-technical Systems Theory. AI can be employed to organize the routine and repetitive tasks of creating templates for configuration commands and standardizing test items; thus, teachers’ cognitive resources will be released for more demanding work, such as evaluating the suitability of the content and its alignment with teaching objectives. At the same time, to ensure the long-term development and wide application of this system, institutional support in the form of a three-level review mechanism, prompt template libraries, and anti-plagiarism regulations should be provided. According to Cognitive Load Theory, the main goal of human-machine collaboration is to reduce both the teachers’ generative cognitive load and the students’ comprehension cognitive load.

The first division of labor in human-machine collaboration is that teachers act as “architects” to set learning goals, task structures, question types, quality requirements, anti-plagiarism measures, etc., and AI serves as the “implementer” to generate the first draft, provide solutions, and respond to revisions according to instructions. In this way, teachers can use their professional skills and AI will help increase the efficiency.

2.2. Principles for the development of theory-practice integrated course resources

In line with the Systematic Work Process Courses Theory and the philosophy of Action-Oriented Teaching, the construction of theory-practice integrated courses should be based on the following principles: contextualisation of tasks (integrating real-world enterprise project scenarios), transformation of knowledge points into project tasks (changing from a knowledge-based logic to an action-based logic), sufficiency of theory for practical application (limited to guiding the operation of practical training and aligned with operational sequences), and various types of questions with authentic assessments (such as fill-in-the-blank, tables, matching, short-answer and scenario analysis); training manuals should also have anti-plagiarism measures, such as personal identification screenshots and student ID directories ^[7–10].

2.3. General framework for intelligent generation of human-machine collaborative resources

The whole process of teaching resource development in this paper is shown through the framework of human-machine collaborative intelligent generation (**Figure 1**). Teachers need to be responsible for designing the overall courses, prompt engineering, quality assurance, and feedback on teaching materials; AI can help them create the first draft of courses and other materials, and then make several revisions based on feedback. Together, they will form project-based task systems, theory-practice integrated assignment banks, training manual libraries, dynamic question banks, personalized learning materials, and student learning profiles.

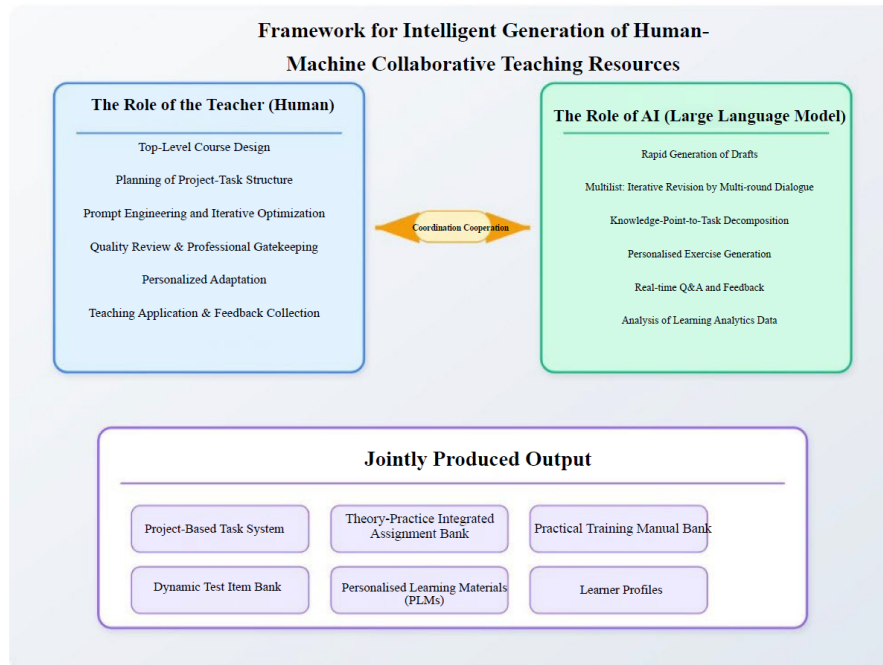


Figure 1. General Framework of Intelligent Generation of Human-Machine Collaborative Teaching Resources.

2.4. Prompt engineering and iterative optimization techniques

The four parts of the proposed prompt engineering method are: 1. Role Empowerment: Design the AI's role, such as a cybersecurity course teacher with 20 years of experience who is familiar with Huawei firewall configuration and has some knowledge of the cognitive patterns of higher vocational college students. 2. Task Decomposition: Break down a large development task into several small, manageable steps; for instance, in the creation of a practical manual, one can first have the AI set training goals, then present the network diagram, next write configuration steps, and finally list verification measures and discussion prompts. 3. Format Constraints: Specify the required structure of the output, such as "Display the results in a table", "Insert an empty line after each question", and "Precede all configuration commands with serial numbers". 4. Iterative Optimization: Repeatedly interact with and revise the prompt to improve the initial results. When teachers find that some parameters are missing or the difficulty of the AI-generated content is not suitable, they can issue targeted corrections, such as adding mode *pat* to the configuration command for Step 3; "Question 2 in the short-answer section is too difficult, please simplify it", etc. Generally, two or three rounds of revision can meet teaching requirements^[4,6].

3. Principal use cases and prompt design

3.1. Conversion of knowledge points into project tasks

Traditional textbooks are organized according to the knowledge-based logic of "protocol principles → equipment configuration → comprehensive application", and theory-practice integrated courses need to be reformed based on the action-oriented logic of "project → task → knowledge points"—the two work together to achieve this goal^[7].

Prompt Template: Based on the list of knowledge points above, design a Project-Based Task Structure for the "Firewall Technology" course. Requirements: 1. Cover the following knowledge points: Cybersecurity basics, Firewall Security Zones, Security Policies, NAT, Dual-Hot Standby, Encryption Technology, PKI, IPSec VPN, SSL VPN. 2. Organize the content according to the sequential workflow of enterprise network security construction and form five projects. 3. Each project should have 3–6 tasks, and the task names should describe what is done. 4. Output Format: A table with the project number, project name, task list, and related knowledge points.

The initial results of the AI were 5 projects and 24 tasks (**Table 1**), and then teachers reviewed and optimized them to create the final course structure.

Table 1. Structure of project-based courses generated with AI assistance (excerpt)

Project number	Project name	Core task	Knowledge points
P1	Introduction to Enterprise Network Security	Concept understanding, communication fundamentals, VRP operation, remote login	Definition of cybersecurity, TCP/IP protocol, VRP commands
P2	Building Enterprise Network Access Control Systems	Threat prevention, security zones, security policies, NAT, NAT Server	Attack types, security zones, policy configuration, NAT principles
P3	Ensuring Business Continuity	VRRP principles, hot standby, AAA authentication, user management	High availability, HRP, authentication & authorization, accounting
P4	Secure Remote Access	Encryption principles, PKI, IPSec VPN, SSL VPN	Symmetric/asymmetric encryption, digital certificates, VPN technologies
P5	Comprehensive Practical Training	Planning, deployment, testing	Comprehensive application

3.2. Intelligent generation of theory-practice integrated tasks

Theory-practice integrated assignments need to turn abstract theories into operational context problems, and the types of questions can be fill-in-the-blank, tables, matching, short-answer, scenario analysis, etc.^[1,10].

Prompt Template: You are a teacher of the “Firewall Technology” course at a vocational college. Design a theory-practice integrated assignment for the task of “[Task Name]” in accordance with the following requirements: 1. Create an actual enterprise introduction based on the scenario of a “Smart Agricultural Technology Company” (about 100 words). Add the following types of questions: 3 fill-in-the-blank questions, 1 table-filling exercise (3–4 rows), 2 short-answer questions, 1 matching exercise, and 1 scenario analysis question (describe a fault situation and have students analyze the cause). All the questions should be closely related to the practical training operations. Give the answer and a score. 5. Format Requirement: Leave a blank line between each question and use numbered labels for all items.

Take P1-T1 “Cybersecurity Concepts and Standards” as an example. In an AI-generated assignment, students need to map the three concerns mentioned by an employer in the AI’s answer— “designs are being stolen”, “the official website has been tampered with”, and “business operations have been disrupted”—to the CIA triad and then explain each objective in a single sentence to the employer. Thus, the abstract knowledge will be applied in practice, and students will be able to learn by taking on the role of a security consultant.

3.3. Intelligent generation and anti-replication mechanism for training manuals

The training manual should present the operating steps clearly and have anti-copying measures to ensure the authenticity of the training^[10].

Prompt Template: Please prepare a practical training manual for the “Firewall Technology” course on the “[Task Name]” (based on the eNSP simulator), and include the following sections: 1. Training goals (3-4 items); 2. Network Topology Diagram (Text Description Only); 3. Configuration Method (Step-by-step explanation); 4. CLI Configuration Commands (complete; add comments after important commands); 5. Web-based Configuration Method (short steps); 6. Verification Methods (at least two verification commands); 7. Anti-copy screenshot requirements: Open Notepad next to the eNSP window, show “Class + Name + Student ID”, and take a screenshot that shows both in the same window; 8. Training Reflection Questions (2 questions); 9. Training Report Submission Requirements.

Anti-copying measure: Open a new Notepad window next to the eNSP window, display “class + name + student ID”, take a full-screen screenshot of the whole window, and thus it is not possible to share screenshots for cheating.

3.4. Intelligent generation and multidimensional annotation of question banks

The question bank should have all sorts of difficulties, various kinds of questions, knowledge points, and supporting multi-dimensional tagging for convenient customization of test assembly^[2,4].

Prompt Template: Design a post-class quiz for the course “Firewall Technology” based on the task “[Task Name]” (total of 15 questions) according to the following requirements: 1. 5 multiple-choice questions (3 basic questions and 2 application-oriented questions); 2. 5 multiple-choice questions (each has at least 4 options, and among them, 2 to 4 should be correct); 3. 5 true/false questions (including 2 common misconceptions); 4. Each question should have a difficulty level (easy/medium/difficult), the corresponding knowledge point, and a brief explanation (30–50 words); 5. All the essential knowledge points of this task

should be covered in the content and linked to actual training.

Teachers review the questions generated by AI and, based on various reasons such as knowledge points, difficulty, question type, etc., add corresponding tags to build a dynamic question bank. Based on the data of incorrect answers, teachers can have the AI quickly create customized practice papers.

3.5. AI-assisted student self- and peer-assessment

After the students have finished writing their training reports, they can provide the information to the AI without any identifying information, and then the AI will give a score based on the reference grade and some suggestions for improvement in terms of the criteria. It should be pointed out that the AI does not grade the work; rather, it offers suggestions. This support reduces teachers' repetitive workload and promptly informs the students of any problems ^[5,10].

4. Implementation results and data analysis

Based on the above framework, 24 task resources have been added to the “Firewall Technology” course.

4.1. The development efficiency has risen

Table 2 shows the comparison of the efficiency of human-machine collaboration and traditional development models.

Table 2. Comparison of efficiency between human-machine collaboration and traditional development (average per task)

Resource type	Traditional development time(hour)	AI-generated draft time(minutes)	Teacher review iteration time (hours)	Total time savings rate
Theory-Practice Integrated Assignment	2.5	3–5	0.5	About 70%
Practical Training Manual	3	5–8	0.8	About 72%
Test Items (15 questions)	1.5	2–3	0.3	About 73%
Total for 24 Tasks	About 168 hours	—	About 38 hours	About 77%

Note: Savings Rate = (Traditional Development Time - Total AI-Assisted Time) / Traditional Development Time × 100%, where Total AI-Assisted Time = AI-Generated Draft Time (converted to hours) + Teacher Review and Iteration Time.

As shown in **Table 2**, the human-machine collaboration model reduced the development time of the 24 tasks from about 168 hours to around 38 hours and increased the efficiency by about 77%. The improvement is that AI has taken over the most time-consuming part—drafting—so teachers can focus on designing the framework and conducting high-quality reviews rather than writing word by word. Thus, the teachers' main responsibilities have shifted from “writing” to “reviewing and editing”, and the amount of repetitive labor has been significantly reduced ^[3,4].

4.2. Quality validation via peer review

Five peer teachers were invited to anonymously review the 24 AI-generated resources, using a five-point rating scale (1 = poor, 5 = excellent). The review results are shown in **Table 3**.

Table 3. Peer review results for resource quality (average score)

Evaluation dimensions	Theory-practice integrated operations	Training manual	Test questions
Accuracy of knowledge points	4.7	4.8	4.8
Contextual Relevance	4.6	4.5	—
Appropriate difficulty level	4.5	4.4	4.6
Effectiveness of anti-copying mechanisms	—	4.9	—
Overall usability	4.6	4.7	4.7

The average rating of all resources is above 4.4, so they are considered “good to excellent”. The score of the anti-copying mechanism in the training manual was relatively high, at 4.9; it was also mentioned that it had been “cleverly designed and was effective in preventing report plagiarism”^[10]. The accuracy of the knowledge points was also relatively high, at 4.7–4.8, so it can be concluded that the professional content produced by AI is generally reliable. The main reasons for the minor deductions were “some questions are slightly too difficult” and “the scenarios could be more detailed”; these comments have been used to improve the system in the next iteration.

4.3. The effect of individualized learning support on students

According to the data on students’ incorrect answers, teachers can have AI filter particular knowledge points and difficulty levels from the question bank and generate personalized practice sheets in minutes^[2]. Out of the 86 students surveyed, 78 percent reported that “personalized exercises have helped me focus on my weaknesses effectively”, so it can be seen that AI-assisted personalized resource generation can address the problem of “uniform exercises not being suitable for all individuals” in traditional classes. Resource Richness and Flexibility were given a satisfaction score of 4.5 out of 5. Given that there are frequent errors in the students’ training reports, the AI will also generate specific troubleshooting exercises, such as “forgetting to enable port translation in NAT address pools” or “reversing the direction of IPSec interest flows”, to address these common mistakes.

4.4. Student self-assessment and peer assessment as support tools

Teachers use AI to help grade students and reduce the time it takes to do so. 82% of the students have been able to identify problems in their reports with the help of AI feedback, and 76% believe that an AI-generated standard for peer review can make the assessment more objective. Therefore, this system can reduce the burden on teachers in grading and enhance students’ self-reflection skills and assessment literacy^[5,10].

5. Conclusion and outlook

The course “Firewall Technology” is taken as the case study in this paper, and we elaborate on how teachers and artificial intelligence can jointly develop theoretical-practical teaching materials.

5.1. Research results

According to the above research, human-machine collaboration has increased the efficiency of teaching resource development; that is, it is now possible to create 24 task-based resources in about 38 hours instead

of around 168 hours, an improvement of about 77%^[3,4]; AI-generated resources meet the requirements of teaching in terms of quality, and the average peer-review score is 4.4 or higher^[1,6]; it provides a fast way to create personalised learning materials, and 78% of the students found them helpful^[2,5]; AI can assist students in self-assessment and peer assessment to reduce the workload of teachers^[10].

This study contributes several innovations. It builds a full-chain intelligent resource generation framework that fills a systematic gap in human-machine collaborative resource generation within vocational education. It also proposes a teacher-led prompt engineering approach, creating a reusable prompt template library^[4]. Additionally, it integrates an anti-plagiarism mechanism into the practical training manuals^[10]. Finally, it explores a new model of AI-assisted self- and peer-assessment for students.

This study has several limitations. Only one course has been offered, and the sample size is relatively small. AI-generated content may have errors that need to be corrected by teachers, and good prompts are required for human-machine interaction^[5]. The above method can be used in other computer networking courses and more general engineering practice courses; its main framework is discipline-independent, the way of prompt engineering is generalizable, and the anti-plagiarism mechanism can be applied to any subject that requires lab report submission. In the future, we could develop a semi-automated prompt-generation tool, explore AI-assisted automated review, and expand this model to more courses^[2,5].

Funding

2025 Guangdong Provincial Vocational and Technical Education Association Research Planning Project: “Research on the Trusted Implementation of OBE Four-Chain Closed Loop Empowered by Full-Process Informatization” (Project No.: 202509G301)

Disclosure statement

The authors declare no conflict of interest.

References:

- [1] Wang Y, Kan D, 2026, Research on the Construction of Teaching Resources for Graphic and Image Processing Courses Empowered by AI. *Computer Knowledge and Technology*, 22(12): 114–117.
- [2] Cao R, Hu B, Guo K, et al., 2026, Design and Application of a Multimodal Large Model-Empowered Personalized Test Generation Platform. *Computer Education*, 2026(03): 277–282.
- [3] Wang G, 2026, Application of DeepSeek Technology in Enhancing the Efficiency and Quality of Teachers’ Lesson Preparation. *Science and Technology and Innovation*, 2026(02): 237–239.
- [4] Chen L, 2025, Construction and Optimization Research on an Intelligent Content Generation Framework for Higher Vocational Computer Courses Based on Key LLM Technologies. *Journal of Hunan Post and Telecommunication College*, 24(04): 30–34.
- [5] Li M, Wei J, 2024, Applications and Challenges of Generative Artificial Intelligence in Vocational College Education and School Governance. *Journal of Guangdong Agriculture Industry Business Polytechnic*, 40(04): 53–58.
- [6] Ou Z, Liu Y, Qin K, et al., 2024, Generation and Evaluation of Artificial Intelligence Multimodal Teaching

Resources—Based on the Application of AIGC in International Chinese Language Education. *Modern Educational Technology*, 34(09): 37–47.

- [7] Jiang D, 2017, Structural Logic of Work Process-Oriented Curriculum. *Education and Vocation*, 2017(13): 5–12.
- [8] Zhang Z, Xu X, Yang W, et al., 2022, Action-Oriented Teaching: Understanding, Practice, and Strategies—A Survey of 650 Vocational School Teachers from Seven Provinces and Municipalities. *China Vocational and Technical Education*, 2022(35): 33–41.
- [9] Yang Y, Lin X, 2021, The Integration of Four Dimensions: The Rational Logic of Action-Oriented Teaching Method. *Education and Vocation*, 2021(02): 80–87.
- [10] Cao Q, 2017, A Study on the Flipped Classroom Teaching Mode of Theory-Practice Integration Courses in Higher Vocational Colleges. *Higher Vocational Education (Journal of Tianjin Vocational Institute)*, 26(03): 74–77+83.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Human-AI Collaborative Decision Model for Low-Voltage Diagnosis and Closed-Loop Mitigation in Distribution Transformer Areas

Biwei Li¹, Caiyu Zhang^{2*}, Xianwen Zhang¹, Yexuan Liang¹

¹Zhaoqing Power Supply Bureau, Guangdong Power Grid Co., Ltd., Zhaoqing 526000, Guangdong, China

²Guangdong Technology College, Zhaoqing 526100, Guangdong, China

Corresponding author: Caiyu Zhang, 1509201014@qq.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Low-voltage mitigation in distribution transformer areas is rarely a single-step technical problem. Field diagnosis, plan selection, and post-implementation review are often handled separately, limiting the reuse of operational evidence and professional judgment. We recast mitigation as a six-stage human-AI closed loop linking state sensing, causal diagnosis, collaborative decision-making, implementation, effectiveness evaluation, and knowledge updating. For each candidate measure, the model considers data reliability, AI diagnostic confidence, human judgment reliability, and implementation risk. An expected-loss criterion assigns a machine-led, human-AI collaborative, or human-led mode. Safety and engineering constraints screen the available measures, after which an uncertainty-aware score identifies the preferred plan-mode pair. Evidence collected after implementation updates causal probabilities and the cause-measure knowledge base. Three constructed scenarios illustrate how decision authority moves from machine-led analysis toward professional control as implementation risk rises or diagnostic evidence weakens. These cases establish the internal consistency of the decision logic; they do not demonstrate improved field accuracy. The framework provides a transparent basis for subsequent calibration and validation with operational data.

Keywords: Distribution transformer area; Low-voltage mitigation; Human-AI collaboration; Decision support; Closed-loop feedback; Artificial intelligence

Online publication: Sept 3, 2026

1. Introduction

Low voltage in distribution transformer areas degrades supply quality, increases losses, and complicates refined operation and maintenance. It also varies across time and location because several factors may interact. Long supply radii, undersized conductors, unsuitable tap positions, and phase imbalance are among the recurrent causes, while violation limits should follow the applicable supply-voltage standard ^[1,2].

Recent work has used customer voltage, load, topology, and environmental data to mine violation patterns, forecast low-voltage events, analyze transformer voltages, and identify low-voltage network topology^[1,3–6]. Intelligent fusion terminals and cloud–edge coordination further support multisource processing^[7,8]. Most studies, however, concentrate on state identification or forecasting; the link from diagnosis to mitigation decisions and outcome feedback remains underdeveloped.

Mitigation decisions must also satisfy safety rules, cost limits, outage arrangements, equipment ratings, and engineering constraints. Machine learning can help grid operators classify events and prepare response options, yet explanation quality, confidence calibration, and generalization to unusual conditions remain concerns^[9–12]. A fixed ‘AI recommendation–human confirmation’ workflow likewise fails to reflect the different risks and authority boundaries of individual measures^[13–15].

This study therefore proposes a six-stage human–AI closed loop. It introduces a dynamic task-allocation rule driven by data reliability, AI diagnostic confidence, human judgment reliability, and implementation risk, together with a plan-evaluation and feedback mechanism that accounts for diagnostic uncertainty. Constructed scenarios are used to test the model’s internal logic under different inputs rather than to claim an improvement in field accuracy.

2. Low-voltage mitigation process and problem definition

2.1. Existing mitigation process

Low-voltage cases are usually triggered by monitoring violations, customer complaints, or operational inspections. Utilities then retrieve data, inspect the site, infer causes, formulate measures, obtain approval, implement the work, and verify the result. This sequential workflow relies heavily on manual records, leaving weak structural links among diagnostic evidence, plan selection, and observed outcomes.

2.2. Causes and multisource data requirements

The main causes can be grouped into source-, transformer-, line-, load-, customer-, and operating-mode factors, which may also act jointly^[1]. Diagnosis requires voltage measurements, transformer loading and tap positions, line parameters, three-phase currents, power factor, low-voltage topology, and maintenance records^[4,7,8]. **Table 1** organizes these elements into a cause–feature–data–measure mapping for probabilistic diagnosis and candidate-plan generation.

Table 1. Mapping of low-voltage causes, observable features, data sources, and candidate mitigation measures

Cause category	Typical cause	Core diagnostic evidence	Main data	Candidate measures and baseline human involvement
Source	Low upstream voltage	Synchronous low voltage across related areas; both near and remote points are depressed	Upstream bus, transformer outlet, and related-area voltages	Recommend upstream voltage regulation, reactive support, or operating-mode changes; human approval for upstream-grid actions
Transformer	Overload, insufficient capacity, or unsuitable tap	Voltage follows loading, or remains area-wide low under light load	Capacity, loading, tap position, and outlet voltage	Capacity expansion, load transfer, or tap adjustment; human approval for investment, outages, and equipment operation

Line	Long radius, undersized conductor, or local impedance anomaly	Voltage declines with distance or drops abruptly between adjacent points	Low-voltage topology, line parameters, end voltages, and currents	Shorten the radius, reinforce the line, or inspect abnormal joints; field verification before engineering work
Load	Phase imbalance, concentrated growth, or insufficient reactive power	Marked phase differences; low voltage tracks load period or reactive demand	Three-phase voltage/current, P/Q, power factor, and load curves	Rephase customers, shift load, or optimize reactive compensation; human review beyond authorized control limits
Customer	Impact load, service-wire issue, or customer-side equipment anomaly	Low voltage is localised and correlated with load switching or customer-side drop	Customer voltage, current, power curves, and work orders	Revise connection, reinforce service wire, or repair equipment; verify load type and ownership boundary on site
Operating mode	Unsuitable load allocation, topology, or equipment mode	Low voltage follows transfer, switching, or voltage-control device changes	DMS/OMS records, area loading, topology, and equipment status	Optimize service area, load transfer, or device mode; human approval for network operations
Combined state	Coupled factors	Several hypotheses have similar probabilities, and no single cause explains the case	Integrated multisource input	Generate combined or staged plans; raise baseline human involvement, with the final mode set by the allocation model

Note. Observable features are evidence for causal inference, not deterministic rules, and the listed measures are candidates rather than final decisions. Human-intervention requirements are operational priors; the actual mode is determined jointly by data reliability, AI confidence, human judgment reliability, and implementation risk.

2.3. Limitations of conventional manual mitigation

Professional staff provide indispensable field knowledge and safety oversight. Even so, cross-system data are costly to associate, coupled causes are difficult to distinguish, experience is hard to reuse systematically, and candidate measures often lack a structured comparison of effectiveness, economy, and implementation risk^[1,7]. The aim is therefore not to replace practitioners, but to improve multisource integration and sustained reuse of professional knowledge.

2.4. Need for a human–AI closed loop

Machines are well suited to data processing, anomaly screening, probabilistic diagnosis, and plan ranking; professionals are better placed to verify field evidence, review safety constraints, judge engineering feasibility, and assume decision responsibility^[14,15]. Decision authority should therefore vary with data quality, diagnostic confidence, and plan risk, while observed outcomes should return to the diagnostic and decision process.

3. Human–AI closed-loop architecture for low-voltage mitigation

3.1. Overall architecture

To connect the otherwise fragmented stages of low-voltage mitigation, the proposed architecture comprises state sensing, cause diagnosis, collaborative decision-making, measure implementation, effectiveness evaluation, and knowledge updating (**Figure 1**). Operational flows connect the stages, multisource data support diagnosis and evaluation, and measured outcomes and human feedback update the knowledge state^[7,8,13].

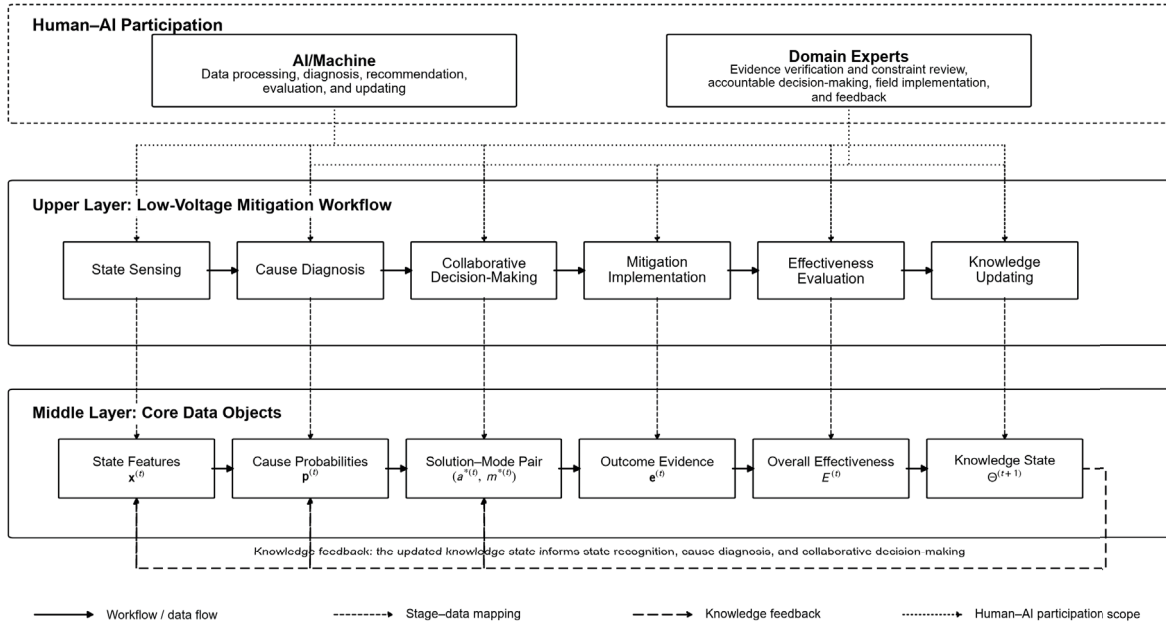


Figure 1. Overall architecture of human–AI collaborative closed-loop low-voltage mitigation in distribution transformer areas.

Note. Solid arrows denote the operational flow, dotted arrows map stages to core data objects, and the feedback loop carries updated knowledge. Knowledge supports the next round of identification, diagnosis, and decision-making but does not alter the physical measurements directly.

3.2. Six-stage closed-loop process

At mitigation round t , sensing produces a feature vector; diagnosis returns a probability distribution; collaborative decision-making selects the best plan–mode pair; implementation and evaluation produce outcome evidence; and knowledge updating revises the basis for subsequent diagnosis and decisions:

$$\mathbf{x}^{(t)} \rightarrow \mathbf{p}^{(t)} \rightarrow (a^{*(t)}, m^{*(t)}) \rightarrow \mathbf{e}^{(t)} \rightarrow E^{(t)} \rightarrow \Theta^{(t+1)} \quad (1)$$

The next diagnosis is driven jointly by new measurements and the updated knowledge state:

Here, $\mathbf{x}^{(t+1)}$ is obtained from the next set of measurements and state sensing, whereas $\Theta^{(t+1)}$ contains revised causal knowledge, mitigation rules, plan outcomes, and records of human–AI collaboration. Together they form the input to the next diagnostic round.

$$\{\mathbf{x}^{(t+1)}, \Theta^{(t+1)}\} \rightarrow \mathbf{p}^{(t+1)} \quad (2)$$

3.3. Functional boundaries between humans and AI

Human and machine roles are not reduced to a fixed ‘recommend–confirm’ pattern. They are assigned dynamically according to information reliability, implementation risk, and operational authority, as summarised in **Table 2** ^[11,12,14,15]. Machine-led operation is limited to authorized, low-risk analytical or operational adjustments. Outages, investment, field electrical work, and network reconfiguration remain

subject to established dispatch, maintenance, and safety procedures.

Table 2. Human and AI responsibilities in the main stages of low-voltage mitigation

Stage	AI responsibilities	Professional responsibilities	Collaborative output
Cause diagnosis	Analyze multi-source features, detect anomalies, estimate causal probabilities and confidence, rank evidence	Verify field evidence, supply missing information, identify exceptional factors, correct abnormal diagnoses	Causal probabilities, key evidence, and correction records
Mitigation decision	Generate candidates, flag constraint conflicts, score plans and modes; handle authorized low-risk tasks	Set authority boundaries and review safety and engineering constraints when human involvement is triggered	Plan-mode pair (a^* , m^*)
Outcome feedback	Calculate effectiveness and prediction error, identify re-diagnosis conditions, propose updates	Confirm actual outcomes, explain deviations, and record revisions or rejections	Outcome evidence, knowledge updates, and re-diagnosis information

3.4. Information exchange and feedback

An AI module should report more than a single diagnosis. Its output should include causal probabilities, key evidence, candidate measures, constraint conflicts, and uncertainty. Human confirmations, revisions, rejections, and reasons should be recorded in structured form alongside implementation records and measured outcomes. Feeding these records back to the knowledge base makes the interaction explainable, correctable, and capable of continued refinement ^[10,11,13].

4. Low-voltage state identification and cause diagnosis

Cause diagnosis links state sensing with mitigation decisions. Rather than training a particular learning algorithm, this study defines a common interface from multisource features to causal probabilities and diagnostic confidence for downstream plan generation and task allocation.

4.1. Low-voltage state representation

Voltage violations are identified against the relevant supply-voltage standard, while their magnitude, duration, and affected range describe the operating state ^[1,2]. The feature vector includes minimum voltage, violation duration and frequency, share of affected users, voltage and current imbalance, transformer loading, power factor, and time-of-occurrence:

$$\mathbf{x} = \left[U_{\min}, T_{lv}, N_{lv}, \rho_u, \eta_U, \eta_I, K_T, \cos \varphi, \tau_t \right] \quad (3)$$

where τ_t denotes the occurrence period or its time category. All indicators are standardized before diagnosis; no single violation value is treated as sufficient evidence of cause.

4.2. Cause classification

The taxonomy distinguishes six physical categories: source, transformer, line, load, customer, and operating mode. Examples include low upstream voltage, transformer overloading, unsuitable tap settings, long supply radii, undersized conductors, local impedance anomalies, phase imbalance, insufficient reactive power, and time-concentrated load growth ^[1]. Voltage regulation, reactive compensation, phase balancing, storage dispatch, and network reinforcement address different mechanisms ^[16–20]. The six categories may occur alone

or in combination; a coupled case is not treated as a seventh physical category. See **Figure 2**.

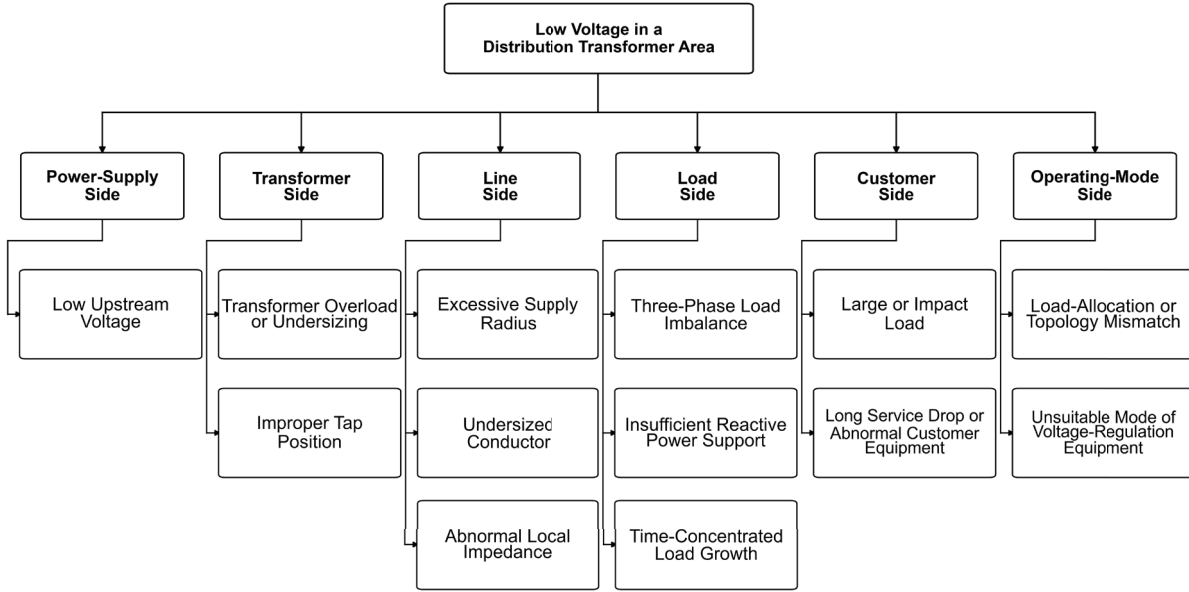


Figure 2. Classification of low-voltage causes in distribution transformer areas.

4.3. Mapping multisource features to causes

No single measurement uniquely identifies a low-voltage cause. Electrical quantities, asset parameters, low-voltage topology, and maintenance records must be considered together ^[3–8]. A normal sending-end voltage accompanied by distance-dependent decline points towards a line factor; coincident peak loading and area-wide voltage decline support transformer overloading; and phase-specific violations warrant checking phase imbalance. **Table 1** supplies diagnostic evidence, while supervised learning, probabilistic inference, or knowledge-based reasoning may implement the mapping. No particular algorithm is prescribed here.

4.4. Causal probabilities and AI diagnostic confidence

Let $C = \{C_1, C_2, \dots, C_n\}$ denote mutually exclusive diagnostic-state hypotheses. A hypothesis may represent one basic cause or a predefined combination of causes. Given $\mathbf{x}$, the diagnostic module returns the posterior probability of each hypothesis:

$$p_j = P(C_j | \mathbf{x}), \sum_{j=1}^n p_j = 1 \quad (4)$$

The leading diagnostic state is $C^* = \arg \max_j p_j$.

To reflect both the advantage of the leading hypothesis and the concentration of the full distribution, this study defines AI diagnostic confidence as:

$$q = \alpha \max_j p_j + (1 - \alpha) \left(1 + \frac{\sum_{j=1}^n p_j \ln p_j}{\ln n} \right), \quad 0 \leq \alpha \leq 1 \quad (5)$$

The convention $0 \ln 0 = 0$ is used, and $n \geq 2$. The second term equals one minus normalized entropy; it approaches 1 as the distribution becomes more concentrated. Thus q measures diagnostic certainty for the current input rather than empirical accuracy^[10–12,21]. This confidence measure is a theoretical interface whose parameter α and functional form require calibration with historical diagnostic data. Section 3 therefore provides $x \rightarrow (p, q)$; Section 4 combines p and q with data reliability d , human judgment reliability h , and plan risk r_i to obtain (a^*, m^*) .

5. Human–AI collaborative mitigation decision model

For each candidate measure, the model combines data reliability, AI diagnostic confidence, human judgment reliability, and implementation risk to assign decision authority and select the best feasible plan–mode pair.

5.1. Data, machine, and human reliability

Data reliability d , AI confidence q , and human judgment reliability h are normalized to $[0, 1]$. Completeness, timeliness, consistency, and measurement credibility are aggregated as:

$$d = \sum_{k=1}^K \alpha_k d_k, \quad \sum_{k=1}^K \alpha_k = 1 \quad (6)$$

Here d_k and α_k are the k th data-quality indicator and its weight. The product dq is the effective machine reliability after accounting for data quality. Human reliability h is not an administrative score; it represents agreement between prior professional judgments and subsequent outcomes in comparable cases, the match between expertise and the current task, and the adequacy of field verification^[11,12,14]. To avoid double counting, h excludes completeness, timeliness, and consistency already represented by d .

5.2. Expected loss under alternative decision modes

Building on the principle that automation should vary with task properties, the model considers machine-led (M), human–AI collaborative (HM), and human-led (H) modes^[14,15]. Their reliabilities are defined as:

$$\rho_M = dq, \quad \rho_H = h, \quad \rho_{HM} = \beta dq + (1 - \beta)h, \quad 0 \leq \beta \leq 1 \quad (7)$$

The coefficient β is the machine's relative contribution in collaborative decisions and can be calibrated by task and historical performance. The model does not assume that ρ_{HM} must exceed ρ_M or ρ_H . Collaboration combines the two information sources and, in higher-risk cases, uses human review to limit risk exposure. The expression is a calculable theoretical interface; its form and parameters require validation with historical mitigation records. For candidate plan a_i , the expected loss under mode m is:

$$L_m = C_e (1 - \rho_m) + C_s \gamma_m r_i + C_m \quad (8)$$

$$C_M = 0, \quad C_H = C_t, \quad C_{HM} = C_t + C_c \quad (9)$$

C_e and C_s denote the losses associated with an incorrect decision and safety risk; $r_i \in [0, 1]$ is the implementation risk of plan a_i ; and γ_m is the mode-specific exposure coefficient. If human review reduces procedural exposure, $\gamma_M > \gamma_{HM} \geq \gamma_H$. C_m is the incremental human and coordination cost above an existing intelligent system. Accordingly, $C_M = 0$ means zero incremental human cost, not cost-free computation; C_t and C_c represent human review and coordination costs. These loss terms are defined here for allocating

decision authority in low-voltage mitigation.

5.3. Candidate-plan generation

From the posterior probabilities $p_j = P(C_j | x)$, retain the diagnostic hypotheses that satisfy:

$$J = \{j | p_j \geq \tau_p\} \quad (10)$$

where τ_p is the screening threshold. The cause–measure mapping in Table 1 then generates:

$$A = \bigcup_{j \in J} A(C_j) \quad (11)$$

Candidates include tap adjustment, phase-load balancing, reactive compensation, load transfer, feeder reinforcement, and transformer capacity expansion. The AI produces a set rather than a single prescription. Hypotheses below τ_p are excluded from separate evaluation, and the probabilities retained in J are renormalized but continue to be denoted p_j [16–19].

5.4. Dynamic allocation of human and AI tasks

For each candidate, operational attributes, institutional rules, and baseline human-intervention requirements define the admissible mode set Ω_i . The selected mode minimizes expected loss:

$$m_i^* = \arg \min_{m \in \Omega_i} L_m \quad (12)$$

A hard constraint applies to high-investment plans, outage work, and measures that affect operational safety:

$$r_i \geq r_{high} \Rightarrow M \notin \Omega_i \quad (13)$$

Thus Ω_i and m_i^* are assigned to each measure rather than to a cause category in advance. Different measures for the same diagnosed cause may carry different authority settings [9,11,12,14].

5.5. Constraints and plan evaluation

Feasible plans A_f are first screened against nodal-voltage limits, conductor ampacity, transformer capacity, budget, allowable outage duration, and safe-operation requirements. Plan a_i then receives a score $S(a_i | C_j)$ under hypothesis C_j based on expected voltage improvement, economy, engineering feasibility, reliability benefit, and implementation risk. Accounting for diagnostic uncertainty, its expected score is:

$$\bar{S}_i = \sum_{j \in J} p_j S(a_i | C_j) \quad (14)$$

Let i^* identify the highest-scoring feasible plan. The selected plan and its associated mode are:

$$i^* = \arg \max_i: a_i \in A_f \bar{S}_i, \quad (a^*, m^*) = (a_{i^*}, m_{i^*}^*) \quad (15)$$

The final output is therefore the plan–mode pair (a^*, m^*) . Plan evaluation answers which measure should be adopted; dynamic task allocation determines how its decision authority should be shared. Expected loss configures authority, whereas the comprehensive score ranks engineering plans. They are not added, which avoids mixing different units or counting risk twice.

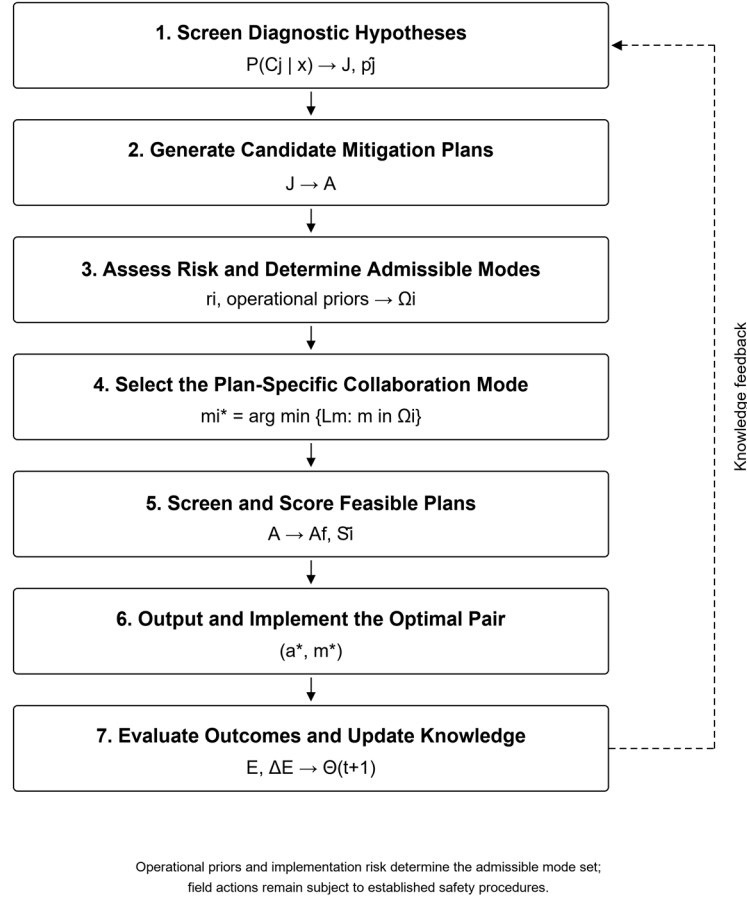


Figure 3. Human–AI collaborative decision-making process for low-voltage mitigation.

Note. ‘Decision-making and implementation under the selected mode’ refers to plan review and authority routing. Any field electrical work remains governed by existing dispatch, maintenance, and safety rules.

6. Effectiveness evaluation and knowledge updating

To move beyond one-off recommendations, observed mitigation outcomes are returned to diagnosis and decision-making through effectiveness evaluation, knowledge updating, and re-diagnosis.

6.1. Evaluation indicators

Effectiveness is evaluated in seven dimensions: voltage improvement, shorter violation duration, fewer affected users, lower network loss, reduced phase imbalance, mitigation cost, and residual operational risk^[1,16,17,19]. All indicators are normalized to [0, 1]. The first five are benefits, whereas cost and residual risk are penalties:

$$E = \mu_1 E_U + \mu_2 E_T + \mu_3 E_N + \mu_4 E_L + \mu_5 E_I - \mu_6 E_C - \mu_7 E_R$$

$$\sum_{k=1}^7 \mu_k = 1, \quad \mu_k \geq 0 \quad (16)$$

EU, ET, EN, EL, and EI denote improvements in voltage, violation duration, affected users, network loss, and phase imbalance; EC is mitigation cost; and ER is the operating-safety risk remaining after implementation. Hence, r_i in Section 4 is an ex ante implementation risk used for authority allocation, whereas ER is an ex post outcome indicator.

6.2. Judging plan effectiveness

Let $E_{min} < E_{tar}$. A plan meets its target when $E \geq E_{tar}$ and no key indicator deteriorates. It is partly effective when $E_{min} \leq E < E_{tar}$, or when voltage recovers while network loss, imbalance, or another key measure worsens. It fails when $E < E_{min}$. The prediction error is:

$$\Delta E = E_{act} - E_{pre} \quad (17)$$

6.3. Updating causal probabilities and knowledge rules

Following principles of calibrated trust, productive delegation, and correctable interaction, both outcomes and human feedback enter the update process^[12–15]. Let $e(t)$ be the multi-indicator evidence after round t . For the implemented plan a^* , causal probabilities are updated by Bayes' rule:

$$p_j^{(t+1)} = \frac{P(e^{(t)} | C_j, a^*) p_j^{(t)}}{\sum_k P(e^{(t)} | C_k, a^*) p_k^{(t)}} \quad (18)$$

$P(e(t) | C_j, a^*)$ is the likelihood of observing the current outcome if C_j holds and a^* is implemented. It should be estimated from comparable historical cases; expert elicitation may initialize the model when samples are scarce. Agreement between predicted and observed outcomes raises confidence in the corresponding cause–measure rule. A marked discrepancy lowers that confidence and triggers a check for omitted or coupled causes. Human confirmations, revisions, and rejections are also retained for rule correction and trust calibration.

6.4. Closed-loop re-diagnosis

Re-diagnosis is triggered when $E < E_{min}$, $|\Delta E| > \varepsilon$, or new low-voltage violations appear after mitigation. Additional information updates d , q , and $P(C_j | x)$ before the process repeats. Successful cases are stored as ‘scenario–cause–measure–mode–outcome’ records. The loop is not indefinite: it ends when the target is met, or a professional terminates the process; insufficient data suspend automation and prompt field review. Subsequent electrical work remains subject to established operating and safety rules.

7. Analysis of constructed scenarios

7.1. Scenarios and parameter settings

Three scenarios are constructed to test authority allocation under different data quality, diagnostic confidence, and implementation risk; no claim is made about AI accuracy. The settings are $\beta = 0.60$, $C_e = 1.00$, $C_s = 0.40$, $C_t = 0.10$, $C_c = 0.02$, $\gamma_M = 1.00$, $\gamma_{HM} = 0.25$, $\gamma_H = 0.20$, $h = 0.80$, and $rh_{high} = 0.70$. Expected plan scores in **Table 3** are normalized illustrative values based on effectiveness, economy, feasibility, reliability benefit, and risk. All parameters serve only to demonstrate model behavior and require calibration from historical cases, project costs, and decision records before field use.

Table 3. Human–AI collaborative decisions in typical low-voltage mitigation scenarios

Scenario and focal plan	Leading raw posterior(s)	d/q/h	Risk r_i	Admissible modes Ω_i	LM/LHM/LH	Selected mode m_i^*	Expected score $\bar{S}_i$
A1: Low-risk operational adjustment	Excessive supply radius 0.90	0.95/0.90/0.80	0.10	{M, HM, H}	0.185/0.297/0.308	M	0.70
A2: Low-voltage feeder reinforcement	Excessive supply radius 0.90	0.95/0.90/0.80	0.40	{HM, H}	—/0.327/0.332	HM	0.82
B: Transformer capacity expansion	Transformer overload 0.90	0.95/0.90/0.80	0.85	{HM, H}	—/0.372/0.368	H	0.84
C: Staged combined mitigation	Three-factor 0.38; imbalance–radius 0.34; radius–impedance 0.18	0.55/0.40/0.80	0.65	{HM, H}	—/0.733/0.352	H	0.79

Note. Causal probabilities are raw posteriors. Probabilities exceeding the threshold are renormalized as specified in Section 4.3. Risk, loss, and plan scores are constructed values rather than field statistics or engineering accuracy. A dash indicates a mode excluded by operational priors. Machine-led mode covers only authorized low-risk analysis or operational adjustment, not direct field switching. Scenario C reports the three leading mutually exclusive diagnostic hypotheses; the remaining hypotheses sum to 0.10, and a hypothesis may itself contain several basic causes.

7.2. Single-cause scenario

In Scenario A, the sending-end voltage is normal, but remote users remain undervoltage; the posterior probability of an excessive supply radius is 0.90^[1]. Low-risk operational adjustment admits all three modes and selects M because LM is the smallest. Feeder reinforcement has engineering consequences, so M is excluded and HM is selected. Its higher expected score yields the final pair ‘feeder reinforcement–human–AI collaboration’. The example shows that two measures for the same cause can carry different authority settings.

7.3. High-risk mitigation scenario

Scenario B has diagnostic inputs close to Scenario A, but transformer capacity expansion entails investment, outage work, and a change in supply arrangements, raising r_i to 0.85. The hard constraint removes M. The remaining losses are LHM = 0.372 and LH = 0.368, so H is selected. Their difference of 0.004 places the case near the HM–H boundary; under the illustrative parameters, human-led decision-making is only slightly preferred. High AI confidence therefore indicates a concentrated diagnosis, not independent authority over a high-risk plan. The boundary must be calibrated from operational records and safety requirements.

7.4. Coupled causes and low confidence

Scenario C combines phase imbalance, an excessive supply radius, and a local impedance anomaly with incomplete measurements. Among predefined mutually exclusive states, the posteriors are 0.38 for the three-factor combination, 0.34 for imbalance plus excessive radius, and 0.18 for excessive radius plus local impedance. The similar values indicate that the available evidence cannot yet distinguish the coupled state. With $d = 0.55$ and $q = 0.40$, effective machine reliability is:

$$\rho_M = dq = 0.55 \times 0.40 = 0.22$$

Collaborative reliability is:

$$\rho_{HM} = 0.60 \times 0.22 + 0.40 \times 0.80 = 0.452$$

The resulting losses, LHM = 0.733 and LH = 0.352, strongly favour H. When data quality is poor and competing compound hypotheses remain unresolved, the model deliberately shifts authority towards professionals. Field checks of conductor size, phase allocation, and connection condition can then update d , q , and $P(C_j | x)$, after which the AI regenerates and reranks the plans. This sequence—human inspection, data completion, AI re-diagnosis, and plan re-evaluation—closes the loop ^[1,10–12,17].

7.5. Comparison of collaboration modes

Table 3 traces an $M \rightarrow HM \rightarrow H$ shift. Machines carry more analytical and decision work when information is reliable, and plan risk is low. Engineering consequences move authority to collaboration; greater implementation risk or diagnostic uncertainty moves it further towards professionals. These results establish logical consistency under the stated inputs, not higher diagnostic accuracy or field effectiveness. Loss parameters, risk thresholds, evaluation weights, and mode boundaries still require calibration with real operating and mitigation data.

8. Conclusions and outlook

A six-stage human–AI closed loop was developed for low-voltage mitigation in distribution transformer areas. The model allocates tasks based on data reliability, AI confidence, human judgment reliability, and implementation risk, and it links uncertainty-aware plan evaluation with outcome-based knowledge updating. Constructed scenarios demonstrate coherent shifts in decision authority as information quality and risk change. The work remains a theoretical and logical validation: human reliability, loss parameters, and evaluation weights require real business data. Future work will test the model on operating data, train diagnostic algorithms, examine parameter sensitivity, and develop a decision-support platform.

Author biographies

Biwei Li (born 1996) received the M.S. degree and is an engineer. His research focuses on distribution-network fault management. Email: 498315697@qq.com. Corresponding author: Caiyu Zhang (born 1994) received the M.S. degree. Email: 1509201014@qq.com

Disclosure statement

The authors declare no conflict of interest.

Reference

- [1] Shen S, Zhong Q, Xu Z, et al., 2025, Mining User Low-Voltage Violation Patterns Using Hierarchical Affinity-Propagation Clustering. *Electric Power Engineering Technology*, 44(1): 30–38.

- [2] General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of China, 2008, Power Quality—Deviation of Supply Voltage. Standards Press of China.
- [3] Guo S, Zhao X, Sun G, et al., 2025, Analysis and Prediction of Over-Limit Characteristics for Regional Distribution-Transformer Voltage Using Bilayer Clustering and Correlation-Feature Screening. *Southern Power System Technology*, 19(2): 19–27.
- [4] Huang Y, Yang S, Cao X, et al., 2023, Topology Identification of Low-Voltage Distribution Networks Using Segmented Current Features and a Random-Forest Algorithm. *Power Demand Side Management*, 25(2): 63–69.
- [5] Wei J, Tian S, Chen W, et al., 2025, Low-Voltage Prediction in Distribution Transformer Areas Using a GA–BP–LSTM Model. *Electric Engineering*, 46(1): 15–20.
- [6] Zhou Y, Zhang B, Huang W, et al., 2023, Low-Voltage Prediction for Distribution Transformer Areas Using an LSTM–BP Combined Model. *Journal of Electric Power Science and Technology*, 38(5): 177–186.
- [7] Chen W, Chen J, Fan Y, et al., 2024, Design and application of smart distribution station areas based on intelligent fusion terminals with cloud-edge coordination. *Electric Power*, 57(4): 190–199.
- [8] Fan Y, Wu H, Li Z, et al., 2024, A flexible orchestration of lightweight AI for edge computing in low-voltage distribution network. *Frontiers in Energy Research*, 12: 1424663.
- [9] Campos V, Klyagina O, Andrade J, et al., 2024, ML-Assistant for Human Operators Using Alarm Data to Solve and Classify Faults in Electrical Grids. *Electric Power Systems Research*, 236: 110886.
- [10] Machlev R, Heistrene L, Perl M, et al., 2022, Explainable artificial intelligence techniques for energy and power systems: review, challenges and opportunities. *Energy and AI*, 9: 100169.
- [11] Naiseh M, Al-Thani D, Jiang N, et al., 2023, How the Different Explanation Classes Impact Trust Calibration: The Case of Clinical Decision Support Systems. *International Journal of Human-Computer Studies*, 169: 102941.
- [12] Yamada S, 2022, Trust Calibration as Rationality for Human–AI Cooperative Decision Making. *Cognitive Studies*, 29(3): 364–370.
- [13] Amershi S, Weld S, Vorvoreanu M, et al., 2019, Guidelines for Human–AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.
- [14] Fügener A, Grahl J, Gupta A, et al., 2022, Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, 33(2): 678–696.
- [15] Parasuraman R, Sheridan T, Wickens C, 2000, A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 30(3): 286–297.
- [16] Guo X, Zheng F, Wang S, 2024, Mobile Energy-Storage Scheduling in Distribution Networks Considering Voltage-Violation Risk. *Power Demand Side Management*, 26(4): 37–42.
- [17] Li B, Li Y, Hai Z, et al., 2025, Online Mitigation of Three-Phase Imbalance in Distribution Networks Using Multi-Agent Deep Reinforcement Learning. *Proceedings of the CSEE*, 45(5): 1729–1741.
- [18] Yang Y, Wen F, Zhou X, et al., 2023, Review of Coordinated Voltage Control in Distribution Systems with High Photovoltaic Penetration. *Electric Power Automation Equipment*, 43(10): 48–58.
- [19] Yuan B, Yang P, Wang T, et al., 2025, Multi-Objective Dispatch of Distributed Energy Storage in High-PV Distribution Transformer Areas Considering Multiple Stakeholders. *Modern Electric Power*, 42(6): 1268–1277.
- [20] Zhang X, Wu Z, Sun Q, et al., 2024, Application and Progress of Artificial Intelligence Technology in the Field of Distribution Network Voltage Control: A Review. *Renewable and Sustainable Energy Reviews*, 192: 114282.

- [21] Schemmer M, Kühl N, Benz C, et al., 2023, Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of the 28th ACM International Conference on Intelligent User Interfaces, 410–422.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on the Application Path of Dual-Mode API Data Market for Enabling Cross-Domain Multi-Scenario Data Circulation

Mutong Chen, Xiong Ni, Chuqiao Xiao

Shanghai Municipal Big Data Center, Shanghai 200072, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: In response to the industry pain points such as scattered data sources, inconsistent standards, lack of trust, restricted high-sensitive data circulation, high costs for small and medium-sized entities to access, and the limitations of the traditional centralized and point-to-point single data circulation model in cross-domain scenarios, this paper combines the theories of data market, data space, and data intermediary, and proposes a “point-to-point + adapter” dual-mode API data market circulation path. By constructing a core architecture that separates the management flow from the data flow, the point-to-point mode ensures the local controllable circulation of highly sensitive data, and the adapter mode reduces the access threshold for existing systems and small and medium-sized entities. This paper builds an exclusive application framework and studies its migration paths in areas such as cargo trade document verification, cross-departmental government approval, financial joint risk control, supply chain collaboration, and industrial internet. The research shows that this model can achieve trusted, compliant, and large-scale cross-domain data circulation without aggregating the original data, providing a feasible infrastructure solution for the release of value of data elements in multiple scenarios.

Keywords: Dual-mode API data market; Cross-domain data circulation; Adapter; Data space; Data intermediary; Shipping trade; Separation of management flow and data flow

Online publication: Sept 3, 2026

1. Introduction

The full realization of the value of data elements relies on efficient and trustworthy circulation across entities, systems, and domains. This is also the core support for the high-quality development of the digital economy and the digital transformation of industries. Currently, global data circulation governance has evolved from simple technical optimization to a comprehensive system that integrates technical architecture, rights and responsibilities systems, market mechanisms, and scenario governance. Simply relying on technological transformation is no longer sufficient to solve the problems of large-scale, compliant, and ecological data

circulation. The EU's "Data Governance Act" and "Data Act" clearly define the institutional value of data intermediaries and establish rules for data ownership, circulation, and intermediary services. The relevant reports of the OECD and the World Bank also clearly point out that insufficient trust among entities, incompatible heterogeneous systems, and imbalanced incentives for data supply and demand are the core bottlenecks restricting the efficient reuse and value realization of data elements.

Shipping trade is the most complex typical scenario for cross-domain data circulation. It involves multiple entities such as cargo owners, shipping companies, ports, customs, and financial institutions, covering various types of heterogeneous data such as bills of lading, customs declaration forms, AIS trajectories, and letters of credit. It also has characteristics of cross-border, strict supervision, and high sensitivity. The data circulation problems in shipping trade are widely present in fields such as government affairs, finance, supply chain, and industrial internet.

The traditional data circulation model has obvious shortcomings: the pure point-to-point direct connection has a high degree of customization and is difficult to manage on a large scale, and it is prone to form data silos; the centralized data platform is likely to cause the aggregation of raw data, leading to compliance risks, data leakage, and unclear rights and responsibilities. Based on this, this paper constructs a "point-to-point + adapter" dual-mode API data market system, focusing on solving core issues such as cross-domain heterogeneous data circulation, high-sensitivity data security control, and low-cost access by small and medium-sized entities, and exploring its multi-scenario implementation paths and promotion value.

2. Related research review

2.1. Data market and API economy

The core of the data market lies in establishing a system for matching data supply and demand, productizing data, setting prices, conducting transactions, and governing the system. It aims to address the issue of insufficient incentives for data sharing and the dispersion of resources, and to promote the transition of data from raw data sets to ecological delivery of data services and analytical capabilities. The API economy, through standardized interfaces, enables the service and productization of enterprises' capabilities and data, serving as the core support for cross-system and cross-entity interconnection. The combination of these two forms a new type of data circulation infrastructure, the API data market, which is not merely an interface display platform, but integrates authorization, invocation, measurement, auditing, and security control.

2.2. Data space and data intermediary

The data space adopts the lightweight integration concept of "on-demand integration and stepwise iteration", breaking the traditional barrier of unified global modeling. Practices such as European IDSA and Gaia-X have further constructed a distributed data sovereignty, interoperability, and usage control system. The data intermediary, as a core institutional innovation, has a core value that lies in neutral mediation, reducing transaction costs, and avoiding risks in the centralized original data. It only undertakes governance functions such as directory management, authorization verification, and audit evidence storage, without touching the original data. This is also the core theoretical support of the dual-mode architecture in this paper^[1].

2.3. Cross-domain security and industry standards

The zero-trust architecture and the ISO series of information security standards provide paradigmatic support

for boundary penetration and security control across private networks, clouds, and the internet. The DID/VC distributed identity system realizes decentralized and traceable authorization verification. The DCSA data standard in the shipping and trade field and the UNCITRAL electronic transferable record model law provide technical and legal support for the standardized circulation of industry data and the legal validity determination of electronic documents, and offer a compliance basis for the implementation of the dual-mode mechanism scenarios.

3. Analysis of issues in cross-domain multi-scenario data circulation

3.1. Multi-source heterogeneity and prominent network boundary barriers

Industry data is scattered across government and private networks, financial intranets, enterprise private clouds, and various business systems. Data formats include API interfaces, messages, document files, real-time device streams, and many other types, exhibiting strong heterogeneity. Networks of different entities are isolated from each other. Forcing unified interfaces and data aggregation would incur extremely high transformation costs and compliance resistance, making it difficult to adapt to large-scale circulation needs.

3.2. Lack of trust among entities and ambiguous definition of rights and responsibilities

The core pain points of data circulation are “not daring to share, not willing to share, and not knowing how to share”. Government departments are concerned about unauthorized access and liability risks, enterprises are wary of the leakage of business secrets and the abuse of data by competitors, and financial institutions are concerned about the authenticity of data and compliance responsibilities. Traditional platforms only solve the problems of “discoverability and connectivity” of data, but they cannot clearly define the authorization entity, usage scope, and accountability basis, and it is difficult to build a trustworthy circulation system ^[2].

3.3. Single model cannot adapt to multi-level data security requirements

Cross-domain data has a clear hierarchical security classification: for instance, high-sensitivity data such as customs declaration forms, financial credit, and core orders must not be centrally transferred; while low-sensitivity data such as logistics status, public directories, and de-identified data aim for efficiency in circulation. The single point-to-point model is prone to forming data silos, and the single centralized model amplifies security and compliance risks. The industry urgently needs differentiated and scenario-based circulation mechanisms.

3.4. Weak access capability of small and medium-sized participants

At present, the construction of the domestic data ecosystem is mostly led by large leading enterprises and various public departments at all levels. The industry resources, technical capabilities, and data access rights are highly concentrated. However, the numerous small and medium-sized freight forwarders, small and medium-sized manufacturing enterprises, grassroots service providers, and small and medium-sized trading enterprises in the industry generally lack professional API connection operation and maintenance capabilities, data standardization governance, and cybersecurity compliance control capabilities. They also have limited technical reserves and capital investment. The strict access standards, complex adaptation processes, and high renovation costs of traditional data platforms have led to the data market being confined to a closed circulation among a few large entities for a long time. It fails to form an all-encompassing industrial network

effect, severely restricting the large-scale circulation and value release of data elements throughout the entire industrial chain ^[3].

4. Mechanism design of the dual-mode API data market

4.1. Overall architecture

The dual-mode API data market constructed in this paper adopts a hierarchical architecture, consisting of six major modules: data directory layer, authorization governance layer, API access layer, secure transmission layer, audit measurement layer, and operation service layer. The core innovation lies in the separation of management flow and data flow. The platform only carries management flow data such as directory registration, identity authentication, authorization management, invocation auditing, and measurement evidence storage. The original data flow flows between the supply and demand parties in accordance with the model, without passing through the platform for transfer. This fundamentally avoids the risk of data concentration and guarantees the sovereignty of the data holder ^[4].

4.2. Point-to-point mode: Trusted direct connection for highly sensitive data

This model is applicable to high-sensitivity, highly compliant, and high-value scenarios such as financial credit granting, customs declaration verification, and core supply chain data. In terms of the process, the data user discovers the data product on the market platform and submits an application; the data holder reviews the data usage, calling scope and security conditions; the authorizer completes the authorization confirmation through DID/VC or other trusted credentials; the platform synchronizes the authorization status and calling credentials; the data user sends a request to the data holder's API with the authorization code; the data holder verifies the authenticity and validity period of the authorization through the platform or local verification service; after verification is successful, the holder responds with the data or results in the original path; the platform only records the calling summary, authorization status and necessary audit information. The advantage of this model lies in that the data does not pass through the platform, and sensitive information is not concentrated; the data can be controlled in terms of access and risk management; supervision can be traced through chain summaries, hashes, and authorization status. The disadvantage is that it requires certain API integration capabilities from the participating parties, and it is more suitable for high-value data circulation among large institutions, financial entities, and leading enterprises ^[5].

4.3. Adapter mode: Lightweight access for legacy systems

This model is designed for scenarios where the existing systems are outdated, the transformation capabilities are weak, and there are a large number of small and medium-sized entities. In this model, lightweight adapters are deployed at the boundary of the data holder to undertake five core functions: First, protocol adaptation, converting legacy databases, files, and various interfaces into standard API services; second, authorization verification, verifying the calling entity, purpose, scope, and validity period; third, format mapping, converting unstructured data into standardized data products; fourth, log reporting, synchronizing call summaries, abnormal information, and audit hashes; fifth, security proxy: achieving cross-domain secure transmission through keys, authentication, and whitelist. The core value of this model lies in significantly reducing the access threshold. There is no need to modify the business system; it can be implemented by simply configuring. Small and medium-sized institutions can be deployed in a lightweight manner, and

industry platforms can be used as data space connectors for practical applications.

4.4. Mode selection rules

The dual-mode API data market should not allow users to arbitrarily choose the mode, but should form rules based on data sensitivity levels, entity capabilities, business timeliness, and compliance requirements. The following selection logic can be established. As shown in **Table 1**.

Table 1. Mode selection rules

Dimension of judgment	Peer-to-peer model	Adapter pattern
Data sensitivity	High sensitivity, strict supervision, business secrets, financial risk data	Moderately low sensitivity, status-related, event-related, desensitized data or locally fieldable data
Original data control	The original data is always directly controlled by the holder.	The original data remains on the holding party's side, and the adapter is responsible for the conversion and proxying.
Accessing entity capability	Both parties possess strong API and security operation capabilities.	A large number of legacy systems, weak renovation capabilities, and a large number of small and medium-sized entities
Typical scenario	Financial credit granting, customs declaration verification, confidential government affairs supervision	Logistics status synchronization, warehouse document circulation, reuse of government affairs materials, and data integration of industrial equipment

4.5. Lightweight toolset and operational mechanism

Four lightweight tool components—access, authorization, archiving, and supervision—are deployed, supporting visual configuration and automated operations. At the same time, a standardized operation system is established to regulate data product listing, version management, frequency control, service quality evaluation, and dispute resolution mechanisms, clearly defining the platform's neutral operational positioning to prevent it from acting as both referee and player, thereby avoiding conflicts of rights and responsibilities.

5. Cross-domain application scenarios

5.1. Shipping trade: Document verification and industry-finance collaboration

As the core pilot scenario, the shipping trade domain achieves differentiated circulation through the dual-mode system: highly sensitive documents such as bills of lading, customs declarations, and financing credit approvals adopt the point-to-point mode to ensure the security of trade and financial data; standardized low-sensitivity data such as ship schedules, port operations, and logistics trajectories are uniformly accessed through adapters. Based on industry standards, trusted verification of electronic documents is achieved, connecting trade, logistics, and financial data chains, solving problems such as cumbersome manual verification, document forgery, and financing risk control difficulties, supporting the implementation of scenarios such as offshore trade, commodity trading, and warehouse receipt pledge financing.

5.2. Government services: Cross-departmental approval and benefit fulfillment

To address issues such as cross-departmental government data barriers, repeated document submission, and low approval efficiency, government data is standardized and cataloged. Sensitive data such as business operations and personal privacy adopt point-to-point authorized invocation, while government public information and status data are reused across departments through adapters. In the scenario of preferential

policy benefit fulfillment, the platform automatically matches enterprise qualification data through authorization verification, with full traceability and auditability, significantly improving approval and policy fulfillment efficiency while ensuring government data compliance.

5.3. Financial risk control: Minimal necessary data collaborative risk control

The financial industry strictly adheres to the data minimization principle. High-risk scenarios such as credit approval and anti-fraud adopt the point-to-point mode for precise invocation of core sensitive data; auxiliary data such as supply chain transactions and logistics fulfillment are accessed in batches through adapters. The platform defines data sources and usage boundaries through authorization archiving and log auditing, supporting financial institutions in compliant risk control, while also enabling result-based data interaction to reduce plaintext data exposure and mitigate data misuse risks.

5.4. Supply chain collaboration: Full-chain data linkage

Given the characteristics of fragmented data in the supply chain and strong demands for commercial privacy protection, an adapter pattern is adopted to uniformly adapt heterogeneous ERP, WMS, and TMS systems from multiple enterprises, achieving large-scale linkage of orders, logistics, and inventory status. Confidential data such as pricing, customer resources, and credit limits of core enterprises are managed through a point-to-point mode. This model does not require the reconfiguration of the business systems of each entity in the supply chain, enabling full-chain collaboration while balancing the efficiency of collaboration and commercial privacy security.

5.5. Industrial internet: Cross-domain industrial data collaboration

Industrial equipment data, process parameters, and quality inspection data are highly frequent, heterogeneous, and highly sensitive. They are connected to edge gateways, MES, and IoT platforms through adapter mode, and standardized event streams are aggregated. Sensitive data such as core processes, fault traceability, and high-end production parameters are directly connected in a point-to-point mode. It meets the interconnection requirements of industrial parks and equipment from different manufacturers, builds a lightweight industrial data space, and supports industrial digital collaboration.

6. Promotion path and evaluation indicators

6.1. Phased implementation path

The first stage is the pilot verification of scenarios, with a focus on implementing scenarios with clear boundaries and definite value, such as customs trade document verification, reuse of government materials, and supply chain pledge financing. The verification of dual-mode adaptation rules and auditing systems is also conducted. The second stage is the replication of industry standardization, where data product specifications, API interface standards, and authorization templates for various industries are consolidated to form reusable industry access solutions. The third stage is the expansion of the ecosystem across the entire domain, relying on government guidance, leading enterprises' traction, and service providers' empowerment to promote large-scale access by small and medium-sized entities, and to improve the ecosystem governance system.

6.2. Role division of entities

The government is responsible for rule-making, compliance supervision and public data authorization; industry leaders provide core data products and scenario guidance; financial institutions offer risk control and financing needs; third-party service providers provide technical support such as adapter deployment and operation maintenance; small and medium-sized enterprises participate in the ecosystem through lightweight access; platform operators remain neutral, dedicated to governance and auditing, to ensure the fairness and trustworthiness of the ecosystem.

6.3. Core evaluation indicators

A comprehensive evaluation system is constructed from four dimensions: Firstly, business efficiency indicators, which cover the access cycle, approval duration, and frequency of manual verification; Secondly, security compliance indicators, including the rate of unauthorized interception, the completeness rate of logs, and the rate of sensitive data transfer through the platform; Thirdly, ecosystem expansion indicators, covering the number of access entities, the scale of data products, and the access costs for small and medium-sized entities; Fourthly, value creation indicators, including the availability of financing, the efficiency of customs clearance and approval, and the data reuse rate, etc., to verify the implementation effectiveness of the model through quantitative dimensions.

7. Challenges and outlook

At present, there are still several challenges in the implementation of the dual-mode API data market: Firstly, the semantic and interface standards across industries have not been unified, which may lead to new issues of data fragmentation; Secondly, the adapter security certification system is missing, and the compliance and anti-tampering capabilities of its operation need to be standardized and controlled; Thirdly, the continuous zero-trust verification mechanism across network domains needs to be improved; static boundary protection cannot adapt to dynamic circulation scenarios; Fourthly, the rights and responsibilities of data intermediaries and the pricing incentive system have not been fully institutionalized; Fifthly, the mutual recognition of cross-border data circulation rules is difficult.

In the future, this model will be deeply integrated with privacy computing, verifiable credentials, AI risk control, and digital trade rules, continuously optimizing the minimum disclosure of data and the ability of trusted result interaction. In the long term, it will adhere to the core principle of “platform managing governance, data belonging to the entities, and adapters reducing barriers”, and build a replicable, auditable, and manageable cross-domain data circulation infrastructure.

8. Conclusion

In response to the industry pain point that it is difficult to balance security, efficiency, and compliance in cross-domain multi-scenario data circulation, this paper proposes a “point-to-point + adapter” dual-mode API data market application path. Through the core architecture of separating management flow from data flow, the point-to-point mode solves the problem of secure circulation of highly sensitive data, ensuring data sovereignty and compliance as a bottom line; the adapter mode breaks through the bottleneck of difficult system adaptation and high access costs for small and medium-sized entities, and realizes large-scale

ecological expansion. Its theoretical value lies in improving the scenario-based implementation mechanism of the data market and data space. Its practical value lies in providing a new type of data circulation solution without the risk of original data centralization for sectors such as shipping and trade, government affairs, finance, supply chain, and industrial internet, and can provide important support for the market-oriented allocation of data elements.

Funding

2024 National “Blockchain” Key Special Project “Key Technologies for Cross-departmental Data Access and Circulation Management in Shipping Trade” (Project No.: 2024YFB2705403)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] International Data Spaces Association, 2024, Dataspace Protocol 2024-1.
- [2] Castro Fernandez R, Subramaniam P, Franklin M, 2020, Data Market Platforms: Trading Data Assets to Solve Data Problems. arXiv:2002.01047.
- [3] Azcoitia S, Laoutaris N, 2022, A Survey of Data Marketplaces and Their Business Models. arXiv:2201.04561.
- [4] Zhang M, Beltran F, Liu J, 2023, A Survey of Data Pricing for Data Marketplaces. arXiv:2303.04810.
- [5] Gunturu N, 2022, Enterprise API Transformation: Driving Towards API Economy. *International Journal of Computer Trends and Technology*, 70(6): 44–50.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on a Trustworthy Collaborative Computing Mechanism for Cross-Domain Sensitive Data Based on Secure Multi-party Computation

Hao Yang, Mutong Chen, Xiong Ni

Shanghai Municipal Big Data Center, Shanghai 200072, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: In the digital economy era, cross-domain data circulation has shifted from traditional data exchange to value-based collaborative computing. Cross-institutional data in scenarios such as shipping trade, financial risk control, government supervision, and supply chain collaboration possess characteristics of high sensitivity, strong ownership, and high compliance constraints. Direct sharing of raw data can easily lead to issues such as leakage of commercial secrets, exposure of customer resources, violation of cross-border data regulations, and ambiguity in liability determination. Secure Multi-Party Computation (MPC), with its characteristics of “data not leaving the domain, value can be calculated, and results can be trusted”, has become the core privacy-enhancing technology for resolving the core privacy challenges in cross-domain sensitive data collaboration. This paper takes secret sharing, obfuscation circuits, and inadvertent transmission as the core protocols, integrates blockchain, DID, and smart contract technologies, builds a trusted collaborative computing mechanism for sensitive data across various domains, and constructs a standardized task lifecycle and on-chain-off-chain collaborative governance system. By verifying the effectiveness of the mechanism in four typical industry scenarios, it systematically analyzes the core bottlenecks in the implementation of the MPC project, such as performance, standardization, malicious protection, and result inference, and proposes template compilation, verifiable auditing, and privacy governance optimization solutions, providing a reference for the construction of a trusted collaborative infrastructure for privacy data across various domains.

Keywords: Secure multi-party computation; Cross-domain data circulation; Sensitive data; Secret sharing; Garbled circuits; Blockchain; DID; Trusted collaborative computing

Online publication: Sept 3, 2026

1. Introduction

The core contradiction of cross-domain data collaboration lies in the conflict between data value extraction and the protection of original data privacy. Entities in sectors such as shipping, finance, government affairs,

and supply chain have independent ownership rights and prominent data barriers. Various business data are highly sensitive and subject to strict compliance constraints. Traditional blockchain and DID technologies can only achieve process evidence and identity authorization, but cannot solve the privacy protection problem of joint computing of multiple private data. Therefore, this paper takes MPC as the core cryptographic computing layer and builds a cross-domain trusted collaborative computing mechanism. Under the premise of retaining the original data locally, it realizes multi-party joint verification, risk scoring, statistical analysis, and rule matching, effectively balancing the demand for data value release and privacy security control^[1].

2. Related work

2.1. MPC basic protocols

The core goal of MPC is to enable multiple mutually untrusting parties to collaboratively complete a pre-specified function computation, with each participant only obtaining the agreed-upon result and unable to steal or infer others' private data. Classical theories form a complete technical foundation: Shamir's threshold secret sharing enables secure data splitting and threshold reconstruction, ensuring static privacy security; Yao's garbled circuits enable Boolean privacy computing without data sharing; GMW and BGW protocols extend secure computation to general multi-party scenarios; Oblivious Transfer (OT) provides underlying support for privacy-preserving matching and circuit computation; Beaver triple pre-processing significantly reduces online computational overhead; and maliciously secure protocols like SPDZ, relying on offline pre-processing and message authentication mechanisms, can resist input forgery, result tampering, and node collusion attacks, making them suitable for high-security-level cross-domain sensitive data trusted collaboration scenarios.

2.2. MPC engineering systems and hybrid protocol frameworks

Currently, MPC (Multi-Party Computation) has moved from theoretical demonstration to practical implementation, forming a mature framework for executing mixed protocols. Mainstream open-source systems such as ABY, MP-SPDZ, and ABY2.0 integrate multiple cryptographic protocols, supporting dynamic mixed arrangements of arithmetic sharing, Boolean sharing, and Yao circuits. They can be seamlessly adapted to complex business functions such as numerical statistics, size comparison, conditional branching, rule matching, and set retrieval. The systems generally come equipped with optimization capabilities such as offline preprocessing, GPU parallel acceleration, communication compression, and dynamic bandwidth adaptation. They can flexibly switch between semi-trustworthy and malicious security protection mechanisms according to the security level of the business. Compared with a single protocol architecture, a hybrid protocol can automatically select the most suitable protocol based on the characteristics of sub-tasks, reducing the cost of domain transitions and significantly reducing the computational power consumption and communication delay of complex cross-domain tasks. This engineering system provides a reusable and iterative technical foundation for highly complex cross-domain scenarios, effectively promoting MPC from a laboratory prototype to industry-scale commercial application^[2].

2.3. MPC data analysis and standardized governance

Since 2020, MPC has been deeply integrated with big data and artificial intelligence, giving rise to the privacy machine learning system. Systems such as CrypTFlow, Delphi, and CrypTen can perform model

privacy inference and lightweight training, and are suitable for intelligent risk control, anomaly identification, and other business applications. Compared with federated learning, which relies on model parameter aggregation and has the risk of gradient leakage, MPC can achieve privacy-preserving collaborative computing at the original feature level, with finer data protection granularity and stronger security. At the standardization level, NIST has incorporated MPC into the privacy-enhanced cryptographic tool system. The DID/VC standard of W3C provides regulatory support for cross-domain identity authorization. Combined with the ISO security and privacy system, it forms a dual guarantee of technology and system. Compared with homomorphic encryption, TEE, and zero-knowledge proof, MPC is more suitable for multi-party peer-to-peer joint computing. It can deeply collaborate with blockchain and smart contracts to build a complete cross-domain sensitive data trusted collaboration system, effectively solving the cross-domain trusted collaboration pain points of missing mutual trust among multiple parties, complex data ownership, and strict compliance requirements.

3. Problem analysis of trusted collaborative computing for cross-domain sensitive data

3.1. Participants and data characteristics

Participants in cross-domain sensitive data trusted collaboration can be categorized into five types: data holders, data users, data authorizers, regulatory auditors, and technical service providers. Each entity has independent rights and responsibilities, providing mutual checks and balances. Collaborative data—such as shipping trade documents, financial transactions, government regulatory data, and supply chain logistics—possesses characteristics of low individual-source value but high multi-source fusion value, strict ownership, and strong confidentiality requirements. Traditional centralized data sharing models easily lead to privacy breaches and compliance risks, failing to meet the business and security needs of cross-domain multi-party collaboration.

3.2. Threat model

This paper constructs a dual-layer threat model combining algorithms and governance, covering core risks in cross-domain trusted collaboration. The algorithmic aspect includes two types of attacks: semi-honest and malicious. The semi-honest model follows the protocol but attempts to steal logs and infer privacy; the malicious model can forge inputs, tamper with results, terminate the protocol, and collude for espionage, and is the core protection target for sensitive business. The governance and operation maintenance aspect also has derivative risks, including leakage of task metadata revealing business relationships, false inputs generating invalid results, fine-grained queries triggering differential inference, inconsistent log data across private networks, and ambiguity in multi-party rights and responsibilities making accountability difficult ^[3].

3.3. Core requirements and constraints

By integrating cross-domain business and security standards, this article has distilled six core requirements: First, local input confidentiality - the original data does not leave the local domain, and cross-domain data transmission only involves protocol protection; second, trustworthy calculation results - accurately resist malicious tampering and false input, ensuring calculation consistency; third, verifiable and controllable authorization - relying on DID/VC to clearly define the scope, scenarios and validity period of data usage;

fourth, minimum result disclosure - only output the coarse-grained results necessary for the business, preventing detailed leakage and reverse inference; fifth, full traceability - store all process information on the chain for evidence, enabling traceability of anomalies and determination of responsibility; sixth, deployment compatibility and adaptation - can be adapted to multiple types of private network environments, supporting stable access of multiple formats of data.

4. Design of MPC-based trusted collaborative computing mechanism

4.1. Overall architecture

This paper constructs a four-layer closed-loop collaborative architecture to achieve deep coupling of privacy computing and governance processes, as well as hierarchical division of responsibilities and full-process controllability. The governance management layer serves as the top-level constraint system. It relies on DID/VC to realize trusted identity identification and refined permission definition across institutions and systems. Through the unified registration of various sensitive data in the global data directory, the semantics of fields, confidentiality level, applicable scenarios, and authorization constraints are defined. The task access rules, frequency limits, data scope, and compliance boundaries are solidified by smart contracts. The MPC task layer, as the core computing core, is responsible for abstracting fragmented business requirements into standardized and compliant task templates, automatically completing protocol intelligent selection, offline cryptographic material preprocessing, online privacy computing, and output granularity control. The cross-domain node layer enables each entity to complete data cleaning, compliance verification, field de-sensitization, and share encoding within the local security domain. It achieves cross-domain trusted communication through encrypted channels and ensures the retention of original data at the physical level. The result auditing layer realizes minimized disclosure of calculation results, hash chain-based evidence storage, and complete archiving of operation logs, providing fallback support for post-event compliance audits, business dispute resolution, and precise determination of responsibilities ^[4].

4.2. Full task lifecycle

This paper designs an eight-stage standardized task lifecycle to achieve controllable and auditable privacy computing throughout the process. The process is as follows: The data holder completes the registration of data capabilities; The user initiates the task application and completes the authorization of ownership; The smart contract verifies compliance and triggers the task; Each node locally completes data verification and share processing; The system matches the optimal protocol and completes offline preprocessing; Nodes exchange protocol data and conduct online secure computing; The authorization entity obtains the minimized disclosure result; Finally, the task data and logs are archived on the chain, handling abnormal risks, and throughout the process, no original data is transferred across domains.

4.3. Core protocol adaptation paths

For differentiated business scenarios, this paper designs four types of dedicated protocol paths to balance security and performance. The secret sharing path is suitable for arithmetic tasks such as multi-party statistics and risk scoring, and reduces computational costs by relying on threshold sharing and Beaver triplets; the obfuscation circuit path is suitable for Boolean tasks such as document verification, numerical comparison, and rule judgment, ensuring the privacy security of logical computing; the oblivious transmission path

realizes private set intersection based on OT extension, suitable for scenarios such as blacklist matching and duplicate financing identification; the hybrid protocol path uses the ABY architecture to achieve dynamic conversion of multiple protocols, suitable for complex risk control and supervision tasks that involve statistics, judgment, and matching ^[5].

4.4. On-chain/off-chain synergy mechanism

This paper proposes a collaborative model of “chain-based management of processes and offline calculation of privacy”, which balances the requirements of privacy protection and compliance auditing. The blockchain and smart contracts only store management data such as task numbers, DID identities, authorization summaries, result hashes, and audit logs, without retaining original data or secret shares. This avoids the risk of transparent leakage on the chain. The DID/VC system realizes multi-level identity control and traceable authorization, and supports permission revocation. The smart contracts automatically complete task access, process arrangement, and abnormal monitoring, replacing manual review and achieving standardized and auditable operation of the privacy computing process.

4.5. Cross-network deployment and evaluation system

This article adopts a localized node deployment model. Each participant deploys MPC nodes in their own secure domain to achieve local data control and completely avoid the risks of aggregation leakage and over-range usage brought by traditional centralized data platforms. The nodes complete cross-domain trusted communication through security gateways and encrypted dedicated lines. They adopt two-way identity authentication, traffic auditing, and abnormal breakage mechanisms, transmitting only protocol messages and audit summaries, thus eliminating the risk of data leakage in centralized storage. At the same time, a multi-dimensional evaluation system is established. From five dimensions, including privacy security, computational correctness, running performance, business adaptability, and compliance governance, the security, stability, and practicality of the mechanism are quantitatively evaluated. It is adapted to heterogeneous cross-domain deployment scenarios such as government dedicated networks, financial intranets, and enterprise private clouds, effectively solving the data collaboration problems in multi-network isolation environments.

5. Cross-domain application scenarios

5.1. Joint verification of shipping trade document authenticity

Shipping trade document verification is a critical requirement for trade financing and warehouse receipt pledging, where traditional sharing models carry significant forgery and leakage risks. Based on the mechanism proposed in this paper, entities such as shipping companies, ports, customs authorities, and warehousing institutions locally submit authorized fields. Through OT-based set matching, garbled circuits, and secret sharing, they perform document consistency comparisons, interval checks, and pledge risk scoring. The system outputs only the verification conclusion and risk level to banks, without disclosing document details or commercial data, effectively preventing fraudulent trade and duplicate pledging risks while balancing financial security, data privacy, and government compliance requirements.

5.2. Financial joint risk control and credit assessment

Joint financial risk control has long faced the challenges of cross-institutional data silos and compliance constraints. The mechanism proposed in this paper enables privacy-preserving collaborative computing of data from banks, logistics, taxation, and insurance. Each institution retains its data locally throughout the process. Through a hybrid protocol, it achieves multi-dimensional feature fusion, enabling risk control functions such as enterprise credit scoring, duplicate financing screening, and false transaction identification. The system only outputs score ranges, risk levels, and admission conclusions, effectively avoiding the risk of data leakage of customers and transactions, breaking through the limitations of a single data source, and significantly improving the accuracy and compliance of cross-domain risk control.

5.3. Government joint supervision and risk identification

There are problems such as data fragmentation among multiple departments and insufficient collaborative investigation in government supervision. The data of different departments cannot be compliantly shared, resulting in missed detection of abnormal risks for enterprises. Relying on the MPC mechanism, departments such as customs, taxation, market supervision, and maritime affairs do not need to share detailed data. They can jointly identify high-risk behaviors such as abnormal customs declarations, invoice violations, abnormal trajectories, and cumulative penalties across departments, and only output risk clues and regulatory marks. By integrating smart contracts to solidify regulatory authorization and processes, we can achieve full traceability and clear rights and responsibilities, addressing the pain points of “lack of willingness to share and difficulty in collaboration” in government data, and supporting precise and compliant cross-departmental joint supervision.

5.4. Supply chain collaborative analysis

In the financial scenarios of supply chain and industrial chain, enterprise operation data is highly sensitive, and there is a high risk of data sharing. Data barriers restrict the implementation of risk sharing and inclusive finance. Traditional supply chain collaboration mainly relies on the centralized collection of data by core enterprises, which easily leads to the monopoly and abuse of business information by upstream and downstream small and medium-sized enterprises, and poses significant compliance and privacy risks. This mechanism enables collaborative analysis of supply chain privacy. Each participant locally uploads the authorization fields, and through MPC, it can complete order verification, inventory assessment, performance scoring, and accounts receivable verification. Only aggregated indicators, warning levels, and financing suggestions are output, without disclosing the detailed business data of the enterprises. This model can effectively break down the data barriers in the supply chain, balance the risk control requirements of the core enterprises with the privacy rights of small and medium-sized enterprises, and assist in the risk management of the industrial chain. It can be widely applied to multiple industries such as manufacturing, cross-border e-commerce, and energy.

6. Challenges and outlook

At present, there are still significant engineering and governance bottlenecks in the cross-domain large-scale application of MPC, which restrict the large-scale implementation of the technology. Firstly, the heterogeneous network environment across different specialized networks is complex. The isolation degrees

of government affairs, finance, and enterprise networks are high, and the communication delay and jitter are large. The optimal performance in the laboratory cannot be replicated in real cross-domain scenarios, and the overall computing efficiency is limited. Secondly, the industry data standardization system is lacking. The field definitions and risk statistics standards of each institution are not unified, resulting in poor generalizability of MPC task templates, high customization costs, and difficulty in batch reuse. Thirdly, the data authenticity governance is weak. MPC can only guarantee the privacy and security of the computing process, but it cannot actively identify deliberate falsification by participants, field mismatches, or selective submission of data, leading to potentially unsafe but ineffective computing results. Fourthly, fine-grained and frequent repeated queries can be achieved through differential iteration for reverse data inference, and there is a risk of implicit privacy leakage. Fifthly, the balance between chain-level transparent evidence storage and privacy protection is difficult. Blockchain is suitable for evidence storage and auditing, but the chain-level information is permanently visible, and the task frequency, participant relationships, and result summaries may also constitute sensitive metadata, posing a leakage risk. Sixthly, the industry standard evaluation benchmarks, performance specifications, and regulatory explainability systems are not yet complete, which is not conducive to compliance implementation and industry promotion.

To address these bottlenecks, future optimization and iteration can be carried out in four aspects: Firstly, by combining GPU acceleration and batch processing optimization to reduce computing power and communication costs, and improving deployment stability; Secondly, unifying industry data semantics and computing standards, and building a reusable privacy computing function library that can be compiled and reused; Thirdly, constructing a dual mechanism of protocol protection and governance constraints, through input commitment, on-chain verification, DID responsibility binding, anomaly detection, post-event inspection, and differential privacy to prevent data fraud and privacy inference risks; Fourthly, improving the industry standard evaluation and auditing system, and building a traceable and supervisable cross-domain trusted collaborative infrastructure.

7. Conclusion

This paper addresses the challenges of cross-domain sensitive data collaboration, confidentiality, and auditability by constructing an MPC-based trusted collaborative computing mechanism. The mechanism integrates blockchain, DID, and smart contract technologies to achieve notarization, auditing, process control, and identity authorization, forming a closed-loop model of “on-chain governance, off-chain privacy computation”. It achieves the goals of data not leaving its domain, value being computable, processes being auditable, and accountability being traceable. This alleviates the contradiction of cross-domain data that “cannot be shared but needs collaboration”, and is particularly suitable for scenarios such as joint verification of shipping trade document authenticity, financial joint risk control and credit assessment, government joint supervision and risk identification, and supply chain collaborative analysis. The paper identifies core issues in MPC deployment, including performance, standardization, and data quality, and clarifies future optimization directions, providing both theoretical and practical support for compliant cross-domain privacy data collaboration and large-scale privacy-preserving computation deployment.

Funding

2024 National “Blockchain” Key Special Project “Key Technology Research on Cross-Sector Data Access and Circulation Management for Shipping Trade” (Project No.: 2024YFB2705403)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Shamir A, 1979, How to Share a Secret. *Communications of the ACM*, 22(11): 612–613.
- [2] Yao A, 1982, Protocols for Secure Computations. *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science*, 160–164.
- [3] Yao A, 1986, How to Generate and Exchange Secrets. *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, 162–167.
- [4] Goldreich O, Micali S, Wigderson A, 1987, How to Play Any Mental Game. *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, 218–229.
- [5] Ben-Or M, Goldwasser S, Wigderson A, 1988, Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation. *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, 1–10.

Publisher’s note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Current Status and Development Trends of Biomedical Big Data

Juan Zhang

North Health Medical Big Data Technology Co., Ltd., Jinan 250000, Shandong, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: The deep integration of digital technologies has ushered biomedical science into a new era driven by data, making biomedical big data a cornerstone for precision medicine, medical research innovation, and public health prevention and control. Compared to traditional medical data, modern biomedical big data is characterized by multimodality, high dimensionality, and strong interconnectivity, fundamentally transforming conventional empirical and fragmented approaches to research and clinical practice. This paper systematically reviews the current applications of biomedical big data from key medical use cases, analyzes existing challenges in data governance, technological integration, and ethical compliance, moves beyond superficial generalizations, and identifies the main development directions of the field across three dimensions: technological evolution, clinical translation, and system building. It provides theoretical support for the standardized application, industrial implementation, and advancement of biomedical big data in the medical domain.

Keywords: Biomedical big data; Medical applications; Data governance; Precision medicine; Development trends

Online publication: Sept 8, 2026

1. Introduction

The deep integration of life sciences and information technology has liberated biomedical research from the constraints of traditional experimental observations and clinical experience, ushering it into a new paradigm of data-intensive science. Clinical records, omics data, medical imaging, and health monitoring information are continuously accumulating, forming a comprehensive biomedical big data system that spans the entire human lifespan and covers all stages of disease. These data are no longer mere byproducts of diagnosis and treatment but have become key production factors for decoding biological mechanisms, conquering complex diseases, and optimizing healthcare resource allocation. Biomedical big data is increasingly penetrating core areas such as basic medical research, precision clinical care, and public health management. However, significant challenges remain in data integration, technological adaptation, and secure governance. By systematically reviewing the current state of industry development through the lens of medical application needs, numerous existing issues can be identified. Addressing these challenges with foresight on future trends

will facilitate the transformation of data value into tangible clinical outcomes, driving the modern medical system toward greater intelligence and precision.

2. Current development status of the biomedical big data industry

2.1. Diversified data systems covering all medical application scenarios

Currently, biomedical big data has established a multidimensional and comprehensive data architecture, breaking through the limitations of traditional medical data that were previously fragmented and limited in scope. It now covers nearly all medical scenarios, including basic research, clinical diagnosis and treatment, public health, and health management ^[1]. In terms of data types, these mainly include genomics data, clinical care data, medical imaging data, and dynamic health monitoring data. Genomics data represent the microscopic mechanisms of life, encompassing biological information at various levels such as genes, proteins, and metabolites, serving as fundamental data for elucidating disease mechanisms and enabling personalized medicine. Clinical care data consist of patient records, treatment plans, medication logs, follow-up information, and other materials generated from daily operations across healthcare institutions, forming the core resources for evidence-based clinical research. Medical imaging data include ultrasound, MRI, CT, and other imaging modalities, providing direct support for non-invasive disease diagnosis and disease progression assessment. Real-time health status data collected via intelligent medical devices also play a crucial role in daily health management and chronic disease prevention and control.

2.2. Deep implementation in application scenarios, empowering innovation in medical models

The application of biomedical big data has evolved from simple data storage and statistics to deeply empowering the entire medical chain, driving traditional empirical medicine toward precision medicine, evidence-based medicine, and preventive medicine. In clinical practice, integrating and analyzing multimodal data can help healthcare professionals uncover hidden disease characteristics, identify disease progression patterns, improve diagnostic and therapeutic strategies, and reduce misdiagnosis and missed diagnoses, making it particularly suitable for managing complex cases such as difficult-to-diagnose conditions and malignant tumors. In basic medical research, a data-driven approach is replacing the previous reliance on single experiments for validation, enabling researchers to discover connections within vast biological datasets, rapidly identify disease targets and mechanisms, and significantly shorten the time required for medical scientific breakthroughs. In public health, comprehensive integration and analysis of healthcare data allow for dynamic assessment of disease trends, supporting rapid response and precise prevention and control during public health emergencies, and optimizing the allocation of public health resources ^[2].

2.3. The technical system has gradually become more robust, enabling efficient data processing

With the continuous advancement of information technology, a processing framework tailored to the characteristics of biomedical data has gradually emerged, overcoming the limitations of traditional medical data processing, such as low efficiency and limited analytical dimensions. In terms of data collection and storage, high-throughput acquisition technologies and distributed storage systems have been widely adopted, effectively addressing the massive volume, multi-source nature, and heterogeneity of biomedical data,

enabling unified aggregation and long-term preservation of diverse fragmented data ^[3]. In data processing and analysis, intelligent algorithms and machine learning techniques have been integrated into biomedical applications, facilitating multimodal data cleaning, matching, fusion, and in-depth mining to accurately identify underlying life cycle patterns or disease manifestations. On the application side, innovations such as medical knowledge graphs and intelligent diagnostic support systems have been developed, significantly improving the efficiency of translating big data analytics into clinical practice and scientific research. Through synergistic integration of these technologies, biomedical big data is being transformed into a valuable and practically usable medical resource.

3. Core challenges in the development of biomedical big data

3.1. Prominent data silo problem, making integration and sharing difficult

Data fragmentation and isolation represent the most critical challenges in today's biomedical big data field. Data collection standards, storage formats, and statistical methodologies vary significantly across medical institutions, research organizations, and health management agencies at different levels, preventing interoperability among healthcare data from different regions and tiers. Moreover, some institutions maintain a closed mindset regarding data, lacking proactive willingness to share information, which results in high-quality medical data being scattered and underutilized, failing to form large-scale, systematic data resources. Due to the heterogeneity of multi-source data, issues such as matching discrepancies and information redundancy arise during data integration, greatly reducing data usability. The existence of data silos not only limits the overall value of big data but also prevents medical research and clinical practice from conducting comprehensive analyses based on holistic datasets, severely hindering the large-scale implementation of precision medicine ^[4].

3.2. Lack of data governance framework, resulting in inconsistent data quality

Biomedical data is directly linked to life and health, and the authenticity, completeness, and standardization of data are crucial for ensuring effective applications. However, the industry has yet to establish a unified and comprehensive data governance system. During data collection, issues such as manual entry errors, equipment measurement biases, and inconsistent clinical documentation often lead to problems in raw data, including missing values, inaccuracies, and redundancy. In the data processing stage, standardized procedures for cleaning, filtering, and calibration are lacking, and different entities apply varying data processing standards, resulting in biased analytical outcomes that cannot be reliably applied across different scenarios. Due to the absence of routine quality control mechanisms and an inadequate system for data updating, maintenance, and traceability, some older data no longer meet current medical development needs, thereby diminishing their application value. The prevalence of low-quality data negatively impacts clinical decision support and scientific research, and may even lead to errors in medical decision-making.

3.3. Poor technical adaptability and low clinical translation efficiency

Currently, most big data technologies rely on general information technology frameworks, resulting in delayed development of specialized tools tailored to biomedical applications and insufficient compatibility between technology and clinical settings. General-purpose intelligent algorithms have limited capacity to analyze the complexity of biomedical data, failing to accurately capture nonlinear relationships within

biological systems or uncover hidden patterns underlying complex diseases. As a result, many analytical outcomes suffer from poor interpretability and low relevance to clinical practice. Most big data research achievements remain at the experimental stage, lacking refinement and optimization for real-world clinical use, creating a clear gap between scientific discoveries and practical application, hindering the development of implementable diagnostic methods and interventions. Moreover, primary-level medical institutions often lack technical expertise and sufficient personnel, making it difficult for them to adopt intelligent big data solutions. Consequently, high-quality data and technological resources fail to achieve comprehensive coverage, leading to significant regional and hierarchical imbalances in industry development.

3.4. Inadequate safety ethics system increases compliance risks

Biomedical big data contains a significant amount of personal privacy information and core genetic data, making it highly sensitive. However, current industry security protection and ethical oversight systems still have notable shortcomings. In terms of data security, some institutions lack robust safeguards in data storage, transmission, and application, with insufficient encryption, data anonymization, and access control mechanisms, leading to risks such as data breaches, misuse, and tampering. Regarding ethical regulation, guidelines for data collection, sharing, and usage remain unclear, informed consent mechanisms for patients are inadequate, and issues such as excessive data collection and unauthorized use of private information persist in both research and commercial applications. Additionally, mechanisms for data traceability and accountability are underdeveloped, making it difficult to accurately assign responsibility when data security incidents occur. The absence of a comprehensive security and ethics framework not only violates individuals' rights to privacy but also hinders the development of the biomedical big data industry.

4. Future trends in biomedical big data

4.1. Deepening of data standardization and governance, with comprehensive data integration and sharing becoming the norm

The future of the biomedical big data industry will focus on standardization, fully addressing the issue of data silos and achieving seamless, cross-domain data interconnectivity. The industry will gradually establish unified standards for data collection, storage, classification, and transmission, standardizing data production processes across various medical and research institutions. It will also unify multimodal data formats and statistical methodologies to eliminate data heterogeneity at the source. Leveraging emerging technologies such as privacy-preserving computing and blockchain, secure and controllable data-sharing mechanisms will be built, breaking down data barriers between institutions and regions while safeguarding data privacy. This will enable collaborative data access across different entities and domains. A routine data quality governance framework will emerge, continuously improving data quality through intelligent validation, dynamic updates, and end-to-end traceability systems. Ultimately, this will create a high-quality, trustworthy biomedical big data resource repository, laying a solid foundation for applications across diverse industries.

4.2. Deep integration with technical scenarios, continuously enhancing clinical translation efficiency

The integration of information technology and biomedical sciences will deepen, leading to the gradual establishment of specialized, context-specific data technology systems. In the future, efforts will focus

on addressing key challenges in medical settings by developing dedicated intelligent algorithms tailored for multi-omics biology, clinical diagnosis and treatment, and public health prevention and control. These algorithms will enhance the ability to analyze complex medical data and improve the interpretability of results, precisely meeting the needs of clinical research and patient care. A major priority will be breaking down barriers between research and clinical practice, transforming big data analysis outcomes into practical tools such as intelligent diagnostic support systems, personalized treatment plans, and disease risk prediction models, thereby closing the loop from data mining to real-world clinical application. Lightweight and user-friendly big data technologies for primary healthcare will continue to evolve and become widely adopted, helping to reduce disparities in healthcare across regions and levels, promoting nationwide sharing of health information, and making medical services increasingly intelligent.

4.3. Application boundaries continue to expand, forming a full-cycle health service system

The application of biomedical big data will extend from traditional disease diagnosis and treatment, as well as scientific research, to comprehensive health management throughout the entire life cycle, establishing an integrated intelligent healthcare system that covers prevention, diagnosis, rehabilitation, and care. In prevention, long-term collection and analysis of diverse health information enable prediction of disease susceptibility risks, allowing medical interventions to be implemented proactively before illness develops, shifting healthcare focus from disease treatment to health promotion. In diagnosis and treatment, multimodal data integration facilitates fully personalized precision medicine, achieving refined medical models such as “same disease, different treatments” and “different diseases, same treatment.” For rehabilitation and nursing, dynamic health monitoring data are used to develop individualized rehabilitation assessment and intervention systems, making chronic disease management and post-surgical recovery more scientific. Big data will also be deeply integrated into public health emergency response systems, improving rapid reaction and efficient handling of sudden outbreaks or epidemic events.

4.4. Standardization of the safety ethics system leads to orderly and controllable industry development

As the industry develops in the future, a compliance framework integrating technical safeguards, institutional constraints, and ethical oversight will emerge, achieving a synergistic effect between data value realization and security and privacy protection. Technologically, security measures such as data anonymization, encrypted transmission, access control, and privacy-preserving computing will be widely adopted, enabling data to be usable yet invisible, thereby mitigating risks of data breaches and misuse. Institutionally, more detailed regulations will be established for every stage of biomedical data collection, sharing, application, and trading, clearly defining the rights and responsibilities of all stakeholders and establishing mechanisms for accountability in data security. The ethical review system for medical data applications will be strengthened, with improved informed consent processes and privacy protection mechanisms to prevent data abuse and excessive data collection. Comprehensive standardization efforts will be implemented, helping the biomedical big data industry move away from unstructured growth and transition toward high-quality, sustainable development.

5. Peroration

Biomedical big data is the fundamental driving force behind the transformation and development of modern medicine. It has reshaped the approaches to medical research, clinical practice, and public health services, offering new solutions for diagnosing complex diseases, optimizing healthcare resources, and ensuring population-wide health. Although the industry has achieved initial scale, numerous practical challenges remain, including incomplete data integration, inconsistent data quality, weak technological translation, and inadequate attention to security and ethics. In the future, with strengthened standardized data governance, improved adaptability of technologies to specific applications, more comprehensive application systems, and increasingly robust compliance frameworks, biomedical big data will fully realize its core value. It will support the advancement of precision medicine, smart healthcare, and public health initiatives, while providing a strong underpinning for innovation in life sciences and safeguarding public health.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Li J, Yang N, 2025, Transformation of Biomedical Research Models Driven by Big Data. *Medicine and Philosophy*, 46(8): 1–6.
- [2] Ke Q, Lu M, Chen F, et al., 2023, Current Status and Reflections on Curriculum Design for Biomedical Big Data Majors. *Chinese Journal of Health Statistics*, 40(6): 946–949.
- [3] Zhang G, Zhao G, Li Y, 2023, Realizing the Value of Production Factors in Biomedical Big Data: Starting from Data Elements. *Life Sciences*, 35(12): 1545–1552.
- [4] Yang S, You D, Shao F, 2022, Exploring the Cultivation of Biomedical Big Data Analysis Capabilities in Biostatistics Majors. *Education Teaching Forum*, 2022(41): 41–44.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research on Plateau Cattle and Sheep Detection Based on QHCSM-GAN Image Completion and Improved YOLO-OC Network

Xin Zhang

Qinghai Minzu University, Xining, Qinghai, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Highland animal husbandry is a core pillar of Qinghai Province's agricultural and pastoral economy, and computer vision-based intelligent monitoring is a key approach to achieving refined and digital management of highland pastures. However, the complex environment at high altitudes presents challenges such as variable lighting, complex terrain, frequent target occlusion, and blurred imaging, leading to severe degradation in the quality of cattle and sheep images collected in the field and the loss of effective features. Furthermore, existing publicly available livestock datasets and general detection models have poor adaptability to open grazing scenarios on high plateaus, exhibiting shortcomings such as low detection accuracy and weak generalization ability. To address these issues, this paper proposes an integrated intelligent detection framework that combines image completion and target detection optimization. First, a dedicated Qinghai cattle and sheep dataset (QH-Livestock) covering various complex plateau scenes is constructed. Second, a generative adversarial network (QHCSM-GAN) for missing images of plateau cattle and sheep is designed, relying on a multi-path feature fusion encoder and a theoretically reliable discriminator to achieve degraded image restoration and missing feature completion. Finally, using YOLOv8n as the baseline model, a lightweight and high-precision YOLO-OC detection model is constructed by introducing full-dimensional dynamic convolution (ODConv) and convolutional block attention module (CBAM), enhancing multi-scale feature extraction capabilities and anti-interference capabilities against complex backgrounds. Experimental results show that QHCSM-GAN outperforms mainstream image completion algorithms in terms of structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and learned perceptual patch similarity (LPIPS). The proposed YOLO-OC model achieves an accuracy of 84.6%, a recall of 81.2%, and an mAP50 of 82.5% on the QH-Livestock test set, with an inference speed of 109 FPS. This method effectively solves the core problems of poor image quality and insufficient detection accuracy of cattle and sheep targets in complex high-altitude scenes, and can provide reliable technical support for intelligent monitoring of high-altitude pastures and digital transformation of animal husbandry.

Keywords: Plateau animal husbandry; Image completion; Deep learning; Object detection; YOLO-OC; Intelligent monitoring

Online publication: Sept 4, 2026

1.Introduction

Traditional manual inspection methods suffer from drawbacks such as low efficiency, high labor costs, and the inability to conduct large-scale, all-weather monitoring, making them unsuitable for the needs of modern ranch management. Intelligent sensing technologies based on deep learning and computer vision enable non-contact, real-time detection and status monitoring of cattle and sheep, and have become a core technological direction for the intelligent upgrading of plateau animal husbandry.

Current mainstream object detection algorithms have achieved good results in standardized, enclosed livestock farms with simple scenes and controllable lighting. However, their applicability is severely limited in open grazing scenarios on plateaus. First, the extreme environment of high altitude, low temperature, frequent sandstorms, and strong sunlight on plateaus leads to instability in image acquisition equipment, resulting in widespread degradation issues such as overexposure, underexposure, blurring, and partial occlusion in acquired images, causing a significant loss of key features of cattle and sheep. Second, the background of plateau grasslands, mountains, and shadows blends closely with the color of cattle and sheep, and the large scale, dispersed distribution, and varied postures of grazing cattle and sheep significantly increase the difficulty of accurate detection. Finally, existing publicly available livestock datasets are mostly built based on standardized inland farming scenarios, lacking labeled samples from extreme plateau environments. This makes general detection models prone to false positives and false negatives in plateau pasture scenarios, resulting in extremely poor generalization performance.

To address the aforementioned industry pain points, existing research primarily focuses on improvements in two dimensions: data augmentation and network structure optimization. At the data optimization level, traditional image inpainting methods can only handle minor image imperfections and cannot repair feature loss caused by large-area occlusion or strong light degradation. While traditional Generative Adversarial Networks (GANs) possess image generation capabilities, they suffer from problems such as pattern collapse, gradient instability, and unquantifiable uncertainty in generated results, easily leading to texture distortion and false targets ^[1]. At the detection model level, mainstream YOLO series algorithms, with their fixed convolutional kernels and single feature enhancement mechanisms, cannot dynamically adapt to multi-scale target changes in high-altitude environments, struggle to effectively suppress complex background interference, and their detection accuracy falls short of actual monitoring needs ^[2]. Therefore, there is an urgent need to construct an integrated solution for high-quality data construction, adaptive image inpainting, and scene-based detection optimization adapted to high-altitude scenarios.

The main research contributions of this paper are as follows: (1) A high-precision labeled QH-Livestock cattle and sheep dataset containing 77,506 images was constructed, covering typical difficult scenarios such as strong light, weak light, occlusion, and distant small targets, filling the gap in the high-altitude livestock-specific dataset; (2) The QHCSM-GAN image completion algorithm was proposed, which quantifies the uncertainty of image generation through a multi-path feature fusion coding structure and a reliable discriminator based on evidence theory, and realizes high-fidelity restoration of high-altitude degraded cattle and sheep images; (3) Based on the YOLOv8n baseline model, a lightweight YOLO-OC detection model was designed by integrating ODConv full-dimensional dynamic convolution and CBAM attention mechanism, which takes into account multi-scale feature extraction capability, background anti-interference capability and real-time inference performance ^[3,4]; (4) Through multiple sets of comparative experiments, ablation experiments and cross-domain generalization experiments, the effectiveness, robustness and universality of

the algorithm in this paper were fully verified.

The structure of this paper is as follows: Chapter 2 reviews the current research status in the fields of image completion and livestock target detection, and analyzes the shortcomings and research gaps of existing technologies; Chapter 3 details the construction process of the QH-Livestock dataset and the principle of the QHCSM-GAN image completion algorithm; Chapter 4 elaborates on the optimization strategy of the YOLO-OC model structure, experimental settings and result analysis; Chapter 5 summarizes the research results of the entire paper, analyzes the limitations of the research and looks forward to future research directions.

2.Related work

2.1. Research on general target detection based on YOLO series

Single-stage YOLO algorithms, with their advantages of end-to-end inference, strong real-time performance, and excellent balance between accuracy and speed, have become the mainstream algorithms for real-time object detection in industrial scenarios. From the first generation YOLOv1 to the latest YOLOv12, the model has been continuously iterated and optimized in terms of feature fusion mechanism, detection head structure, and training strategy. Early YOLO models had problems such as weak small object detection capability and unbalanced multi-scale feature fusion. Subsequent versions have effectively improved multi-scale object detection performance by introducing modules such as Cross-Stage Local (CSP) structure, Path Aggregation Network (PANet), and decoupled detection head^[5,6]. Among them, YOLOv8n, as a lightweight representative model, uses the C2f module to replace the traditional C3 module, optimizes gradient propagation and feature fusion effects, and combines anchorless prediction and classification-regression decoupled branches, achieving an excellent balance between lightweight deployment and detection accuracy, and is widely used in various mobile detection scenarios.

While YOLO models perform well in general scenarios, their fixed convolutional kernels and single feature enhancement mechanisms make them difficult to adapt to the extremely complex high-altitude environments. General-purpose YOLO models cannot dynamically adapt to the large-scale variations in cattle and sheep targets on the plateau, lack targeted suppression capabilities for complex background noise, and are prone to false detections such as missed detections of small distant targets and target-background confusion, severely hindering their practical application in intelligent monitoring scenarios on plateau pastures.

2.2. Current status of research on livestock target detection

With the rapid development of smart agriculture technology, deep learning object detection algorithms have been widely applied to tasks such as livestock and poultry monitoring, counting, and behavior recognition. Existing livestock detection research mostly focuses on standardized farm scenarios with controlled environments, targeting pigs, poultry, and penned cattle and sheep. Researchers have optimized models such as Faster R-CNN, Mask R-CNN, YOLOv5, and YOLOv8 to address issues like dense occlusion and small object detection. This type of research relies on experimental scenarios with simple backgrounds, stable lighting, and uniform target scale, resulting in highly adaptable algorithms, but they cannot be directly transferred to open grazing scenarios.

In the field of image data optimization, traditional image inpainting methods based on partial differential equations, total variational models, and block matching can only repair simple image defects such as minor

scratches and slight blurring, and cannot handle severe image degradation problems such as large-area occlusion and overexposure in high-altitude scenes. Deep learning inpainting methods based on GANs have become the mainstream solution for agricultural image restoration due to their powerful feature modeling and data generation capabilities. However, traditional GAN models suffer from problems such as unstable training gradients, mode collapse, and uncontrollable generation results, leading to texture distortion, structural breaks, and background artifacts in restored cattle and sheep images. While existing attention-enhanced GAN models can strengthen the ability to focus on target features, they lack uncertainty quantification mechanisms and adaptive feature fusion strategies specifically designed for high-altitude scenes, resulting in significant shortcomings in restoration effectiveness.

In summary, the current field of intelligent livestock monitoring lacks dedicated datasets and targeted algorithm optimization schemes adapted to open grazing scenarios on plateaus, and has not yet formed a complete technical system integrating “data augmentation-image restoration-precise detection”. This is the core research gap that this paper focuses on addressing.

3. Research methods

3.1. Construction of the QH-livestock plateau cattle and sheep dataset

To address the issues of scarce, limited, and poor-quality data in plateau livestock farming scenarios, this paper constructs a dedicated QH-Livestock plateau cattle and sheep dataset, using a typical plateau pastoral area in Xia Dawu Township, Maqin County, Golog Tibetan Autonomous Prefecture, Qinghai Province as the data collection region. The data collection area covers diverse scenarios including open natural pastures, intensive breeding areas, and herders' courtyards. High-definition monitoring equipment at multiple locations was used to collect images of cattle and sheep activities under different seasons, time periods, and weather conditions. Additionally, a small number of edge scene samples from publicly available datasets were supplemented to enrich the diversity of target poses, scales, and scenarios (See **Figure 1**).



Figure 1. Image data collected from the pastures in Qinghai.

The dataset contains 77,506 high-precision labeled RGB images, categorized into cattle and sheep targets. To unify data distribution and improve sample quality, the original images underwent standardized preprocessing: uniformly cropped and scaled to 640×640 pixels; adaptive histogram equalization was used for illumination correction to eliminate overexposure in strong light and underexposure in weak light; Gaussian filtering was used to suppress wind noise and equipment salt-and-pepper noise; and severely

blurred, distorted, or targetless invalid samples were removed. The dataset is randomly divided into training, validation, and test sets in an 8:1:1 ratio to ensure balanced sample distribution across various scenes, scales, and lighting conditions. Three specialized test subsets, small targets, strong light, and occlusion, are also included to accurately validate the model's detection performance in challenging high-altitude scenarios. Due to privacy concerns regarding herders, this dataset is not currently open-source.

3.2. Design of QHCSM-GAN image completion algorithm

To address the issues of strong light degradation, target occlusion, and feature loss in images of cattle and sheep on the plateau, this paper proposes the QHCSM-GAN image completion algorithm. The overall structure consists of a multi-path feature fusion encoder, a decoding module, and an evidence-theoretic trusted discriminator. The algorithm restores details in degraded images through multi-scale feature fusion and relies on evidence theory to quantify generation uncertainty, effectively suppressing artifacts and distortion. **Figure 2** shows the overall network architecture of QHCSM-GAN, fully illustrating the multi-path encoding branches, the CGA-fusion attention fusion module, the decoding output, and the trusted discriminator structure ^[7].

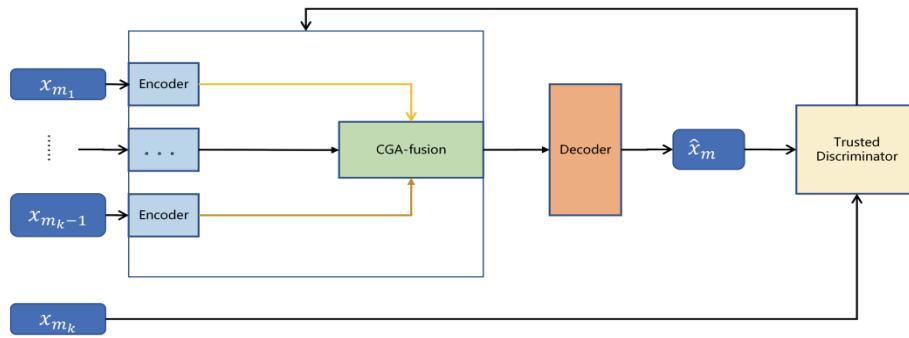


Figure 2. QHCSM -GAN.

3.2.1. Multi-channel feature fusion encoder

The encoder employs a K-way parallel autoencoder structure, comprising K-1 feature fusion branches and 1 guidance branch. Each branch extracts high-level semantic features and low-level texture detail features from the image through multi-layer coding modules, avoiding feature aliasing and detail loss. The network embeds a content-guided attention fusion module (CGA-fusion) that adaptively fuses cross-modal and multi-scale features, weighting and enhancing the effective features of cattle and sheep targets while suppressing interference from complex background noise in the plateau region (See **Figure 3**).

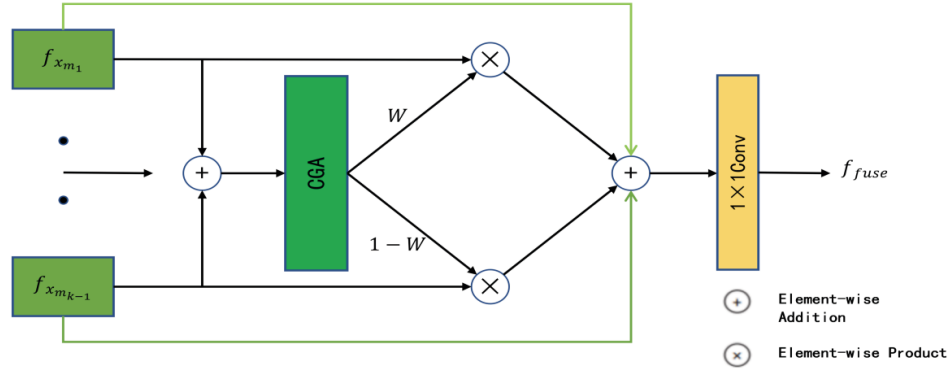


Figure 3. CGA -fusion module.

The guidance branch takes a real, clear image as input, extracts high-quality reference features, and performs deep supervision on the fusion encoder through feature alignment loss. The feature alignment loss is calculated as follows:

$$L_{feat} = \sum_{k \in \{h, l\}} \| F^k - \tilde{F}^k \|_1 \quad (3-1)$$

In the formula : F^k represents the features extracted by the multi-path feature fusion encoder, $\tilde{F}^k$ represents the ground truth features extracted by the guide autoencoder, $\| \cdot \|_1$ and represents the L1 paradigm .

The training of the entire multimodal conversion network is constrained by multiple loss mechanisms, including image-level reconstruction loss L_{img} , reconstruction loss for each modality feature L_{rec} , and the aforementioned feature alignment loss λ_{rec} . The total loss is defined as:

$$L_{MIN} = L_{img} + \lambda_{rec} L_{rec} + \lambda_{feat} L_{feat} \quad (3-2)$$

Among them , λ_{rec} and λ_{feat} are the balance coefficients.

The CGA fusion encoder effectively suppresses feature shifts caused by uneven lighting and image blurring in high-altitude areas through adaptive high- and low-frequency feature fusion; it enhances the stability of cattle and sheep target features through content-aware attention, mitigating the impact of posture and scale changes; and it weakens complex background interference through global attention, achieving early enhancement of foreground targets and background denoising, providing robust and clean feature representations for subsequent reconstruction.

3.2.2. Evidence theory credibility checker

Traditional generative adversarial networks (GANs) typically use Softmax to output classification confidence scores for their discriminators, which can easily lead to misclassifications due to “overconfidence” in the generated images. To address this issue and quantify the uncertainty in the generated results, this study employs a reliable classifier based on evidence theory as the discriminator D.

Evidence theory is an important mathematical tool for modeling and fusing uncertain information. It can quantify the conflict of image features, noise interference, and information loss through basic probability assignment, confidence function, and likelihood function. In the cattle and sheep image completion task, this

theory can effectively model the uncertainty caused by texture blur, occlusion, and illumination distortion, and improve the reliability of the discriminator in distinguishing between real and fake images.

The Dirichlet distribution, as the conjugate prior of a multinomial distribution, is often used for distribution modeling of probability vectors and classification confidence. Introducing it into a confidence discriminator can constrain the probability distribution of the evidence vector output by the network, calibrate the confidence level of the generated image, suppress unreasonable generation and artifacts, and make the completion result more consistent with the real distribution of the plateau cattle and sheep target in terms of structure, texture, and semantics.

This discriminator treats true/false binary classification as a subjective logical problem. For the input image, it doesn't directly output classification probabilities, but instead predicts a set of evidence vectors $\mathbf{e}$ to parameterize a Dirichlet distribution $Dir(p|\alpha)$. Here, the concentration parameter is denoted by $\hat{\alpha} = \mathbf{e} + 1$, and the class probability is represented by a simplex. Through this distribution, both the quality of classification belief p_k and the overall uncertainty u can be obtained simultaneously.

$$u = \frac{S}{S+1}, b = \frac{e_k}{S+1} \quad (3-3)$$

Where $S = \sum_{k=1}^K (e_k + 1)$ K is the number of categories.

The discriminator's training loss L_D consists of two parts: first, an ensemble cross-entropy loss based on the Dirichlet distribution L_{ECE} , designed to gather more evidence for the correct class; and second, a KL divergence loss L_{KL} , used to penalize evidence for incorrect classes, preventing it from approaching zero.

$$L_D = L_{ECE} + \lambda_{KL} L_{KL}$$

The balancing factor λ_{KL} is gradually increased during training to ensure the network fully explores the parameter space. Correspondingly, in addition to the adversarial constraints from D L_{adv} , the generator G is also subject to an additional evidence loss L_{evi} . This loss ensures that the images produced by the generator not only closely approximate reality in terms of discrimination, but also that the evidence contained in their internal features supports the "real" category, thereby achieving feature-level alignment.

$$L_{evi} = \left\| \mathbf{e}_{fake} - \mathbf{e}_{real} \right\|_2^2 \quad (3-5)$$

Finally, the overall loss function of the generator is:

$$L_G = L_{MIN} + \lambda_{adv} L_{adv} + \lambda_{evi} L_{evi} \quad (3-6)$$

By leveraging the dual constraints of the trusted discriminator (adversarial loss and evidence loss), this method can effectively quantify and utilize uncertainty during image generation, thereby synthesizing structurally accurate, detailed, and reliable missing modal images.

3.3. Improved YOLO-OC detection model design

This paper uses the lightweight YOLOv8n as the baseline model. To address the problems of difficulty in adapting to multi-scale targets and strong background interference in plateau scenes, the backbone network and the neck network are optimized respectively to construct the YOLO-OC detection model. The feature extraction capability is optimized by using ODConv full-dimensional dynamic convolution, and the key feature focusing is enhanced by combining the CBAM attention mechanism. **Figure 4** shows the complete

network structure of YOLO-OC, with the embedding positions of ODConv and CBAM and the overall network flow marked.

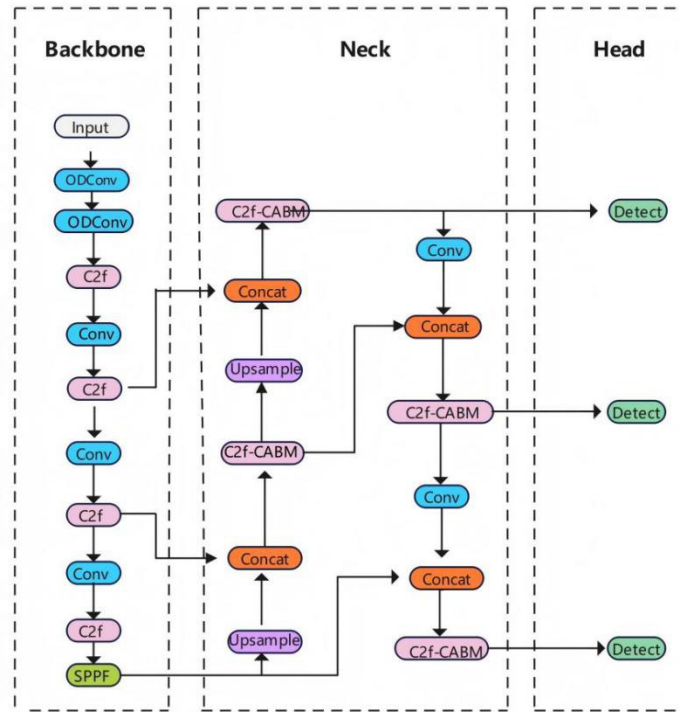


Figure 4. Optimization structure diagram of YOLO-OC model.

3.3.1. Backbone network optimization based on ODConv

Traditional fixed convolutional kernels can only extract features at a single scale, failing to adapt to the scale variations of large near-field targets and small distant targets such as cattle and sheep in the plateau region. This paper replaces the shallow fixed convolutions in the backbone network with full-dimensional dynamic convolutions (ODConv). Through a multi-dimensional attention mechanism involving space, channels, and convolutional kernels, convolutional weights are dynamically generated. This allows for adaptive adjustment of the response mode based on the input image features, effectively enhancing the extraction of detailed features from small targets and the aggregation of semantic features from large targets, thus solving the problem of insufficient scale adaptability of fixed convolutions (See **Figure 5**).

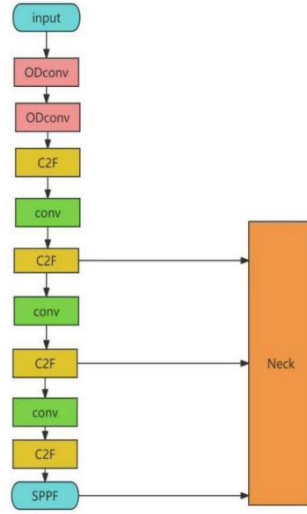


Figure 5. Structure diagram of ODC-FPN network.

ODConv generates dynamic weights through an attention mechanism, enabling adaptive adjustment of the convolutional kernel. The formula for generating dynamic weights is as follows:

$$W_{dynamic} = \sum_{i=1}^N \alpha_i \cdot W_i \quad (4-1)$$

Where is $W_{dynamic}$ the generated dynamic convolution kernel weight, is the number of basic weight vectors, α_i is the dynamic weight coefficient generated by the attention mechanism, W_i and is the i -th basic weight vector .

$$\sum_{i=1}^N \alpha_i = 1 \quad (4-2)$$

Through the above mechanism, ODConv can dynamically adjust the convolution kernel parameters based on the input features, thereby improving the model's adaptability to targets of different scales.

3.3.2. CBAM-based neck network optimization

To address the challenges presented by the QH-Livestock dataset, such as complex backgrounds in high-altitude pasture scenes, livestock coat colors closely resembling environmental textures, and the tendency for target features to weaken under strong light, this chapter optimizes the feature fusion process in the Neck section of YOLO v8 by embedding Convolutional Block Attention Modules (CBAMs) before and after the feature concatenation positions within the C2f module. By introducing a joint modeling mechanism of channel attention and spatial attention, the model's responsiveness to key regions and discriminative features of cattle and sheep is enhanced, while suppressing irrelevant interference from complex backgrounds and lighting variations. This improves the model's feature focusing ability and detection stability in high-altitude scenes.

CBAM is a lightweight attention mechanism module. Its core idea is to adaptively model the importance of different channels and spatial locations in the feature map through a concatenated structure of channel attention and spatial attention, and dynamically adjust the feature response weights to highlight target-related information, suppress background noise interference, and improve the model's feature representation ability in complex scenes. Compared with methods that only focus on single-dimensional feature enhancement,

CBAM can optimize feature representation from both “channel selection” and “spatial localization” levels, making it more suitable for feature enhancement in target detection tasks in complex environments.

CBAM first models the importance of the input feature map in the channel dimension through the Channel Attention Module (CAM). This module performs global average pooling and global max pooling on the input features, and then maps the resulting features into a shared multilayer perceptron to generate channel attention weights. The calculation process can be represented as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (4-3)$$

Where $\sigma()$ is the Sigmoid activation function, $MLP()$ is the multilayer perceptron, $AvgPool()$ is the global average pooling, and $MaxPool()$ is the global max pooling.

After channel-weighting, CBAM further models the important regions of the feature map in the spatial dimension through the Spatial Attention Module (SAM). This module first performs average pooling and max pooling operations on the channel-weighted feature map, then concatenates the results and feeds them into a convolutional layer to generate the spatial attention weight map. The calculation process can be represented as follows:

$$M_S(F) = \sigma(Conv([AvgPool(F); MaxPool(F)])) \quad (4-4)$$

$Conv$ This represents a convolution operation, $[;]$ indicating feature map concatenation.

Embedding this module into the feature fusion process of the Neck part helps to improve the effectiveness of multi-scale features in the transmission and fusion stages, thereby enhancing the model’s ability to distinguish targets under complex backgrounds, lighting changes, and local occlusion conditions.

As shown in **Figure 6**, by introducing CBAM into the Neck part, the model can further enhance its attention to the target region while maintaining the basic stability of the original network framework, providing more discriminative fusion features for subsequent detection heads, thereby improving the false detection and false negative detection problems in plateau scenes.

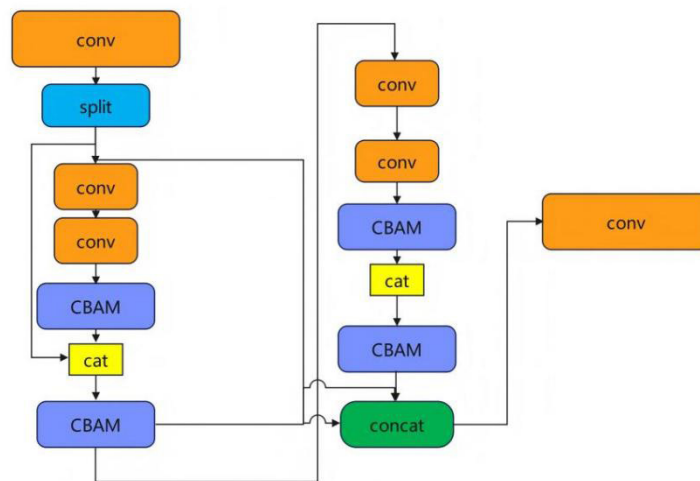


Figure 6. Network structure diagram of C2F-CBAM-Neck.

4. Experiment and results analysis

4.1. Experimental setup

All experiments in this paper were conducted on an Ubuntu 22.04 LTS system, using the PyTorch 2.1.2 deep learning framework and CUDA 11.6 parallel computing environment. The hardware configuration consisted of an Intel i7-8750H CPU, 32GB of RAM, and an NVIDIA RTX 4060Ti GPU. The experiments were conducted on the QH-Livestock dataset, with 37% of the training samples enhanced and optimized using the QHCSM-GAN algorithm. The model training parameters were uniformly set as follows: input image size 640×640 , SGD optimizer, initial learning rate 0.01, batch size 8, momentum coefficient 0.937, and a maximum of 300 training epochs.

For the image completion task, root mean square error (RMSE), structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and learned perceptual patch similarity (LPIPS) are selected as evaluation metrics; for the object detection task, precision (P), recall (R), mAP50, and inference speed (FPS) are selected to comprehensively evaluate the model's accuracy and real-time performance.

4.2. Experimental results of QHCSM-GAN image completion

To verify the superiority of the QHCSM-GAN algorithm, eight mainstream image completion algorithms, including StarGAN, Pix2Pix, and TokenFusion, were selected for comparative experiments. The tests were completed on the QH-Livestock, BraTS2020, and RaFD datasets^[8–14].

Table 1 presents the quantitative comparison results of the QH-Livestock dataset. In the visible light to infrared modal restoration task, QHCSM-GAN achieves an SSIM of 0.9345, a PSNR of 28.92 dB, and an LPIPS as low as 0.0825, significantly outperforming the comparison algorithms in all metrics. In the infrared-to-visible light restoration task, QHCSM-GAN achieves SSIM and PSNR of 0.8983 and 27.14 dB, respectively, demonstrating the best overall performance.

Table 1. Comparison of modal synthesis performance of various methods on the QH-Livestock dataset

Methods	Visible to Infrared			Infrared to Visible		
	SSIM↑	PSNR↑(dB)	LPIPS↓	SSIM↑	PSNR↑(dB)	LPIPS↓
StarGAN	0.8365	22.48	0.1615	0.8395	22.12	0.1735
CollaGAN	0.7887	21.43	0.2272	0.8091	22.88	0.1892
Hi-net	0.8521	23.15	0.1426	0.8478	23.02	0.1568
Pix2Pix	0.8798	24.89	0.1087	0.8645	24.83	0.1245
AttentionGAN	0.8568	23.37	0.1086	0.8231	21.23	0.1322
TSIT	0.8335	22.46	0.1625	0.8392	22.25	0.1735
CUT	0.8172	21.18	0.1334	0.8134	20.52	0.1591
TokenFusion	0.9011	25.67	0.0932	0.8435	23.58	0.1396
Ours(QHCSM-GAN)	0.9345	28.92	0.0825	0.8983	27.14	0.1160

Table 2 presents the ablation experiment results of QHCSM-GAN. Feature alignment loss and evidence loss are crucial to model performance; removing any constraint term leads to a significant decrease in image restoration accuracy, demonstrating that the dual constraint mechanism effectively ensures the structural consistency and generation reliability of the restored image. Cross-dataset experiments further validate the algorithm's generalization ability.

Table 2. Results of the QHCSM-GAN ablation experiment

Model configurations	SSIM↑	PSNR↑(dB)	LPIPS↓
QHCSM-GAN - feat -evi	0.9345	28.92	0.0855
QHCSM-GAN - feat	0.9113	26.81	0.1011
QHCSM-GAN - evi	0.9064	26.53	0.1022

Figure 7 shows the visual comparison results of image completion by various algorithms. The algorithm presented in this paper can effectively repair defects such as occlusion and overexposure in plateau images, restore the complete outline and texture details of cattle and sheep, and provide high-quality training samples for the detection model.



Figure 7. The image on the left is the original image; the image on the right is the one generated by QHCSM-GAN.

4.3. Results of YOLO-OC comparison experiment

To verify the detection performance of the YOLO-OC model, a comparative experiment was conducted with seven mainstream detection models, including Faster R-CNN, YOLOv5n, YOLOv7-tiny, YOLOv8n, YOLOv9-tiny, and YOLOv12^[15].

Table 3 presents the quantitative comparison results of the models. The YOLO-OC model in this paper achieved a precision of 84.6%, a recall of 81.2%, and a mAP50 of 82.5%, representing improvements of 6.1%, 8.9%, and 7.3%, respectively, compared to the baseline YOLOv8n. Compared to the latest lightweight YOLO model, YOLO-OC achieves a significant improvement in accuracy with minimal speed loss (10⁹ FPS), fully meeting the real-time monitoring needs of high-altitude pastures. Compared to the two-stage Faster R-CNN model, it has an absolute advantage in both accuracy and speed.

Table 3. Performance comparison of each model on the dataset

Model name	Accuracy	Recall rate	mAP50	Fps	Params
Faster R-CNN	72.3 %	58.6 %	65.4 %	15	41.8M
YOLO v5 n	74.5 %	68.2 %	71.3 %	84	1.9M
YOLO v7-tiny	76.8 %	71.5 %	74.2 %	87	6.2M
YOLO v8 (benchmark)	78.5 %	72.3 %	75.2 %	90	3.2M
YOLO v 9 -tiny	79.2%	74.8%	77.5%	115	2.0M
YOLO v 12	0.81.4%	76.5%	78.9%	120	2.6M
YOLO-OC (Optimized)	84.6 %	81.2 %	82.5 %	109	3.3M

Figure 8 shows the visualization comparison results of different detection models. Mainstream algorithms have serious problems with missed detections and false detections in small targets, strong lighting, and complex background scenes, while YOLO-OC can accurately identify cattle and sheep targets at multiple scales, and its anti-interference ability and scene robustness are significantly better than the comparison models.

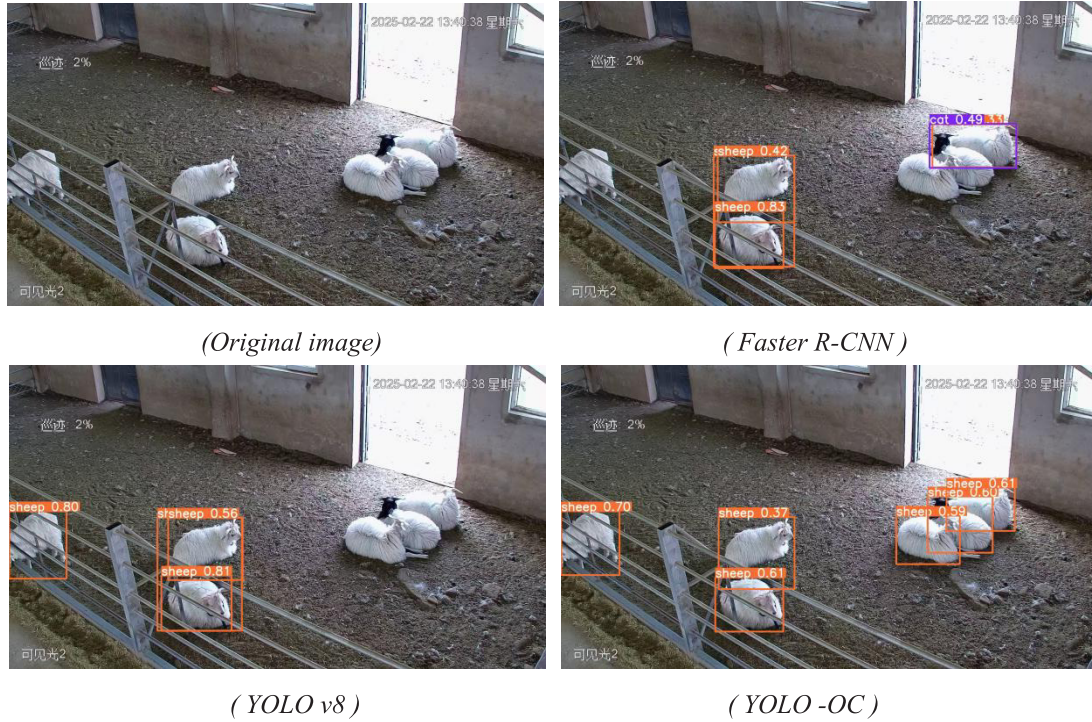


Figure 8. Comparative image visualization results.

4.4. Results of YOLO-OC ablation experiments

To verify the effectiveness and synergistic effect of the ODConv and CBAM modules, an ablation experiment was set up.

Table 4 presents the quantitative results of the ablation experiments. When only the ODConv module is introduced, the model's mAP50 is improved by 2.6%, demonstrating that dynamic convolution can effectively enhance multi-scale feature extraction capabilities. When only the CBAM module is introduced, the mAP50 is improved by 3.2%, verifying that the attention mechanism can efficiently suppress background noise. The model performance reaches its optimal level after the fusion of the two modules, forming a collaborative optimization mechanism of “dynamic feature extraction + precise feature focusing”.

Table 4. Results of the ablation experiment

Model configuration	Accuracy(P)	Recall rate(R)	mAP50	Reasoning speed(FPS)
YOLO v8 (benchmark)	78.5 %	72.3 %	75.2 %	90
YOLO v8 - ODConv	80.3 %	75.6 %	77.8 %	85
YOLO v8 - CBAM	81.7 %	74.2 %	78.4 %	86
YOLO-OC	84.6 %	81.2 %	82.5 %	109

Figure 9 shows the visual comparison results of the ablation experiment, further demonstrating that the dual-module improvement can significantly enhance the detection effect in challenging high-altitude scenarios.



Figure 9. Ablation image visualization results.

4.5. Training process analysis

Figure 10 shows the training loss and accuracy curves of the YOLO-OC model. The model converges quickly and is highly stable, with all accuracy metrics rising rapidly within 100 rounds and stabilizing after 200 rounds, showing no significant oscillations or overfitting. Thanks to the high-quality data augmentation and refined module optimization of QHCSM-GAN, the model possesses excellent generalization ability and stability for engineering deployment.

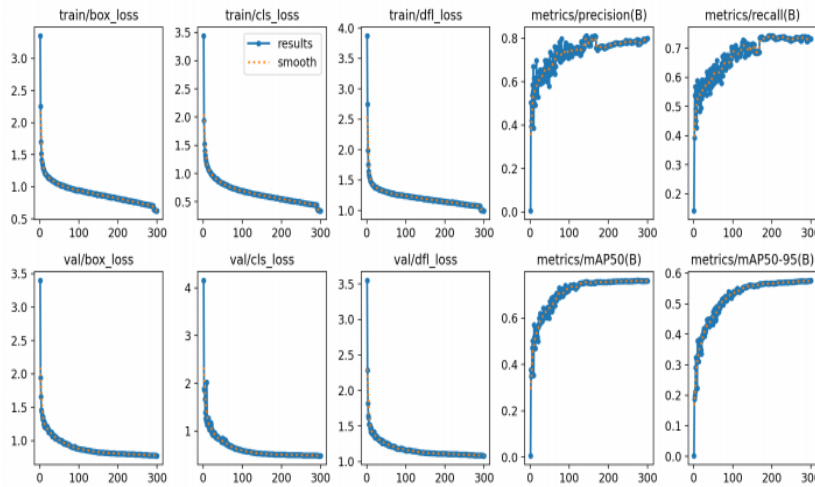


Figure 10. Performance trend chart.

5. Conclusions and outlook

To address the industry pain points in intelligent monitoring of livestock in high-altitude areas, such as poor image quality, scarcity of dedicated datasets, and low detection accuracy in complex scenes, this paper constructs an intelligent detection framework for cattle and sheep in high-altitude areas, integrating dataset

construction, image completion optimization, and target detection improvement. First, the QH-Livestock dataset, specifically designed for cattle and sheep in high-altitude areas, is constructed, filling the data resource gap in high-altitude grazing scenarios. Second, the QHCSM-GAN image completion algorithm is proposed, achieving high-fidelity restoration of degraded images in high-altitude areas through multi-path feature fusion and uncertainty constraints based on evidence theory, effectively improving the quality of training samples. Finally, a YOLO-OC detection model is designed based on YOLOv8n, integrating ODConv dynamic convolution and CBAM attention mechanisms to significantly improve the model's multi-scale detection capability and resistance to interference in complex backgrounds. Multiple experimental results demonstrate that the proposed framework exhibits high accuracy, good real-time performance, and strong robustness, effectively adapting to intelligent monitoring scenarios in open pastures in high-altitude areas.

This study still has certain limitations: the dataset does not adequately cover extreme rain and snow weather and low-light nighttime scenarios; the model's lightweight design still has room for improvement, and its ability to adapt to low-power edge deployments needs to be enhanced. Future research will focus on three aspects: first, expanding the dataset to enrich samples for challenging scenarios such as extreme weather and nighttime monitoring; second, further lightweighting the model structure to reduce the number of parameters and computational load, making it suitable for embedded edge device deployments; and third, integrating target tracking and behavior recognition algorithms to construct a comprehensive and multifunctional intelligent monitoring system for plateau animal husbandry, thereby facilitating the intelligent and digital transformation and upgrading of plateau animal husbandry.

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Goodfellow I, Pouget-Abadie J, Mirza M, et al., 2014, Generative Adversarial Nets. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2672–2680.
- [2] Tian Y, Ye Q, Doermann D, 2025, YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv*.
- [3] Li C, Zhou A, Yao A, 2022, Omni-Dimensional Dynamic Convolution. *arXiv:2209.07947*.
- [4] Chen M, Zhou G, Ma Y, 2024, An Improved YOLOv8 Small Target Detection Algorithm With Embedded CBAM Attention. *Computer Engineering and Science*, 46(03): 478–485.
- [5] Wang Y, Xu D, Wang Y, et al., 2025, TAF-ELAN: Tri-Attention Fusion Enhanced ELAN Network for Single Image Super-Resolution. *Proceedings of the 2025 7th International Academic Exchange Conference on Science and Technology Innovation*, 163–167.
- [6] Liu S, Qi L, Qin H, et al., 2018, Path Aggregation Network for Instance Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768.
- [7] Ma M, Liang L, Sheng X, 2025, MDF-Net: An Attention-Guided Multi-Scale Dual-Fusion Network for Retinal Vessel Segmentation. *Measurement*, 257(02): 118695.
- [8] Choi Y, Choi M, Kim M, et al., 2018, StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8789–8797.

- [9] Isola P, Zhu J, Zhou T, et al., 2017, Image-to-Image Translation with Conditional Adversarial Networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 5967–5976.
- [10] Tang H, Xu K, Li X, et al., 2019, AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5647–5656.
- [11] Jiang L, Li M, Lin L, et al., 2021, TSIT: A Simple and Versatile Framework for Image-to-Image Translation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 423–432.
- [12] Park T, Efros A, Zhang R, et al., 2020, Contrastive Learning for Unpaired Image-to-Image Translation. European Conference on Computer Vision, 319–334.
- [13] Kim J, Kim M, Kim D, et al., 2019, CollaGAN: Collaborative Generative Adversarial Network for Missing Image Data Imputation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–256.
- [14] Zhou T, Li Z, Wang H, et al., 2020, Hi-Net: Hybrid-Fusion Network for Multi-Modal MR Image Synthesis. IEEE Transactions on Medical Imaging, 39(12): 4014–4025.
- [15] Girshick R, 2015, Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440–1448.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research and Practice on Construction Pathways of Emerging Engineering Programs for the Intelligent Manufacturing Industry

Miaomiao Xin, Wanzhen Wang, Na Li, Ting Wang, Man Qiu, Xichao Yan, Dandan Zhang, Yuanyi Liu*

School of Intelligent Manufacturing and Control Engineering, Qilu Institute of Technology, Jinan, China

**Author to whom correspondence should be addressed.*

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Driven by the new round of the industrial revolution, the intelligent manufacturing industry is undergoing rapid iteration. As a national key strategic industry, its transformation and upgrading generate an urgent demand for compound and application-oriented emerging engineering talents, posing new requirements for the engineering education reform of local universities. Newly built undergraduate universities generally suffer from weak disciplinary accumulation, insufficient interdisciplinary integration of curriculum, lack of teachers' engineering practical experience, and superficial industry-education integration. Taking five intelligent manufacturing-related undergraduate majors in Qilu Institute of Technology as the research sample, this study constructs a professional construction system of "one core, two wings, and four dimensions" based on the emerging engineering education philosophy. Through optimizing talent training schemes, building interdisciplinary curriculum modules, cultivating dual-qualified teachers, and constructing industry-education integrated practical platforms, this paper explores a distinctive path for emerging engineering major construction in newly-built undergraduate universities. Practical verification indicates that the proposed system effectively solves the problems of disconnection between talent training and industrial demands and homogenized talent cultivation, improves the quality of application-oriented innovative talents, and provides practical demonstration and reference for emerging engineering construction in peer universities.

Keywords: Intelligent manufacturing; Emerging engineering education; Specialty construction; Industry-education integration; OBE concept

Online publication: Sept 3, 2026

1. Introduction

Currently, a new round of scientific and technological revolution and industrial transformation is accelerating at an unprecedented pace. Emerging technologies, including artificial intelligence, big data, and the industrial Internet, are deeply empowering the manufacturing sector. Leveraging initiatives such as the

Intelligent Manufacturing Engineering program and the “AI + Manufacturing” special action plan, the nation is cultivating new quality productive forces, advancing new-type industrialization, and driving the comprehensive digital, intelligent, and green transformation of manufacturing ^[1].

Intelligent manufacturing has emerged as the core lever for manufacturing upgrading. The rapid iteration of the industry has triggered a surging demand for compound engineering talents equipped with interdisciplinary knowledge, hands-on engineering capabilities, and innovative literacy, posing new and higher requirements for the reform of local higher engineering education. As a pivotal initiative aligned with national industrial strategies and responsive to industrial transformation, Emerging Engineering Education (EEE) focuses on breaking down the silos of traditional engineering education through interdisciplinary integration and deep industry–education collaboration ^[2]. Its core objective is to precisely cultivate application-oriented, innovative engineering talents tailored to the future intelligent manufacturing industry.

Newly established undergraduate institutions serve as the core providers of regional intelligent manufacturing talent. However, constrained by limited educational heritage, they face systemic challenges: weak disciplinary foundations, outdated curricula, a shortage of dual-qualified faculty, and shallow university-enterprise collaboration. The traditional, siloed engineering education model is ill-suited to meet the occupational demands of the intelligent manufacturing industry, where computing, communication, and control technologies converge, resulting in a pronounced industry-education disconnect.

Against this backdrop, this paper takes five intelligent manufacturing-related undergraduate programs at Qilu Institute of Technology as its practical foundation. It addresses the pain points of EEE construction in newly established undergraduate institutions through systematic pedagogical reform. Specifically, it builds an integrated education ecosystem uniting industry, academia, research, and application, and explores distinctive program-development pathways tailored to regional industrial needs. The ultimate goal is to provide a replicable practical paradigm for EEE advancement in local universities and the high-quality development of the regional intelligent manufacturing sector.

2. Research status and significance

2.1. Research status

The construction of EEE and the cultivation of intelligent manufacturing talents represent current research frontiers in global engineering education reform. Chinese academia has conducted extensive systematic research on the connotation, construction pathways, and talent cultivation models of EEE.

Professor Lin Jian from Tsinghua University systematically reviewed the core connotation, defining features, construction objectives, and overall framework of EEE. He clarified the priorities and development directions of EEE construction for different types of higher education institutions, providing a theoretical foundation for engineering program reform ^[3]. Lyngdorf et al. provided a systematic overview of digital transformation frameworks in engineering education, demonstrating how interdisciplinary fusion between emerging technologies and traditional engineering disciplines underpins the development of future-oriented engineering programs ^[4]. As highlighted in both the 2026 Global Future Higher Education Summit, universities must address the new talent cultivation requirements brought by industrial transformation through three key pathways: interdisciplinary integration, industrial alignment, and collaborative education ^[5]. Together, these strategies enable the development of EEE talent cultivation models characterized by sustainable competitiveness

and practical relevance. At present, domestic research predominantly focuses on EEE construction in key universities, while studies on targeted program development in newly established undergraduate institutions serving the regional intelligent manufacturing industry remain relatively scarce.

Engineering education reform in Europe and the United States began earlier and has evolved into a more mature system. The Harvard John A. Paulson School of Engineering and Applied Sciences focuses on frontier challenges in human societal development, prioritizes emerging interdisciplinary fields, dissolves the boundaries of traditional engineering programs, and has established a cross-domain engineering talent cultivation system^[6]. American engineering education has developed standardized, systematic models for ethics education and innovation education, widely adopted and highly innovative, that serve as a benchmark for global engineering education reform. The Massachusetts Institute of Technology (MIT) pioneered the vision of engineering education reform for future industries, advocating that the focus of engineering education should shift from adapting to existing industries to anticipating future ones. MIT aims to build a forward-looking, innovative new engineering education system that places strong emphasis on cultivating students' ability to solve complex engineering problems and their creative thinking^[7]. Overall, European and American engineering education emphasizes interdisciplinary integration, industry–education collaboration, and innovation empowerment. However, as these studies are grounded in the industrial systems and educational models of developed countries, they cannot be directly applied to the regional industrial characteristics and operational realities of newly established undergraduate institutions in China.

2.2. Research significance

Currently, the development of intelligent manufacturing-related EEE programs in newly established undergraduate institutions still faces systemic shortcomings that constrain the quality of regional industrial talent cultivation.

First, disciplinary support is weak. With limited educational heritage, traditional engineering programs in these institutions are dominated by single manufacturing disciplines, lacking interdisciplinary curricula adapted to intelligent manufacturing development. This makes it difficult to align with the demands of the new-generation information technology and intelligent equipment industries^[8].

Second, faculty development lags behind. Most full-time instructors emphasize theoretical instruction and lack hands-on engineering experience. Their insufficient grasp of cutting-edge intelligent manufacturing technologies and industry standards renders them unable to meet the talent cultivation requirements of emerging industry formats^[9].

Third, industry–education integration remains superficial. The university–enterprise collaborative education mechanism is underdeveloped, and high-quality enterprise projects and authentic engineering scenarios have not been deeply integrated into teaching. As a result, students' complex engineering problem-solving abilities and innovative literacy are insufficiently cultivated^[10].

Against this backdrop, this paper is grounded in the developmental needs of the intelligent manufacturing industry, taking five intelligent manufacturing-related programs at Qilu Institute of Technology as its practical carriers. It implements systematic pedagogical reform across four dimensions: talent cultivation plans, curriculum systems, faculty development, and practical teaching—addressing the gaps in EEE construction at newly established undergraduate institutions. The research findings can precisely match regional industrial talent demands and effectively improve the quality of applied innovative talent

cultivation. Concurrently, they provide a referable practical paradigm and replicable model for EEE program construction and transformation upgrading in similar local universities.

This study takes serving the development of the regional intelligent manufacturing industry and cultivating high-quality applied innovative talents as its core objectives. It deeply advances interdisciplinary integration across mechanical engineering, automation, computer science, electronic communication, and related fields, and comprehensively promotes reform and innovation in teaching content, curriculum systems, faculty teams, teaching methodologies, and practical platforms.

Focusing on the core requirements of EEE construction, the study strives to achieve the “Five First-Class” construction objectives:

- (1) Building a first-class dual-qualified faculty team.
- (2) Establishing a first-class undergraduate education system.
- (3) Developing first-class distinctive intelligent manufacturing programs.
- (4) Creating first-class interdisciplinary integrated courses and cultivating first-class industry-adapted engineering talents.

Ultimately, this research aims to construct a replicable and scalable EEE program construction paradigm for newly established undergraduate institutions, comprehensively elevating the quality of regional intelligent manufacturing talent cultivation.

3. Reform scheme design and implementation methods

3.1. Overall reform scheme design

Grounded in the educational positioning of newly established undergraduate institutions and the developmental needs of the intelligent manufacturing industry, this study innovatively constructs an EEE program construction reform system defined as “One Core, Two Wings, Four Dimensions”.

Taking the EEE construction philosophy as the core guidance and leveraging the university’s educational resources and enterprise industrial resources as the two supporting wings, comprehensive reforms are implemented across four dimensions: talent cultivation plans, curriculum system development, faculty team cultivation, and practical innovation platforms. This forms a closed-loop, systematic program construction pathway that fully integrates the education chain, talent chain, industrial chain, and innovation chain—achieving synchronized advancement between EEE program construction and industrial development.

3.2. Specific reform implementation methods

3.2.1. Interdisciplinary integration empowerment: Optimizing the “five integrations” talent cultivation system

Guided by industrial demands, the talent cultivation plan is optimized through research validation, university–enterprise collaboration, and expert empowerment. By visiting intelligent manufacturing enterprises, investigating peer institutions, and inviting industry experts and enterprise technical backbone to participate in program validation, a four-in-one cultivation framework is reconstructed—comprising general education, disciplinary foundational education, specialized direction education, and practical innovation education.

Five major integration reforms are deeply advanced:

- (1) Integrating program characteristics with industry needs.

- (2) Cutting-edge technologies with traditional knowledge.
- (3) Innovation and entrepreneurship education with specialized education.
- (4) Practical education with enterprise resources, and in-class cultivation with extracurricular development.

This thoroughly breaks the homogeneous model of traditional talent cultivation. Focusing on the core competencies of intelligent manufacturing positions, graduation requirements and competency indicators are refined, constructing a new personalized, interdisciplinary, and innovative talent cultivation system that precisely matches the needs of core positions such as intelligent production line operation and maintenance, intelligent equipment R&D, and industrial data processing.

3.2.2. Concurrent recruitment and cultivation: Building a dual-qualified faculty team

Adhering to the faculty construction principle of “rooted in cultivation, focused on recruitment, and shared between university and enterprise”, the faculty structure is comprehensively optimized, and overall faculty capabilities are enhanced.

In terms of internal cultivation, a standardized mechanism for faculty enterprise practice is established. Each year, 1–2 professional faculty members are selected to engage in full-time engineering practice at intelligent manufacturing enterprises, accumulating frontline project experience, mastering cutting-edge industry technologies, and addressing gaps in engineering practical experience. Leveraging on-campus research platforms and educational reform projects, faculty are encouraged to conduct EEE pedagogical research and technological R&D, simultaneously improving their teaching and research capabilities.

Adopting flexible recruitment and elastic employment models, highly skilled enterprise talents, master craftsmen, and technical experts are recruited into the part-time faculty team, undertaking teaching tasks such as practical instruction, course lectures, and project guidance—thereby compensating for the insufficient industry experience of on-campus faculty.

Through two-way talent flow between university and enterprise and dual empowerment of teaching and research, a well-structured, highly capable dual-qualified faculty team adapted to EEE talent cultivation is built.

3.2.3. OBE philosophy guidance: Constructing interdisciplinary curriculum modules

Grounded in the Outcome-Based Education (OBE) philosophy, with student competency outcomes as the core, a modular, interdisciplinary, and cutting-edge curriculum system is reconstructed.

Breaking down the disciplinary silos of traditional mechanical, automation, and communication programs, the three core technology systems—computing, communication, and control—are integrated to construct distinctive interdisciplinary curriculum modules. Cutting-edge intelligent manufacturing technologies such as big data analytics, artificial intelligence, machine vision, 5G industrial communication, and the industrial Internet of Things are fully integrated into the curriculum. Nine interdisciplinary and cross-program elective courses, including “Artificial Intelligence Technology and Applications”, are added to fill the gap in cutting-edge technologies within traditional curricula.

Research findings, enterprise projects, and industry standards are promoted for conversion into teaching resources, and distinctive textbooks such as “Artificial Intelligence Technology and Applications” are developed. Using real engineering projects as carriers, a progressive project case system is established: “in-class foundational experiments → intensive specialized practice → graduation design innovation”. Abandoning traditional theory-cramming pedagogy, project-based, case-based, and inquiry-based teaching

models are implemented. Course assessment methods are optimized, and a process-oriented, diversified, and comprehensive evaluation system is constructed—comprehensively enhancing students’ ability to apply interdisciplinary knowledge and solve engineering problems.

3.2.4. University–enterprise co-construction: Building a four-in-one practical teaching platform

Leveraging provincial-level research and teaching platforms—including the Shandong Provincial Sino-German Smart Factory Application Engineering Research Center, the Shandong Provincial Key Laboratory of Industrial Big Data and Intelligent Manufacturing, and the Shandong Provincial University International Joint Laboratory of Industrial Internet of Things—university-enterprise strategic cooperation is deepened. On-campus and off-campus practical resources are integrated to build a four-in-one practical innovation teaching system: on-campus training platforms, provincial research platforms, industrial college platforms, and enterprise practice platforms.

Authentic enterprise production scenarios, engineering challenges, cutting-edge equipment, and technical projects are fully integrated into the entire process of practical teaching. Outdated practical content is updated, practical teaching models are innovated, and practical teaching management mechanisms are refined.

Students are organized to participate in phased practical activities: enterprise cognitive internships, on-the-job training, project R&D, and graduation design. A full-chain industry–education integration practice system of “cognition → practice → innovation” is constructed, comprehensively tempering students’ engineering practical abilities, teamwork capabilities, and innovative R&D competencies.

4. Reform practice outcomes

4.1. Reform practice achievements

Through systematic and standardized EEE program construction reform practices, the construction quality of five intelligent manufacturing-related undergraduate programs at Qilu Institute of Technology has significantly improved, the adaptability of talent cultivation has been greatly enhanced, and favorable pedagogical reform outcomes have been achieved.

An interdisciplinary curriculum system adapted to the intelligent manufacturing industry has been successfully constructed, breaking down traditional disciplinary silos and forming a new multi-dimensional integrated talent cultivation model. Program characteristics have become more distinctive, and alignment with regional industries has markedly improved.

Through enterprise practice, external recruitment, and internal cultivation, the coverage rate of engineering practical experience among full-time faculty has substantially increased. The proportion of dual-qualified faculty continues to grow, and the part-time faculty team effectively supplements industry teaching resources, resulting in a continuously optimized faculty structure.

The practical teaching system has become more comprehensive, enabling students to deeply engage with authentic engineering scenarios and cutting-edge technologies in intelligent manufacturing. Their complex engineering problem-solving abilities and innovative practical capabilities have significantly improved.

The quantity and quality of student participation in intelligent manufacturing disciplinary competitions and innovation and entrepreneurship projects continue to rise. The proportion of graduation designs aligned with actual enterprise projects has substantially increased. Graduate employment rates, major-job matching

rates, and enterprise satisfaction have steadily grown, effectively alleviating the shortage of interdisciplinary talents in the regional intelligent manufacturing industry.

The “One Core, Two Wings, Four Dimensions” EEE program construction system formed by this reform provides practical reference for the transformation and upgrading of EEE programs in similar newly established undergraduate institutions across the province, and possesses strong scalability value.

4.2. Existing issues and reflective improvement

In the phased reform practices, several issues remain that urgently need improvement.

First, the construction of interdisciplinary course teaching teams still requires strengthening. Some interdisciplinary courses place high demands on faculty’s comprehensive knowledge reserves, and faculty’s interdisciplinary teaching capabilities still need continuous enhancement.

Second, there remains room for deepening university–enterprise cooperation. The long-term incentive mechanism for enterprise participation in talent cultivation has not been fully established, and the integration of some enterprise projects into teaching is not sufficiently thorough.

Third, the personalized talent cultivation system still needs refinement. The stratified cultivation model for students with different academic foundations and different career orientations requires further optimization.

Going forward, targeted efforts will be made to address these shortcomings: improving the long-term mechanism of university–enterprise collaborative education, strengthening interdisciplinary faculty cultivation, refining stratified and categorized talent cultivation plans, and continuously deepening EEE pedagogical reform.

5. Conclusion

Advancing EEE program construction in response to the transformation and upgrading needs of the intelligent manufacturing industry represents a core pathway for newly established undergraduate institutions to adapt to regional economic development, improve the quality of applied talent cultivation, and highlight their distinctive educational identity.

Grounded in the educational shortcomings of newly established undergraduate institutions and the pain points of regional industrial talent demand, this paper takes intelligent manufacturing-related programs at Qilu Institute of Technology as practical carriers and innovatively constructs the “One Core, Two Wings, Four Dimensions” EEE program construction system. By optimizing the Five Integrations talent cultivation plan, building interdisciplinary curriculum modules, cultivating a dual-qualified faculty team, and developing a four-in-one industry–education integration practical platform, this study systematically addresses the core challenges facing engineering programs in newly established undergraduate institutions: low industrial adaptability, weak disciplinary integration, insufficient faculty practical experience, and lagging practical education.

Reform practices demonstrate that this construction pathway aligns with the operational realities of newly established undergraduate institutions and precisely meets the demand for interdisciplinary, innovative talents in the intelligent manufacturing industry. It effectively improves talent cultivation quality and supports the intelligent transformation and upgrading of regional manufacturing.

Going forward, the university will continue to deepen EEE construction reform, grounded in the new

development trends of Shandong Province's intelligent manufacturing industry. It will dynamically optimize the curriculum system and talent cultivation model, continuously deepen industry–education integration and university–enterprise collaborative education, and strengthen interdisciplinary innovative talent cultivation. The university will steadily refine the EEE program construction system for newly established undergraduate institutions, continuously delivering high-quality applied engineering talents for the high-quality development of the regional intelligent manufacturing industry. Furthermore, it will amplify the demonstrative and spillover effects of educational reform, promoting high-quality development of engineering education in local universities.

Funding

2023 Shandong Provincial Undergraduate Teaching Reform Research Project (Project No.: Z2023209); 2025 Shandong Provincial Undergraduate Teaching Reform Research Project (Project No.: M2025723)

Disclosure statement

The authors declare no conflict of interest.

References

- [1] Ministry of Industry and Information Technology, Cyberspace Administration of China, National Development and Reform Commission, et al., 2026, Implementation Opinions on the “AI + Manufacturing” Special Action, visited on 2026-04-18, https://www.miit.gov.cn/cms_files/filemanager/1226211233/attach/20261/735011cb05584824b3cb2a45666a58ce.pdf.
- [2] Hernández-de-Menéndez M, Escobar Díaz C, Morales-Menéndez R, 2020, Engineering Education for Smart 4.0 Technology: A Review. *International Journal on Interactive Design and Manufacturing (IJDeM)*, 14: 863–879.
- [3] Zheng L, Lin J, 2021, Curriculum Design in New Engineering Education: A Case Study of Two Emerging Engineering Programs in China. *Proceedings of REES AAEE 2021*.
- [4] Lyngdorf N, Jiang D, Du X, 2024, Frameworks and Models for Digital Transformation in Engineering Education: A Literature Review Using a Systematic Approach. *Education Sciences*, 14(5): 519.
- [5] UNESCO International Centre for Higher Education Innovation, 2026, Global Future Higher Education Summit Convenes in Shenzhen to Jointly Build a New Ecosystem for AI-Driven Industry-Academia Collaborative Talent Development, visited on 2026-06-10, <https://en.ichei.org/en/news/information/4217.html>.
- [6] Ming X, van der Veen J, MacLeod M, 2025, Competencies in Interdisciplinary Engineering Education: Constructing Perspectives on Interdisciplinarity in a Q-Sort Study. *European Journal of Engineering Education*, 50(2): 406–427.
- [7] Anand M, Abel C, Sabuncu A, 2024, Cultivating Innovators—Unveiling the Hidden Potential of ‘Innovation Through Making’ in Engineering Education. *The Future of Engineering Education 2024 Annual Conference & Exposition*.
- [8] Abrorov A, Saidov A, Meyliyeva N, et al., 2025, Optimizing Mechanical Engineering Curriculum Design Using AI and Machine Learning. *AIP Conference Proceedings*, 3268: 070036.
- [9] Quintero W, Maldonado J, 2024, Competencies of the Engineer in Industry 4.0 Context: A Systematic Literature

Review. Production, 34: e20230051.

- [10] Vuoriainen A, Rikala P, Heilala V, et al., 2025, The Six C's of Successful Higher Education-Industry Collaboration in Engineering Education: A Systematic Literature Review. *European Journal of Engineering Education*, 50(1): 26–50.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A Review of Path Planning Algorithms for Mobile Robots

Xinrui Li*

Zhangjiagang Campus, Jiangsu University of Science and Technology, Zhenjiang 212003, China

*Corresponding author: Xinrui Li, 3783574526@qq.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Path planning for mobile robots is a key technology for achieving autonomous navigation. It aims to search for a collision-free optimal path from the starting point to the target point in complex environments. This paper provides a systematic review of existing path planning algorithms. First, according to the principles of path planning algorithms, the existing methods are categorized into classical algorithms, intelligent optimization algorithms, and artificial intelligence algorithms. Second, the basic principles, advantages, disadvantages, and improvement strategies proposed by other scholars to address the limitations of each type of algorithm are elaborated in detail. Finally, based on the trend of algorithm fusion, the development potential of algorithm fusion mechanisms is prospected. The review concludes that a single algorithm struggles to adapt to complex and dynamic environments, and the integration of the advantages of multiple algorithms has become the current research trend. In the future, improving the real-time performance and robustness of algorithms in dynamic and unstructured environments remains a major challenge.

Keywords: Mobile robot; Path planning algorithm; Classical algorithm; Intelligent optimization algorithm

Online publication: Sept 9, 2026

1. Introduction

With the advent of the “Industry 4.0” era and the intelligent service age, the application of mobile robots in industrial manufacturing, intelligent logistics, medical services, and home assistance has become increasingly widespread^[1]. As a core technology for autonomous navigation, path planning’s primary task is to find a collision-free path or trajectory from the initial state to the target state for a robot with specific kinematic or dynamic constraints in complex environments^[2]. The key to achieving autonomous, efficient, and safe operation lies in advanced motion planning and control algorithms. At the same time, it imposes high requirements on the real-time performance, robustness, and optimality of path planning. How to efficiently plan collision-free paths in unknown or partially known environments has become a research hotspot^[3]. The quality of path planning algorithms directly determines the intelligence level and application scope of mobile

robots. Related research holds significant theoretical and engineering value ^[4].

This paper systematically reviews the mainstream algorithms for mobile robot path planning, covering the principles, advantages, disadvantages, and improvement strategies of three major categories: classical algorithms, intelligent optimization algorithms, and artificial intelligence algorithms (as shown in **Figure 1**). It further discusses algorithm fusion mechanisms and provides an outlook on future trends based on existing research. The aim is to provide reference and inspiration for research on mobile robot path planning and to promote the development of this field.

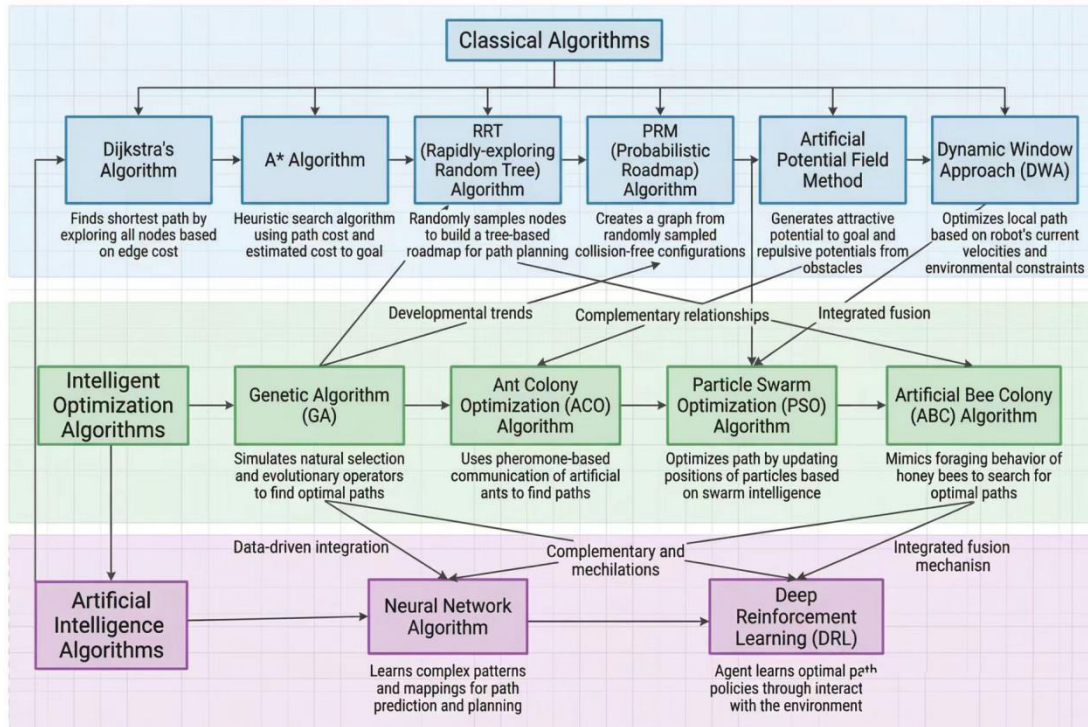


Figure 1. Classification of mobile robot path planning algorithms.

2. Path Planning algorithms for mobile robots

2.1. Classical algorithms

2.1.1. Graph search-based algorithms

With the rapid development of artificial intelligence, path planning has become one of the key technologies in mobile robot applications ^[5]. Graph search-based algorithms solve problems by systematically exploring graph structures, with core strategies including breadth-first search, depth-first search, heuristic search, and Dijkstra's algorithm. This paradigm laid the foundation for mobile robot path planning, among which the most representative algorithms are Dijkstra's algorithm and the A* algorithm ^[6].

Dijkstra's algorithm is a single-source shortest path graph search algorithm. As the theoretical cornerstone of shortest path searching, it employs a breadth-first strategy to expand nodes layer by layer, ensuring the acquisition of the optimal path in known environments ^[7]. However, when dealing with large-scale environments, its unique mechanism leads to high computational complexity and reduced efficiency.

To address the above limitations of Dijkstra's algorithm, Li proposed a relaxed Dijkstra algorithm

based on grid environment modeling, which is used to solve real-time path planning problems for mobile robots in large-scale and densely obstructed working environments ^[8]. Luo et al. proposed an extended Dijkstra algorithm that simulates the terrain environment through Delaunay triangulation, constructs a two-dimensional developable and passable corridor on the surface, and solves for the optimal path of this corridor, thereby improving the accuracy of surface-optimized paths ^[7]. Gong et al. found that when mobile robots move along paths planned by the Dijkstra algorithm in complex environments, frequent turning problems often occur ^[9]. Therefore, they proposed a smooth path planning method based on the Dijkstra algorithm by utilizing geometric topology methods combined with actual scene information. This method uses the Dijkstra algorithm to search for the optimal path in a discrete graph as a guide path, thereby improving algorithm efficiency. However, traditional smoothing techniques often result in poor curvature handling and suboptimal trajectories. Ahmad et al. proposed a Dijkstra algorithm integrated with dynamic waypoint allocation using piecewise cubic Bézier curves ^[10]. Based on the path constructed by the Dijkstra algorithm, this method achieves greater adaptability across different scenarios through context-aware path allocation and iterative optimization based on curve smoothing.

The A* algorithm is a widely used heuristic algorithm. It introduces a heuristic function on the basis of Dijkstra's algorithm. By prioritizing the expansion of nodes with the smallest total cost through the cost function, it significantly improves search efficiency. The cost function of the A* algorithm is defined as:

$$f(n) = g(n) + h(n) \quad (1)$$

where $g(n)$ is the actual cost from the start point to node n , and $h(n)$ is the heuristic estimated cost from node n to the goal.

The A* algorithm performs well in static environments, but its performance highly depends on the design of the heuristic function. In complex environments, it often suffers from excessive node expansion, resulting in increased computational burden, and it struggles to effectively handle dynamic obstacles.

Although the A* path planning algorithm can generate a collision-free path, it still has several shortcomings, such as suboptimal paths, too many redundant points, and insufficient smoothness. To address the deficiencies of the A* algorithm, Zhang et al. adopted a bidirectional optimization method to remove useless nodes, shorten the path length, reduce the number of inflection points, and enhance path smoothness, proposing an improved version of the A* algorithm ^[5]. Zhao et al. eliminated redundant nodes and unnecessary inflection points in the path to obtain a path containing only the start point, necessary inflection points, and the goal point, and further optimized the path planned by the traditional A* algorithm using geometric methods ^[11]. Since the A* algorithm has difficulty adapting to complex scenarios and fails to meet real-time requirements, Ou et al. improved the A* algorithm for obstacle space path planning by equipping the mobile robot platform with LiDAR and an inertial measurement unit, solving the problems of excessive turning points and slow search speed in traditional A* algorithms ^[12]. Yin et al. innovatively adopted a grid environment modeling method to improve the traditional A* algorithm ^[13]. The improved A* algorithm demonstrates higher search efficiency and fewer traversed nodes compared to both the traditional A* and bidirectional A* algorithms, thereby enhancing the efficiency of mobile robots.

2.1.2. Sampling-based algorithms

Sampling-based algorithms efficiently approximate solutions to complex problems by randomly selecting data samples. They do not rely on systematic traversal but utilize probabilistic and statistical principles,

making them suitable for high-dimensional, computationally intensive, or imprecise model scenarios. Their core idea is to trade acceptable error for time and resource efficiency. The results exhibit probabilistic convergence rather than absolute precision^[14]. The Probabilistic Roadmap (PRM) is one of the classic path planning methods. Its advantages include simple principles, probabilistic completeness, fast planning speed, and the ability to form asymptotically optimal paths. However, it performs poorly in dynamic obstacle avoidance^[15].

Francis et al. adopted a hierarchical robot navigation approach, using PRM as a planning tool and reinforcement learning (RL) for robot control^[16]. They proposed a PRM-RL long-range navigation method equipped with AutoRL, which completed short-range obstacle avoidance tasks. To meet various navigation tasks in dynamic scenarios, Liu et al. required mobile robots to achieve high planning success rates, fast planning, dynamic obstacle avoidance, and shortest paths^[15]. They adopted a hierarchical planning concept by introducing D* during the network construction and planning process of PRM, and proposed the PRM-D* method to improve dynamic obstacle avoidance performance. Facing the difficulty of offline path planning algorithms in obtaining available paths in real time, Ye et al. proposed a PRM algorithm based on an obstacle potential field sampling strategy, called the Obstacle Potential Field Probabilistic Roadmap method^[17]. Chen et al. adopted Bézier curves to enhance the smoothness of paths generated by PRM^[14]. To address the shortcomings of PRM in complex environments, they dynamically adjusted PRM sampling points to improve the success rate of path detection. Experimental results showed that the improved PRM algorithm outperforms the original PRM, Rapidly-exploring Random Tree Star (RRT*), and bidirectional random expansion methods in path generation.

Compared with the PRM method, RRT* has attracted widespread attention in trajectory planning due to its asymptotic optimality. Wang et al. addressed the randomness of path generation in the RRT-Connect algorithm and the slow convergence speed of the B-RRT* algorithm by introducing an intelligent sampling function to enhance the directionality of random tree expansion^[18]. They proposed an efficient path planning algorithm based on an improved B-RRT*. This algorithm incorporates a fast expansion strategy into the EB-RRT* framework, allowing the use of RRT-Connect in free space and the improved RRT* in obstacle space, thereby improving expansion efficiency while avoiding local optima. Khuat et al. introduced a new algorithm called MultiRRT to solve collaboration requirements during path searching^[19]. This algorithm extends RRT by incorporating two new mechanisms—node reduction and Bézier interpolation—to ensure the feasibility and optimality of the generated paths. Yu et al. improved the traditional PQ-RRT* algorithm by combining the posture adjustment angle of the mobile robot and introducing a new distance evaluation function, proposing a hybrid planning algorithm named New Potential Fast RRT*^[20]. Bian et al. proposed a Quadtree Rapidly-exploring Random Tree Star algorithm, which shortens the trajectory length of RRT*, improves node quality, and reduces trajectory cost^[21]. It performs excellently in terms of node count, convergence time, and path length, thereby enhancing the potential of the algorithm in path planning.

2.1.3. Local optimization-based algorithms

Local optimization algorithms aim to start from an initial solution and iteratively search and move toward better solutions within its “neighborhood,” quickly converging to a local optimum^[22]. These methods are highly efficient and easy to implement, serving as core solving modules for many complex optimization problems. However, their results heavily depend on the initial point and are prone to falling into local optima

rather than global optima.

The Artificial Potential Field (APF) method is a local approach for real-time robot path planning. It assumes that the goal point generates an attractive force and obstacles generate repulsive forces, together forming a virtual force field. The robot is treated as a point in the field and moves toward the goal under the guidance of the resultant force. However, traditional APF algorithms are highly susceptible to local minima, which reduces motion reliability in complex environments. To address this issue, Ma et al. proposed a novel hybrid obstacle avoidance algorithm called Deflection Simulated Annealing-Adaptive Artificial Potential Field by combining an improved simulated annealing mechanism with an enhanced APF model and embedding a direction deflection mechanism during the annealing process ^[23]. Yang et al. proposed an improved APF algorithm based on paths generated by multi-objective models and quadratic programming, which improved path smoothness and reduced the time required for path planning and trajectory optimization ^[24]. Zhai et al. addressed the problem of avoiding dynamic and static obstacles in unstructured environments by adding virtual target points and establishing a water-drop-shaped repulsive potential field function ^[25]. They proposed a local obstacle avoidance trajectory planning method based on an improved APF algorithm. In large-scale environments, goal searching and obstacle avoidance based on RL dynamic optimization decision-making are still limited by slow convergence and low computational efficiency. To solve this problem, Jin et al. proposed a hybrid framework combining RL and APF, named the SAC-APF hybrid framework, which improved the goal search algorithm ^[26]. Experiments showed that this method significantly outperforms standalone RL algorithms in both goal search efficiency and obstacle avoidance performance.

In recent years, autonomous mobile robots have significantly increased in real-world adoption due to their ability to perform diversified tasks. A critical aspect of their large-scale application is effective local path planning. In this context, the Dynamic Window Approach (DWA) is widely regarded as a robust local path planning solution.

When mobile robots face changes in environmental conditions such as congestion levels, road width, and obstacle density, the evaluation module may select inefficient paths or even cause collisions. To overcome this challenge, Kobayashi et al. proposed dynamic weight coefficients based on Q-learning ^[27]. This algorithm can dynamically select optimal paths and weight coefficients using a pre-learned Q-table to better adapt to different environmental conditions. Chang et al. introduced a bidirectional search strategy and a variable-weight trajectory evaluation function to address the low efficiency of traditional obstacle avoidance algorithms in unknown complex environments ^[28]. They proposed an improved DWA algorithm and integrated it with the Optimal Reciprocal Collision Avoidance method, ultimately presenting an enhanced DWA fusion algorithm. To solve the problem that the original DWA algorithm cannot balance safety and speed due to fixed parameters in complex environments, Gong et al. proposed an improved DWA algorithm by modifying the original velocity cost function and adding a distance cost function for the target point, which is used for local obstacle avoidance of robots ^[29]. Regarding the difficulty of satisfying global optimality and real-time obstacle avoidance in dynamic environments, Duan Qiuhui proposed a fusion algorithm combining an improved D* algorithm and an improved DWA ^[30]. This algorithm first rasterizes the environment map for modeling; then optimizes the cost function in the D* algorithm by establishing a regional obstacle complexity quantification index vector; finally, it designs the evaluation function of the dynamic window method based on key node information from the globally optimal path, thereby quickly

planning a globally optimal smooth path.

2.2. Intelligent optimization algorithms

2.2.1. Genetic algorithm

The Genetic Algorithm (GA) is a meta-heuristic algorithm that simulates the processes of selection, crossover, and mutation in biological evolution. It can be used as a global random search method for optimizing mobile robot paths ^[31]. Its flowchart is shown in **Figure 2**.

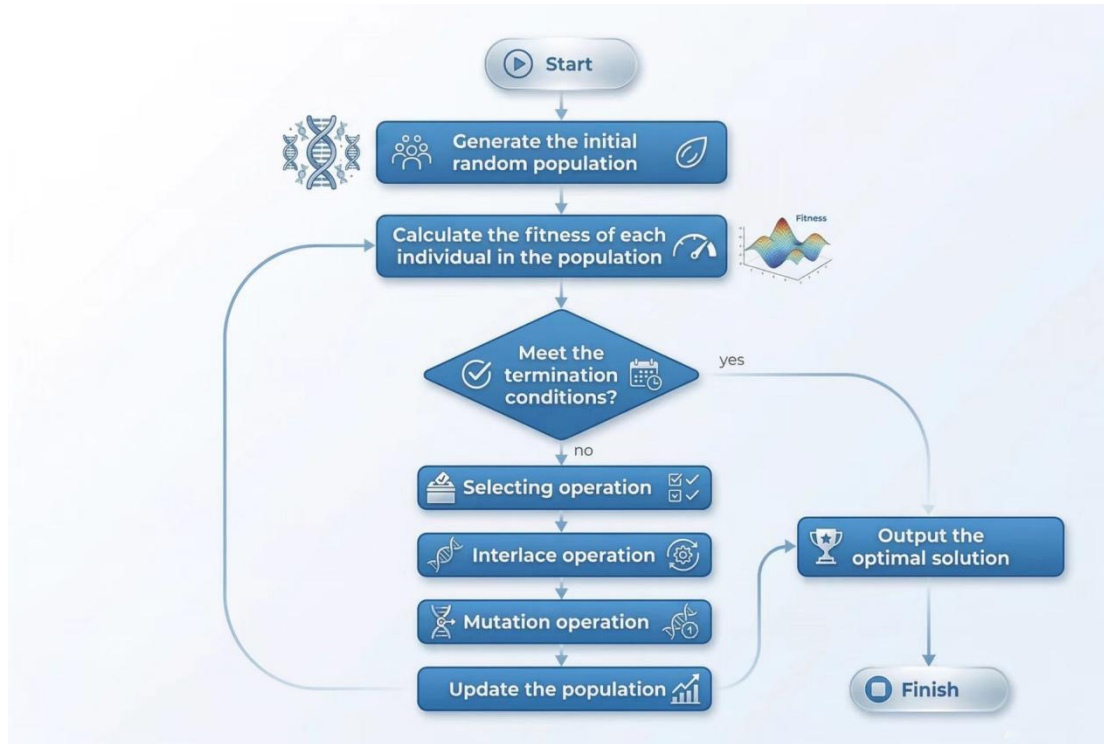


Figure 2. Flowchart of the genetic algorithm.

To address the shortcomings of the standard GA, many improved variants have been proposed. Li et al. incorporated domain-specific knowledge into robot path planning and introduced effective methods for evaluating path quality and accurately detecting collisions ^[32]. They proposed a novel knowledge-based GA that enables the algorithm to find optimal or near-optimal robot paths in both static and dynamic environments. Li et al. proposed an improved multi-objective GA ^[33]. The algorithm uses a heuristic median insertion method to generate the initial population, thereby improving the feasibility of initial paths. A multi-objective fitness function is constructed based on three criteria: path length, path safety, and path energy consumption, to ensure the quality of the planned path. Ab Wahab et al. proposed an enhanced GA that improves previous models by introducing a novel population initialization method and combining various genetic operators, resulting in higher-quality optimal paths ^[34]. In complex environments, when mobile robots perform tasks with different risk levels, various road factors need to be considered. Feng et al. proposed a novel hybrid adaptive genetic algorithm ^[35]. This algorithm integrates optimized dual optimization operators with an enhanced adaptive GA, achieving efficient path planning under diverse tasks and complex road conditions.

2.2.2. Ant colony optimization algorithm

The Ant Colony Optimization (ACO) algorithm is inspired by the foraging behavior of ants and is a probabilistic swarm intelligence optimization method. Ants deposit pheromones along their paths, and subsequent ants select paths based on pheromone concentration, forming a positive feedback mechanism that ultimately guides the colony toward the optimal solution ^[36]. Its principle is illustrated in **Figure 3**. In global path planning for mobile robots, the ACO algorithm is widely adopted due to its swarm intelligence characteristics. However, it still suffers from tendencies to fall into local optima and difficulty in obtaining globally optimal paths.

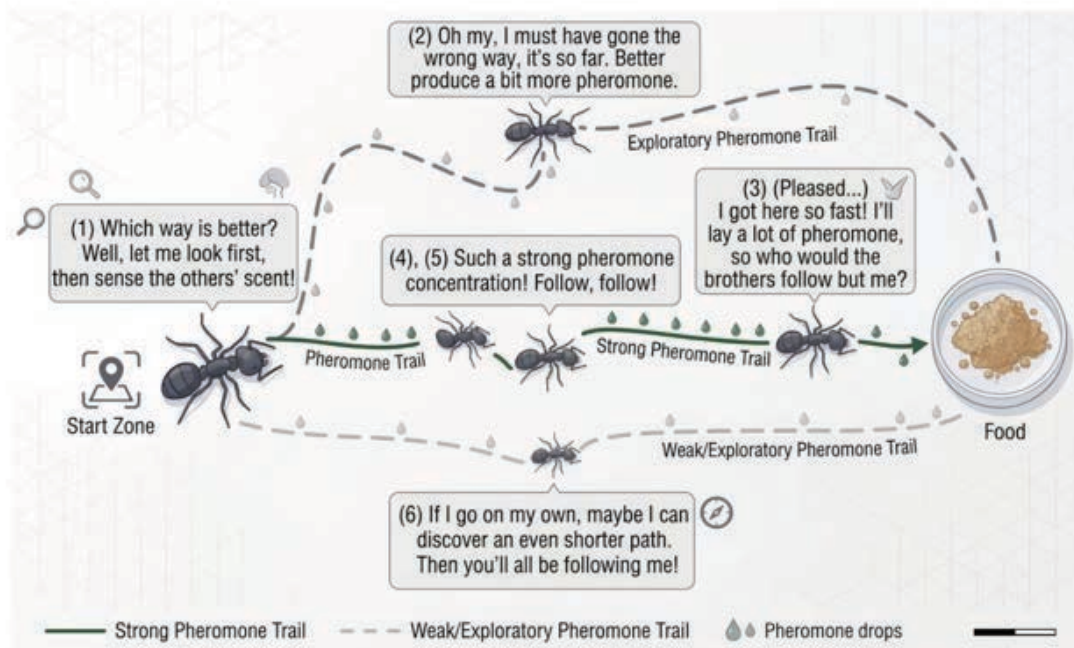


Figure 3. Schematic diagram of the principle of the ant colony optimization algorithm.

To overcome issues such as low search efficiency, slow convergence speed, and susceptibility to local optima, Zhang et al. proposed a novel multi-strategy enhanced Ant Colony Optimization algorithm for mobile robot path planning under non-holonomic constraints ^[37]. Zhao et al. introduced Q-Learning into the traditional ACO algorithm to enhance its learning capability in dynamic environments, proposing an improved ACO algorithm based on Q-Learning ^[38]. This algorithm effectively reduces the number of iterations, shortens path optimization time and path length, and improves other performance metrics, while enhancing global search capability and convergence speed. Li et al. proposed an intelligently enhanced ACO algorithm ^[39]. By implementing non-uniform initial pheromone distribution, adopting an ϵ -greedy strategy to adjust state transition probabilities, and dynamically balancing pheromone and heuristic functions, the algorithm outperforms traditional ACO in terms of path length, convergence speed, and solution stability, highlighting its practical value in solving global path planning problems for mobile robots. Shi et al. addressed problems of the basic ACO algorithm such as excessively long convergence time, poor global path quality, and unsuitability for dynamic and unknown environments ^[40]. They proposed an improved ACO algorithm by designing a new probability transfer function and a novel path exploration strategy for unknown

environments, effectively solving the weak path exploration capability of ACO in unknown environments.

2.2.3. Particle swarm optimization algorithm

The core idea of the Particle Swarm Optimization (PSO) algorithm originates from the social behavior of bird flocks foraging for food. Each potential solution is regarded as a “particle” in the search space, endowed with velocity and position attributes. By tracking individual and global optimal solutions, particles continuously adjust their states during iterations and eventually converge to the optimal solution of the problem. The algorithm is simple to implement, converges quickly, and is commonly used for optimization problems ^[41]. Its flowchart is shown in **Figure 4**.

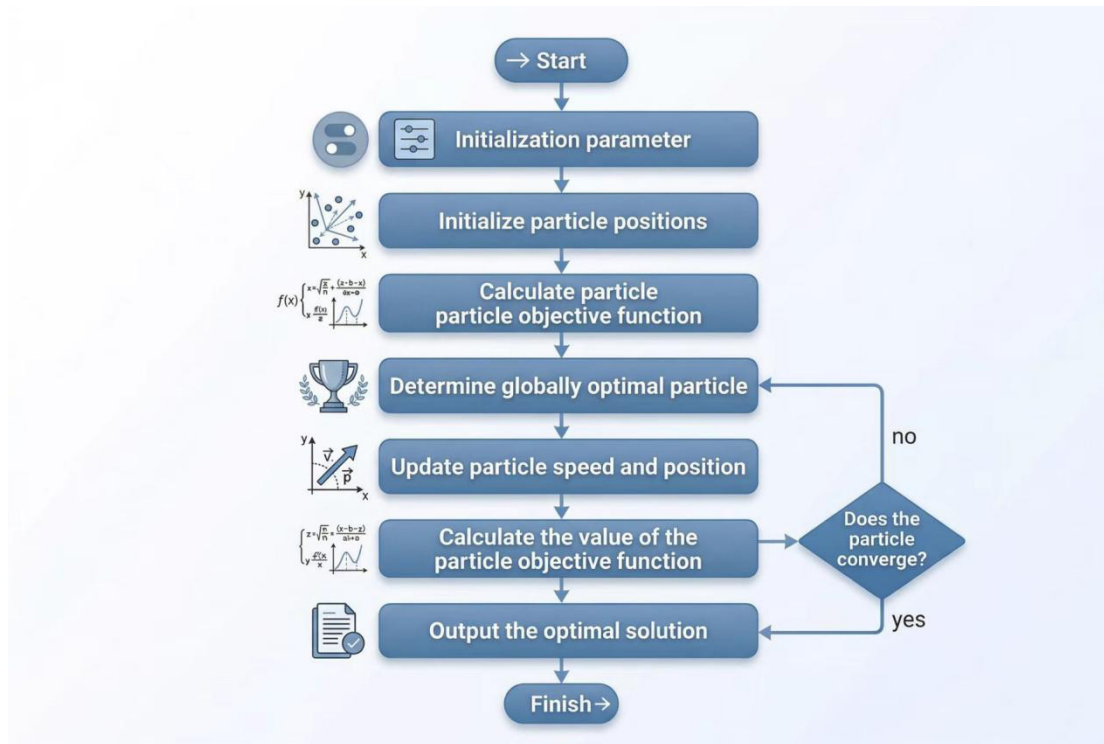


Figure 4. Flowchart of the particle swarm optimization algorithm.

For mobile robots operating under various constraints, finding a near-optimal path remains a challenging optimization problem. Traditional PSO algorithms have limitations when solving constrained problems, such as premature convergence. To address these challenges, Tao et al. proposed a dual-particle PSO algorithm with a random perturbation strategy ^[42]. This algorithm outperforms existing PSO algorithms and other mature path planning algorithms in terms of path quality and running time, demonstrating its feasibility in solving mobile robot path planning challenges. Lian et al. proposed a novel chaotic adaptive particle swarm optimization algorithm, which maintains the smoothness of robot path movement by optimizing control points in cubic spline interpolation ^[43]. Lu et al. proposed a multi-objective optimization algorithm for path planning based on improved particle swarm optimization. The algorithm constructs a multi-objective optimization mathematical model and uses PSO to solve multi-objective nonlinear optimization problems, thereby achieving global optimization of path planning ^[44]. Wang et al. proposed a novel particle swarm optimization method based on a hybrid learning model ^[45]. To prevent the algorithm from falling into local

optima, this method employs a hybrid learning model and a multi-pool fusion strategy, which improves convergence speed while increasing population diversity. Zeng et al. proposed a switching PSO algorithm based on dynamic neighborhoods ^[46]. The algorithm utilizes a new velocity update mechanism and a novel switching learning strategy to fully exploit the population evolution information of the entire swarm and conduct a comprehensive search of the problem space. In addition, it is hybridized with the differential evolution algorithm to alleviate premature convergence. Experimental results show that the algorithm outperforms several existing PSO algorithms in terms of solution accuracy and convergence performance, and is particularly suitable for complex multimodal optimization problems.

2.2.4. Artificial bee colony algorithm

The Artificial Bee Colony (ABC) algorithm is a swarm intelligence optimization algorithm that simulates the foraging behavior of honeybees. It achieves optimization through the division of labor and cooperation among employed bees, onlooker bees, and scout bees: employed bees exploit food sources and share information, onlooker bees select high-quality food sources based on the shared information, and scout bees randomly explore new areas ^[47]. The algorithm flowchart is shown in **Figure 5**. This positive feedback and exploration mechanism gives the algorithm strong global optimization capability, making it commonly used for solving numerical and combinatorial optimization problems ^[48]. However, the ABC algorithm is prone to falling into local optima, suffers from premature convergence, and exhibits poor adaptability when dealing with high-dimensional or discrete problems.

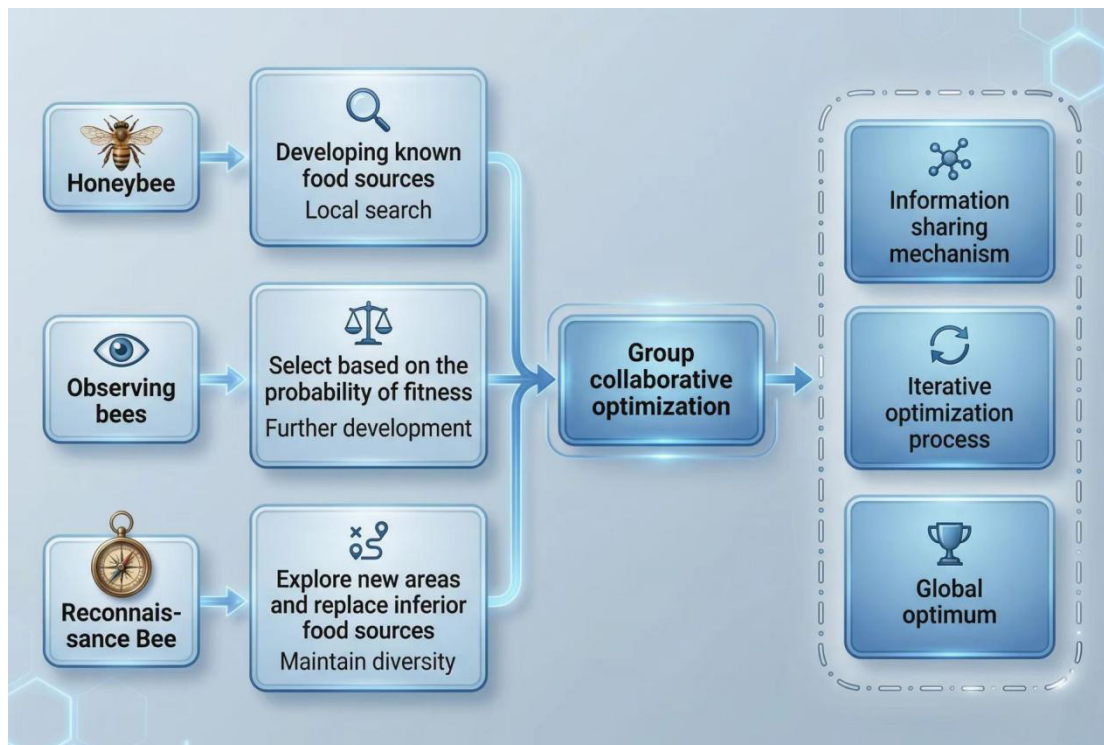


Figure 5. Flowchart of the artificial bee colony algorithm.

To overcome these limitations, Yildirim et al. proposed an improved ABC algorithm by integrating path linearization strategies with local search strategies ^[49]. This algorithm optimizes the traditional ABC,

improves convergence speed, and enhances the algorithm's ability to find global optimal solutions in complex solution spaces, effectively overcoming the tendency of traditional methods to fall into local optima. To solve the multi-objective path planning (PP) problem for mobile robots, Yu et al. proposed an improved ABC algorithm^[48]. To optimize the two objectives of path length and path safety, the algorithm designs a well-constructed environment model and path encoding method, and adopts variable neighborhood local search and global search strategies to enhance exploitation and exploration. Experimental results show that the algorithm has significant advantages in HV and SC metrics, with fast convergence speed and better robustness than Non-dominated Sorting Genetic Algorithm II and Multi-objective Particle Swarm Optimization. Ye et al. proposed a learning-based ABC algorithm for solving multi-objective path planning problems^[50]. The algorithm effectively addresses the challenges of multi-objective path planning by optimizing three key objectives: path length, safety, and smoothness. Cui et al. developed an improved multi-objective ABC algorithm based on a bionic intelligent optimization mechanism for handling multi-objective path planning problems^[47]. The algorithm considers multiple optimization objectives in path planning and improves the adaptability and practicality of the algorithm while ensuring solution quality.

2.3. Artificial intelligence algorithms

2.3.1. Neural networks

A neural network (NN) is a machine learning algorithm that simulates the working mechanism of neurons in the human brain (as shown in **Figure 6**). It serves as the cornerstone of modern deep learning. Its core idea is to construct a network composed of a large number of simple processing units to automatically learn patterns and regularities from data. With its powerful nonlinear fitting capability, it is widely applied in fields such as image recognition and natural language processing^[51].

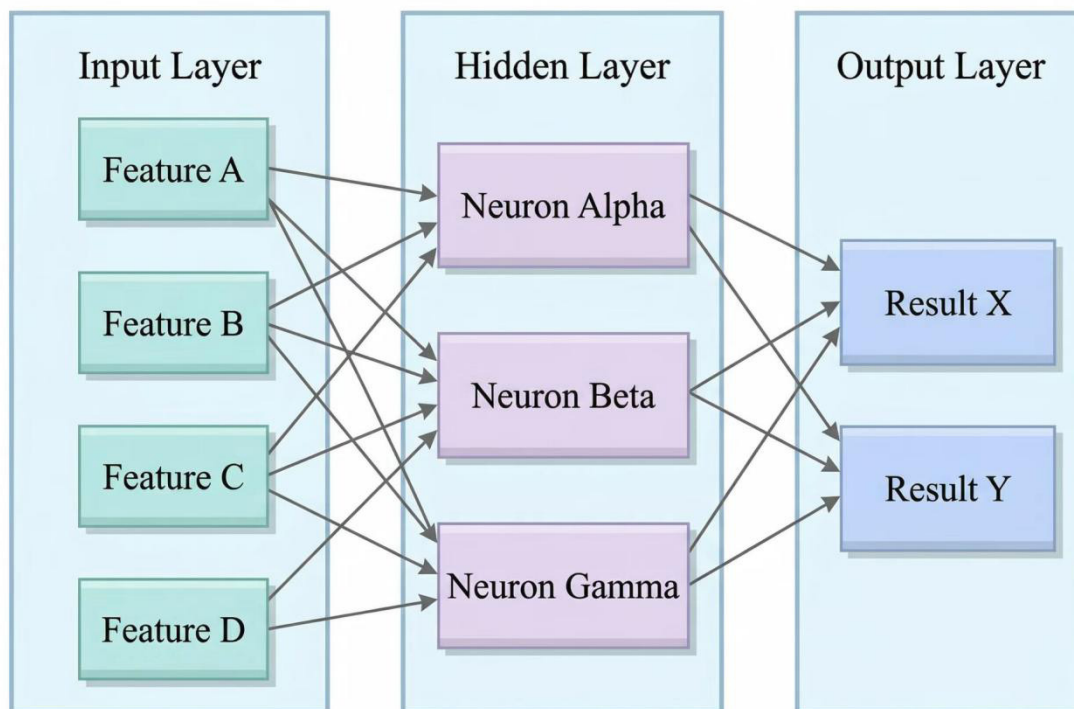
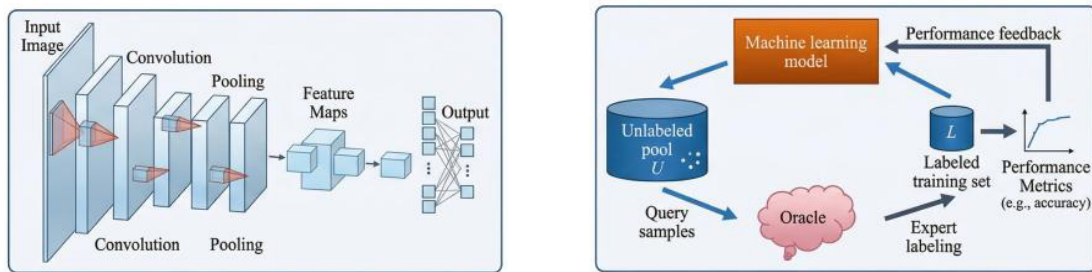


Figure 6. Schematic diagram of neural network structure.

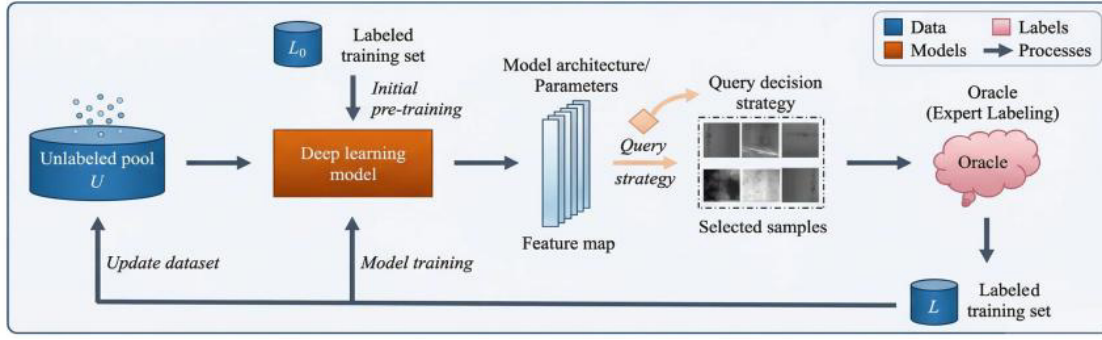
To address the challenges of intelligent path planning for mobile robots in complex road environments, Chen et al. integrated an improved bionic NN with an improved GA to develop a new environment representation and path decision-making model, proposing a hybrid symmetric bionic neural network algorithm^[52]. Mulás-Tejeda et al. proposed a Long Short-Term Memory (LSTM) NN model to realize and evaluate the dynamic obstacle avoidance capability of mobile robots. The model acquires information from the robot's LiDAR, enabling it to plan a path from the starting point to the target point while avoiding two dynamic obstacles and multiple static obstacles^[53]. As the scale and resolution requirements of the environment increase, the rising computational and time costs of bionic neural networks have attracted scholars' attention. Luo et al. first calculated paths using an improved NN algorithm, then introduced a multi-scale map method into the simulation neural network, and subsequently proposed an improved bionic neural network path planning method based on terrain scaling^[54]. This algorithm reduces mathematical complexity and significantly saves time. Molina-Leal et al. proposed an LSTM network with the Adam optimization method for learning trajectory planning strategies for mobile robots in dynamic environments^[55]. The algorithm was first validated in a computational simulation environment and then uploaded to GitHub for model training, enabling the robot to find a trajectory from a specified starting point to the target position while avoiding both dynamic and static obstacles. Romano et al. utilized a new extension of the tree-based AutoML software TPOT—TPOT-NN—to integrate and analyze extremely high-performance models generated by automated machine learning and artificial neural networks^[56]. They proposed an AutoML + NN estimator to evaluate machine learning results and hoped that AutoML + NN could be used as a free open-source tool in the context of TPOT.

2.3.2. Deep reinforcement learning

Deep Reinforcement Learning (DRL) is the integration of deep learning and reinforcement learning. It leverages the powerful perception capabilities of deep neural networks to process complex inputs and extract critical states. Through the decision-making framework of reinforcement learning, the agent learns the optimal policy through trial-and-error and reward signals during interaction with the environment^[57]. Its framework is illustrated in **Figure 7**.



(a) Schematic Diagram of Convolutional Neural Network Structure (b) “Active Learning Loop Based on Pooling”



(c) Classical Example of Deep Active Learning

Figure 7. Framework diagram of deep reinforcement learning.

To address the two deficiencies of slow convergence and excessive randomness in robot path planning, Yang et al. applied the Deep Q-Network (DQN) algorithm within the DRL framework and proposed an improved DQN algorithm^[58]. This algorithm combines Q-learning, an experience replay mechanism, and generative neural network techniques. Its convergence speed is faster than that of classical deep reinforcement learning algorithms and improves the efficiency of robot path planning. Li et al. addressed the issues of dimensional disaster and low search efficiency caused by traditional Q-learning in robot path planning by combining the feature extraction capability of deep convolutional neural networks with the decision-making capability of reinforcement learning, proposing an improved algorithm based on DRL^[59]. To tackle the overestimation problem, Qi et al. designed a deep double Q-learning strategy^[60]. This method uses a replay mechanism to separately calculate the robot's execution actions and decision evaluations, thereby improving algorithm convergence. Yuan et al. introduced an N-step prioritized double DQN algorithm, which integrates double Q-learning and N-step priority strategies to accelerate convergence and improve learning efficiency^[61]. Experimental results verified that the algorithm exhibits significant effectiveness and stability in path planning, notably reducing the number of cycles and improving convergence speed.

Currently, DRL is a popular research field, but the problem of dynamic path planning remains prominent. Focusing on DQN, a subfield of DRL, Gök et al. proposed a Dueling Double Deep Q-Network (D3QN) with a prioritized experience replay extension to further optimize DRL models^[62]. The algorithm employs both Double Deep Q-Networks (D2QN) and D3QN to train the model and conduct experiments on a test platform. The results show superior performance in path length and time to reach the goal compared to D2QN and D3QN.

As shown in **Table 1**, classical algorithms offer clear principles and relatively reliable planning but are often limited by computational cost, map dependence, or local optima. Intelligent optimization algorithms provide stronger global search capability, although their convergence and performance are sensitive to parameter settings. Artificial intelligence algorithms are more adaptable to dynamic environments but require substantial training data and computational resources. This comparison highlights the need for path-planning methods that balance convergence speed, search stability, and environmental adaptability.

Table 1. Comparative analysis of robot path planning algorithms

Category	Algorithm	Core Idea	Contribution	Optimality	Applicable Environment	Computational Efficiency	Main Features and Limitations
Classical algorithms	Dijkstra	Breadth-first search, gradually expanding the current shortest path nodes	Laid the theoretical foundation for shortest path search	Guaranteed optimal	Static	Relatively Low	Features: Optimal solution, simple principle. Limitations: High computational complexity, unsuitable for large-scale graphs
	A*	Introduces a heuristic function, prioritizes nodes with the smallest comprehensive cost	Pioneering work in heuristic search	Guaranteed optimal	Static	Relatively High	Features: High search efficiency, optimal solution. Limitations: Relies on accurate maps, high computational cost in high-dimensional spaces
	RRT	Rapidly constructs a search tree through random sampling to explore space	Solved the planning feasibility problem in high-dimensional spaces	Probabilistically complete	Dynamic	High	Features: Suitable for high-dimensional spaces, fast planning speed. Limitations: Non-optimal paths, instability due to randomness
	PRM	Offline sampling to build a roadmap, online query for feasible paths	Proposed the classical paradigm for multi-query path planning	Probabilistically complete	Static	High	Features: Suitable for multi-query tasks. Limitations: Insensitive to narrow passages; roadmap quality depends on sampling
	APF	Motion guided by the resultant force of attractive force from goal and repulsive forces from obstacles.	Proposed an intuitive physical model for reactive control	No	Dynamic	Very High	Features: Simple computation, good real-time performance. Limitations: Prone to local minima, oscillation in narrow passages
	DWA	Simulates and evaluates trajectories in dynamically feasible velocity space	Deeply integrates robot dynamic constraints with local obstacle avoidance	No	Dynamic	High	Features: Considers dynamic constraints, real-time obstacle avoidance. Limitations: Lacks global view, prone to local optima

Intelligent optimization algorithms	GA	Simulates natural selection, iteratively optimizes population through selection, crossover, and mutation	Classic representative of evolutionary computation	Approximate optimal	Static	Relatively Low	Features: Good generality, strong global search ability. Limitations: Complex parameter tuning, slow convergence
	ACO	Simulates positive feedback mechanism of pheromones, finds optimal path through swarm collaboration	Pioneered the ant colony optimization paradigm	Approximate optimal	Static	Relatively Low	Features: Fast convergence via positive feedback, strong robustness. Limitations: Blind initial search, prone to local optima
	PSO	Simulates social behavior; particles follow individual and global historical best positions.	Typical algorithm of swarm intelligence optimization	Approximate optimal	Static	Medium	Features: Few parameters, fast convergence. Limitations: Prone to premature convergence, weak local search ability
	ABC	Simulates division of labor in a bee colony; optimizes through the cooperation of employed, onlooker, and scout bees.	Proposed an optimization algorithm based on specific foraging behavior	Approximate optimal	Static	Medium	Features: Good global convergence. Limitations: Sensitive to parameter settings; local search ability needs improvement
	NN	Learns mapping from perceptual input to control output through multi-layer network structure	Initiated exploration of end-to-end learning in robot control	No	Dynamic	High	Features: Strong environmental adaptability, fast response. Limitations: “Black-box” model, relies on large amounts of training data
Artificial intelligence algorithms	DL	End-to-end learning of planning strategies from raw data using deep neural networks	Deep Q-Network demonstrates the ability to learn strategies from pixel input	No	Dynamic	High	Features: Strong perception and decision-making ability. Limitations: High demand for data and computing power, poor interpretability
	RL	Agent learns optimal policy to maximize cumulative reward through trial-and-error interaction with the environment.	Marks the paradigm shift from “programming” to “learning”	Approximate optimal	Dynamic	High	Features: Long-term planning capability, adapts to dynamic environments. Limitations: Long training cycles, high cost, difficult reward function design

3. Algorithm fusion mechanisms

3.1. Complementary fusion

Complementary fusion algorithms are computational methods designed to integrate information from multiple sources or modalities. Their core lies in leveraging the complementarity between different information sources to overcome the limitations of a single data source^[63]. Through fusion strategies at the data level, feature level, or decision level, information from different dimensions or modalities is synthesized to generate more comprehensive and robust outputs.

To address the shortcomings of traditional path planning algorithms, Liu et al. proposed a method that combines a kinematics-constrained A* algorithm with the DWA algorithm^[63]. This method improves the node expansion strategy and heuristic function of the A* algorithm by incorporating robot kinematics to plan the global path. Subsequently, it optimizes the trajectory evaluation function of the DWA algorithm, enabling DWA to plan local paths under the guidance of the global path.

Existing methods face issues such as insufficient obstacle avoidance and low planning efficiency. To tackle these challenges, Zhang et al. proposed a hybrid method combining an improved A* algorithm with an optimized DWA^[64]. A tuning factor is introduced to adjust the weight of the heuristic function to enhance the A* algorithm, while the improved DWA algorithm incorporates a path smoothing coefficient and a local target selection strategy, thereby improving the safety and stability of local planning.

3.2. Enhanced fusion

Enhanced fusion algorithms are intelligent processing methods that improve information accuracy and robustness by combining multi-source heterogeneous data. On the basis of traditional fusion, they introduce optimization mechanisms such as adaptive weighting, deep learning, or Kalman filtering to perform deep processing of multi-source data^[65]. In addition to integrating information, these algorithms focus on enhancing key features to generate enhanced fusion results that are more suitable for subsequent analysis and decision-making than the original data.

Traditional Artificial Potential Field (APF) is easy to implement but has very limited computational capability. In dense environments filled with obstacles, it frequently encounters problems such as being trapped in local minima and oscillation. In contrast, PSO's global optimization technique is effective in exploring solutions but is relatively slow in large-scale navigation problems. Benmachiche et al. developed a new path planning method by integrating the PSO and APF algorithms, combining the fast response capability of local planners with the extensive search capability of meta-heuristic techniques^[66]. This helps mobile robots navigate in areas with both static and dynamic obstacles. The method enables real-time optimization of APF parameters through PSO during navigation, allowing the robot to repeatedly reconfigure the force field according to environmental changes, thereby enhancing responsiveness, robustness, and the ability to escape local minima. Pan et al. proposed an algorithm that uses GA to train DL, combining the advantages of both DL and GA^[65]. GA is responsible for collecting states and paths from various scenarios to provide high-quality initial policies or training targets for the DL model, thereby rapidly generating optimized paths. The fusion of "optimization algorithm + learning model" demonstrates strong optimization capability and fast convergence speed.

4. Conclusions and prospects

Path planning for mobile robots, as a core research area in robotics, reflects the profound evolution of artificial intelligence and control science. Classical algorithms form the cornerstone of path planning. Based on rigorous mathematical modeling, they exhibit high computational efficiency and strong interpretability in known environments with static obstacles. These algorithms established the fundamental paradigm of path planning and provided a theoretical reference for subsequent research. The introduction of intelligent optimization algorithms marked the expansion of path planning from precise computation to heuristic search. By simulating self-organizing behaviors in nature, these algorithms demonstrate strong adaptability when dealing with dynamic environments and multi-constraint conditions. They transform path planning problems into multi-objective optimization problems, compensating for the limitations of classical algorithms in handling high-dimensional and nonlinear problems. The surge of artificial intelligence algorithms is triggering a fundamental paradigm shift in path planning. Through end-to-end learning mechanisms, robots can autonomously acquire obstacle avoidance strategies and navigation behaviors in unknown and dynamic scenarios. These algorithms break the absolute dependence on environmental models, enabling robots to possess human-like environmental perception and decision-making capabilities. They particularly demonstrate advantages that traditional algorithms struggle to match in complex semantic scenarios for path understanding.

Looking toward the future, innovation in fusion mechanisms will be a key breakthrough for improving the comprehensive performance of path planning. Specifically, the following directions are worth focused exploration:

(1) Deep fusion at the algorithm level

Constructing hybrid architectures of “knowledge-driven + data-driven” approaches. For example, using classical algorithms such as A* to provide prior knowledge or reward shaping for deep learning, thereby reducing the agent’s ineffective exploration space; or incorporating traditional planning algorithms as “differentiable modules” in neural networks to form an end-to-end joint optimization framework that balances global optimality and real-time responsiveness, achieving more efficient and intelligent path planning.

(2) Robust planning in dynamic environments

Combining predictive AI technologies with receding horizon optimization enables robots not only to perceive current obstacles but also to predict the motion trends of dynamic objects. At the same time, introducing federated learning mechanisms allows multiple robots to share obstacle avoidance experience while protecting data privacy, thereby enhancing the level of swarm intelligence.

(3) Fusion perception of semantics and geometry

Breaking through the limitations of traditional two-dimensional or three-dimensional spatial planning by incorporating semantic information into the cost function of path search. Utilizing the scene understanding capabilities of large language models to make path planning more consistent with human behavioral habits and social norms, achieving a transition from “feasible paths” to “reasonable paths”.

Acknowledgment

This paper is an output of the science project.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Wu J, Yan J, Zhang F, et al., 2026, A Review of Research on Path Planning Algorithms for Mobile Robots. *Application Research of Computers*, 1–13.
- [2] Dias P, Carvalho de Souza J, Solteiro P, et al., 2025, Robot Path Planning: From Analytical to Computer Intelligence Approaches. *Journal of Intelligent & Robotic Systems*, 111(4): 112.
- [3] Niloy M, Shama A, Chakraborty R, et al., 2021, Critical Design and Control Issues of Indoor Autonomous Mobile Robots: A Review. *IEEE Access*, 9: 35338–35370.
- [4] Liu L, Wang X, Yang X, et al., 2023, Path Planning Techniques for Mobile Robots: Review and Prospect. *Expert Systems with Applications*, 227: 120254.
- [5] Zhang L, Li Y, 2021, Mobile Robot Path Planning Algorithm Based on Improved A*. *Journal of Physics: Conference Series*, 1848(1): 012013.
- [6] Luo W, Zhang C, Zheng W, et al., 2025, Review and Outlook of Path Planning Algorithm System for Mobile Robots. *Computer Engineering and Applications*, 1–18.
- [7] Luo M, Hou X, Yang J, 2020, Surface Optimal Path Planning Using an Extended Dijkstra Algorithm. *IEEE Access*, 8: 147827–147838.
- [8] Li X, 2021, Path Planning of Intelligent Mobile Robot Based on Dijkstra Algorithm. *Journal of Physics: Conference Series*, 2083(4): 042034.
- [9] Gong H, Ni C, Wang P, et al., 2024, Smooth Path Planning Method Based on the Dijkstra Algorithm. *Journal of Beijing University of Aeronautics and Astronautics*, 50(2): 535–541.
- [10] Ahmad J, Nadhir Ab Wahab M, 2025, Enhancing the Safety and Smoothness of Path Planning Through an Integration of Dijkstra's Algorithm and Piecewise Cubic Bézier Optimization. *Expert Systems with Applications*, 289: 128315.
- [11] Zhao J, Zhang Y, Ma Z, et al., 2018, Improvement and Verification of A* Algorithm for AGV Path Planning. *Computer Engineering and Applications*, 54(21): 217–223.
- [12] Ou Y, Fan Y, Zhang X, et al., 2022, Improved A* Path Planning Method Based on the Grid Map. *Sensors*, 22(16): 6198.
- [13] Yin C, Tan C, Wang C, et al., 2024, An Improved A-Star Path Planning Algorithm Based on Mobile Robots in Medical Testing Laboratories. *Sensors*, 24(6): 1784.
- [14] Chen S, Yang G, Cui G, et al., 2025, Improved Path Planning and Controller Design Based on PRM. *IEEE Access*, 13: 44156–44168.
- [15] Liu C, Xie S, Sui X, et al., 2023, PRM-D* Method for Mobile Robot Path Planning. *Sensors*, 23(7): 3512.
- [16] Francis A, Faust A, Chiang H, et al., 2020, Long-Range Indoor Navigation With PRM-RL. *IEEE Transactions on Robotics*, 36(4): 1115–1134.
- [17] Ye L, Chen J, Zhou Y, 2022, Real-Time Path Planning for Robot Using OP-PRM in Complex Dynamic Environment. *Frontiers in Neurorobotics*, 16: 910859.
- [18] Wang K, Zeng G, Lu D, et al., 2019, Mobile Robot Path Planning Algorithm Based on Improved Asymptotically Optimal Bidirectional Rapidly-Exploring Random Tree. *Journal of Computer Applications*, 39(5): 1312–1317.
- [19] Khuat T, Bui D, Nguyen H, et al., 2025, Multi-Goal Rapidly Exploring Random Tree With Safety and Dynamic

- Constraints for UAV Cooperative Path Planning. *IEEE Transactions on Vehicular Technology*, 74(9): 13446–13457.
- [20] Yu Z, Xiang L, 2021, NPQ-RRT*: An Improved RRT* Approach to Hybrid Path Planning. *Complexity*, 2021(1): 6633878.
 - [21] Bian P, Fan J, Liu Y, et al., 2025, Quad Rapidly-Exploring Random Tree Star Algorithm with Improved Potential Force for Unmanned Aerial Vehicle Path Planning. *IEEE Access*, 13: 111187–111198.
 - [22] Bi Y, Fang X, 2025, A Hybrid Path Planning Framework Integrating Deep Reinforcement Learning and Variable-Direction Potential Fields. *Mathematics*, 13(14): 2312.
 - [23] Ma B, Ji Y, Fang L, 2025, A Multi-UAV Formation Obstacle Avoidance Method Combined with Improved Simulated Annealing and an Adaptive Artificial Potential Field. *Drones*, 9(6): 391–421.
 - [24] Yang J, Zhang H, Ning P, 2023, Path Planning and Trajectory Optimization Based on Improved APF and Multi-Target. *IEEE Access*, 11: 139121–139132.
 - [25] Zhai L, Liu C, Zhang X, et al., 2024, Local Trajectory Planning for Obstacle Avoidance of Unmanned Tracked Vehicles Based on Artificial Potential Field Method. *IEEE Access*, 12: 19665–19681.
 - [26] Jin F, Ye Z, Li M, et al., 2025, A New Hybrid Reinforcement Learning with Artificial Potential Field Method for UAV Target Search. *Sensors*, 25(9): 2796.
 - [27] Kobayashi M, Zushi H, Nakamura T, et al., 2023, Local Path Planning: Dynamic Window Approach With Q-Learning Considering Congestion Environments for Mobile Robot. *IEEE Access*, 11: 96733–96742.
 - [28] Chang X, Wang J, Li K, et al., 2025, Research on Multi-UAV Autonomous Obstacle Avoidance Algorithm Integrating Improved Dynamic Window Approach and ORCA. *Scientific Reports*, 15(1): 14646.
 - [29] Gong X, Gao Y, Wang F, et al., 2024, A Local Path Planning Algorithm for Robots Based on Improved DWA. *Electronics*, 13(15): 2965.
 - [30] Duan Q, 2026, Path Planning for Unmanned Surface Vehicle Based on D and DWA. *Control Engineering of China*, 33(1): 129–134.
 - [31] Tao Y, Wen Y, Gao H, et al., 2022, A Path-Planning Method for Wall Surface Inspection Robot Based on Improved Genetic Algorithm. *Electronics*, 11(8): 1192.
 - [32] Li J, Hu Y, Yang S, 2025, A Novel Knowledge-Based Genetic Algorithm for Robot Path Planning in Complex Environments. *IEEE Transactions on Evolutionary Computation*, 29(2): 375–389.
 - [33] Li K, Hu Q, Liu J, 2021, Path Planning of Mobile Robot Based on Improved Multiobjective Genetic Algorithm. *Wireless Communications and Mobile Computing*, 2021: 8836615.
 - [34] Ab Wahab M, Nazir A, Khalil A, et al., 2024, Improved Genetic Algorithm for Mobile Robot Path Planning in Static Environments. *Expert Systems with Applications*, 249: 123762.
 - [35] Feng T, Li J, Jiang H, et al., 2024, The Optimal Global Path Planning of Mobile Robot Based on Improved Hybrid Adaptive Genetic Algorithm in Different Tasks and Complex Road Environments. *IEEE Access*, 12: 18400–18415.
 - [36] Yang L, Fu L, Li P, et al., 2022, LF-ACO: An Effective Formation Path Planning for Multi-Mobile Robot. *Mathematical Biosciences and Engineering*, 19(1): 225–252.
 - [37] Zhang C, Ma J, Wang X, et al., 2025, NMS-EACO: A Novel Multi-Strategy ACO for Mobile Robot Path Planning. *Electronics*, 14(17): 3440.
 - [38] Zhao L, Li F, Sun D, et al., 2024, An Improved Ant Colony Algorithm Based on Q-Learning for Route Planning of Autonomous Vehicle. *International Journal of Computers Communications & Control*, 19(3): 5382.

- [39] Li P, Wei L, Wu D, 2025, An Intelligently Enhanced Ant Colony Optimization Algorithm for Global Path Planning of Mobile Robots in Engineering Applications. *Sensors*, 25(5): 1326.
- [40] Shi Y, Zhang H, Li Z, et al., 2023, Path Planning for Mobile Robots in Complex Environments Based on Improved Ant Colony Algorithm. *Mathematical Biosciences and Engineering*, 20(9): 15568–15602.
- [41] Huang C, Zhao Y, Zhang M, et al., 2023, APSO: An A*-PSO Hybrid Algorithm for Mobile Robot Path Planning. *IEEE Access*, 11: 43238–43256.
- [42] Tao B, Kim J, 2024, Mobile Robot Path Planning Based on Bi-Population Particle Swarm Optimization with Random Perturbation Strategy. *Journal of King Saud University – Computer and Information Sciences*, 36(2): 101974.
- [43] Lian J, Yu W, Xiao K, et al., 2020, Cubic Spline Interpolation-Based Robot Path Planning Using a Chaotic Adaptive Particle Swarm Optimization Algorithm. *Mathematical Problems in Engineering*, 2020: 1–20.
- [44] Lu C, Yang J, Leira B, et al., 2023, Three-Dimensional Path Planning of Deep-Sea Mining Vehicle Based on Improved Particle Swarm Optimization. *Journal of Marine Science and Engineering*, 11(9): 1797.
- [45] Wang Y, Wang B, Li Z, et al., 2023, A Novel Particle Swarm Optimization Based on Hybrid-Learning Model. *Mathematical Biosciences and Engineering*, 20(4): 7056–7087.
- [46] Zeng N, Wang Z, Liu W, et al., 2022, A Dynamic Neighborhood-Based Switching Particle Swarm Optimization Algorithm. *IEEE Transactions on Cybernetics*, 52(9): 9291–9301.
- [47] Cui Q, Liu P, Du H, et al., 2023, Improved Multi-Objective Artificial Bee Colony Algorithm-Based Path Planning for Mobile Robots. *Frontiers in Neurorobotics*, 17: 1196683.
- [48] Yu Z, Duan P, Meng L, et al., 2022, Multi-Objective Path Planning for Mobile Robot with an Improved Artificial Bee Colony Algorithm. *Mathematical Biosciences and Engineering*, 20(2): 2501–2529.
- [49] Yildirim M, Akay R, 2025, An Efficient Grid-Based Path Planning Approach Using Improved Artificial Bee Colony Algorithm. *Knowledge-Based Systems*, 318: 113528.
- [50] Ye F, Duan P, Meng L, et al., 2026, A Learning-Driven Artificial Bee Colony Algorithm for Mobile Robot Multi-Objective Path Planning. *Applied Soft Computing*, 192: 114692.
- [51] Mirza N, 2020, Robotic Path Planning and Fuzzy Neural Networks. *The International Arab Journal of Information Technology*, 17(4A): 615–620.
- [52] Chen S, Feng T, Li J, et al., 2025, Research on Intelligent Path Planning of Mobile Robot Based on Hybrid Symmetric Bio-Inspired Neural Network Algorithm in Complex Road Environments. *Symmetry*, 17(6): 836–863.
- [53] Mulás-Tejeda E, Gómez-Espinosa A, Escobedo C, et al., 2024, Implementation of a Long Short-Term Memory Neural Network-Based Algorithm for Dynamic Obstacle Avoidance. *Sensors*, 24(10): 3004.
- [54] Luo M, Hou X, Yang S, 2019, A Multi-Scale Map Method Based on Bioinspired Neural Network Algorithm for Robot Path Planning. *IEEE Access*, 7: 142682–142691.
- [55] Molina-Leal A, Gómez-Espinosa A, Escobedo C, et al., 2021, Trajectory Planning for a Mobile Robot in a Dynamic Environment Using an LSTM Neural Network. *Applied Sciences*, 11(22): 10689.
- [56] Romano J, Le T, Fu W, et al., 2021, TPOT-NN: Augmenting Tree-Based Automated Machine Learning with Neural Network Estimators. *Genetic Programming and Evolvable Machines*, 22(2): 207–227.
- [57] Yu J, Su Y, Liao Y, 2020, The Path Planning of Mobile Robot by Neural Networks and Hierarchical Reinforcement Learning. *Frontiers in Neurorobotics*, 14: 63.
- [58] Yang Y, Juntao L, Lingling P, 2020, Multi-Robot Path Planning Based on a Deep Reinforcement Learning DQN Algorithm. *CAAI Transactions on Intelligence Technology*, 5(3): 177–183.

- [59] Li H, Qi Y, 2020, A Robot Path Planning Method Based on Deep Reinforcement Learning in Complex Environments. *Application Research of Computers*, 37(S1): 129–131.
- [60] Qi R, Wu X, 2022, Robot Path Planning Based on Deep Reinforcement Learning. *Manufacturing Automation*, 44(12): 177–180.
- [61] Yuan G, Zhu B, Hu Y, et al., 2025, Assembly Measurement Path Planning for Mobile Robots Using an Improved Deep Reinforcement Learning. *Applied Sciences*, 15(23): 12406.
- [62] Gök M, 2024, Dynamic Path Planning via Dueling Double Deep Q-Network (D3QN) With Prioritized Experience Replay. *Applied Soft Computing*, 158: 111503.
- [63] Liu Y, Wang C, Wu H, et al., 2023, Mobile Robot Path Planning Based on Kinematically Constrained A-Star Algorithm and DWA Fusion Algorithm. *Mathematics*, 11(21): 4552.
- [64] Zhang J, Ling H, Tang Z, et al., 2025, Path Planning of USV in Confined Waters Based on Improved A* and DWA Fusion Algorithm. *Ocean Engineering*, 322: 120475.
- [65] Pan Y, Yang Y, Li W, 2021, A Deep Learning Trained by Genetic Algorithm to Improve the Efficiency of Path Planning for Data Collection With Multi-UAV. *IEEE Access*, 9: 7994–8005.
- [66] Benmachiche A, Derdour M, Kahil M, et al., 2025, Adaptive Hybrid PSO–APF Algorithm for Advanced Path Planning in Next-Generation Autonomous Robots. *Sensors*, 25(18): 5742.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Discussion on Challenges and Development of Electronic Warfare under Distributed Operations

Bing Luo*, Tianwei Zhou, Yuanming Zhou, Guofeng Shao

No.36 Research Institute, China Electronics Technology Group Corporation, Jiaxing, Zhejiang 314033, China

*Corresponding author: Bing Luo, 416283286@qq.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Against the backdrop of intensifying global strategic competition, distributed operations have evolved into a dominant new warfare paradigm. Based on the latest operational concepts proposed by the U.S. military, this paper elaborates on the origin, evolution trajectory and operational characteristics of distributed maritime and air operations, and analyzes the technical development paths and implementation approaches for future distributed combat systems. Furthermore, it investigates the operational efficacy and inherent dilemmas of electronic warfare (EW) when countering distributed operations, and proposes a roadmap for key EW technological advancements, providing references for the future development of electronic warfare.

Keywords: Distributed operations; Electronic warfare (EW); Anti-access/area denial (A2/AD)

Online publication: Sept 8, 2026

1. Introduction

Previous U.S. military campaigns, including the Gulf War, Kosovo War, and Syria War, mainly targeted adversaries with relatively limited military strength, relying on concentrated combat platforms to achieve overwhelming operational superiority. However, when confronting peer great-power competitors, robust Anti-Access/Area Denial (A2/AD) capabilities of rival forces will severely constrain the effectiveness of traditional centralized combat architectures. To address this vulnerability, the United States has put forward the distributed operations concept, heavily invested in unmanned platforms, artificial intelligence (AI), and other emerging technologies, and integrated them into practical combat deployments^[1]. The U.S. military has restructured its troop formation to build large-scale, low-cost unmanned forces, leveraging AI-assisted command-and-control (C2) decision-making to deliver enhanced combat performance at a lower overall cost. Notably, recent U.S.-Israeli joint operations such as Operation Epic Fury have validated highly intelligent distributed combat architectures. Dispersed reconnaissance and strike nodes are rapidly integrated into an

intelligent kill web via ubiquitous wide-area communication networks.

Distributed operations are overwhelmingly dependent on electronic information systems, while electronic warfare serves as the core means to paralyze and disrupt such information infrastructure. This paper analyzes how EW disrupts three core pillars of distributed combat: sensor intelligence for operational decision-making, communication networks for coordinated command, and unified time-synchronization and friend-or-foe identification data required for multi-platform coordination. On this basis, it explores the developmental trends of EW against distributed operations to offer engineering and theoretical references.

2. Analysis of distributed maritime and air operations concepts

To adapt to great-power competition and meet the operational demands of high-contention maritime and air combat environments, the U.S. military has designed forward-looking warfare frameworks and proposed distributed maritime operations and distributed air operations, laying theoretical foundations for future peer-level conflicts ^[2].

2.1. Distributed maritime operations

Distributed maritime operations disperse naval combat assets across extensive maritime zones to sustain persistent sea control capability. This concept originated from Jeff Kares’ “Distributed Networked Force” theory, with its evolutionary timeline summarized in **Table 1**.

Table 1. Evolution timeline of distributed maritime operations concept

Year	Key development milestones
2006	Retired U.S. Navy officer Jeff Kares published <i>Distributed Networked Operations: The Foundation of Network-Centric Warfare</i> , proposing the “Distributed Networked Force” composed of manned warships, unmanned surface vessels (USVs), and ship-borne unmanned aerial vehicles (UAVs).
2014	U.S. Naval War College formalized the “Distributed Lethality” concept, which mandates every surface combatant to deliver credible offensive threats to hostile forces ^[3,4] .
2020	Senior U.S. Navy leaders, including former Chief of Naval Operations Gary Roughead, identified unmanned surface vessels, unmanned underwater vehicles (UUVs), and drone carriers as top priorities for future naval force structure, designating distributed maritime operations as the primary naval combat paradigm.
2026	The U.S. Navy announced annualization of the Pacific Distributed Operations Experiment; distributed maritime operations have become the standard framework for naval routine training and force generation, with unmanned systems accounting for over 20% of total naval assets.

The core winning mechanism of distributed maritime operations lies in covert, dispersed, and flexible deployment of large numbers of surface combatants armed with heavy offensive ordnance ^[4]. This forces adversaries to allocate excessive detection and firepower resources. (**Figure 1**)

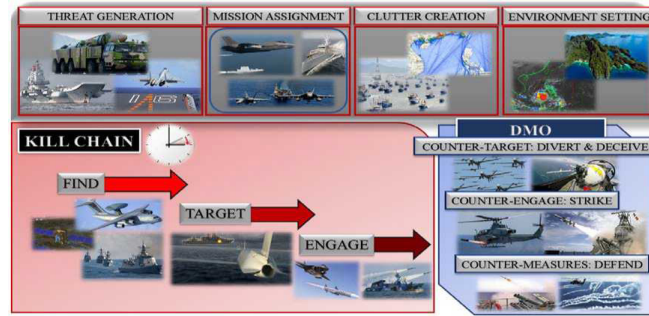


Figure 1. Concept diagram of distributed maritime operations.

(1) Core characteristics of distributed maritime operations

Multi-Domain Integration: Dynamic reorganization of combat platforms, weapons, and sensors fuses traditional fire-and-strike with intelligence warfare and electronic warfare across surface, subsurface, air, and cyber domains, maximizing integrated multi-domain combat performance.

(2) Deep informatization

Communication networking, automatic control, and intelligent signal processing form the technical foundation of distributed maritime combat. Guided by the distributed operations concept, the U.S. military has accelerated core information technology development, enabling manned/unmanned, surface/subsurface, and support platforms to support distributed information processing and networked cross-domain coordinated combat.

(3) System-centric integration

Individual combat platforms suffer catastrophic performance degradation once decoupled from the overall combat system. Inter-platform communication networks are the backbone of distributed architectures. The U.S. military maintains unique network integration advantages via high-speed, low-latency broadband satellite communication constellations represented by Starlink.

2.2. Distributed maritime operations

Distributed air operations were pioneered by the U.S. Defense Advanced Research Projects Agency (DARPA) to innovate aerial combat paradigms^[5,6]. The concept integrates manned fighters and autonomous UAVs, supported by AI-assisted command-and-control, to realize low-cost, high-lethality combat via flexible dynamic reconfiguration of aerial assets. Its evolutionary timeline is shown in **Table 2**.

Table 2. Evolution timeline of distributed maritime operations concept

Year	Key development milestones
2014	DARPA launched the System of Systems Integration Technology and Experimentation (SoSITE) program and formally proposed distributed air operations. The project developed innovative system architectures for heterogeneous aircraft, weapons, sensors and mission systems, distributing aerial combat capabilities across interoperable manned and unmanned aerial platforms.
2017	DARPA introduced the novel “Mosaic Warfare” operational framework.

2020 The U.S. military published *Mosaic Warfare: Decision-Centric Warfare Powered by Artificial Intelligence and Autonomous Systems*. Mosaic Warfare was framed as a concrete implementation pathway for the overarching decision-centric warfare framework, enabling commanders to generate faster, more effective decisions while paralyzing adversary decision cycles.

Post-2026 Next-Generation Air Dominance (NGAD) manned fighters (e.g., F-47) and Collaborative Combat Aircraft (CCA, e.g., YFQ-42A/YFQ-44A, XQ-58A) will achieve full operational deployment, ushering in an era of large-scale manned-unmanned swarm coordination via system-of-systems clusters.

After 2017, distributed air operations were gradually subsumed under the Mosaic Warfare framework, which shares identical core logic: dynamically assembling low-cost, single-function modular assets into scalable combat systems. The two concepts are inherently consistent^[7]. Future distributed air operations will decompose high-end aerial combat capabilities (detection, EW, strike) onto low-cost autonomous UAVs, with a small number of high-performance manned fighters serving as coordination hubs to field scalable, cost-effective modular combat clusters with rapid technology iteration cycles. (**Figure 2**)



Figure 2. Concept diagram of distributed air operations.

Core characteristics of distributed air operations:

(1) Scalable force generation

Functional decomposition is the core of distributed air combat. High-end fighters offload detection, EW, and strike functions to unmanned platforms, forming disruptive manned-unmanned swarm combat led by a small number of manned command aircraft.

(2) Broad environmental adaptability

UAVs can carry diversified ordnance. Supported by AI, they generate real-time multi-dimensional mission solutions and dynamically reconfigure payloads to adapt to high-intensity and low-intensity combat environments.

(3) Diversified offensive tactics

Mass deployment of heterogeneous aerial assets creates severe battlefield fog, obscuring real strike axes and feint decoys for adversaries. Diversified weapon payloads further complicate enemy threat identification and response planning.

2.3. Core implementation paths for distributed operations

Distributed combat capabilities will be realized through breakthroughs and large-scale deployment of four core technologies: manned-unmanned collaborative coordination, modular combat element dynamic reconfiguration, AI-assisted decision-making, and self-organizing anti-jamming communication links.

(1) Advancement of unmanned platforms

Since the 21st century, U.S. unmanned system R&D priorities have shifted toward systematic integration and autonomous intelligence. The U.S. Department of Defense has released five successive editions of the Unmanned Systems Integrated Roadmap to advance swarm networking and autonomous combat capabilities. The 2017–2042 Unmanned Systems Integrated Roadmap, published in August 2018, serves as the U.S. military’s overarching strategic guide for unmanned assets^[8]. Centered on four core themes—interoperability, autonomy, secure networking, and human-machine coordination—the roadmap.

(2) Advancement of artificial intelligence

AI functions as the primary enabler of distributed operations. Current machine learning algorithms process massive multi-source sensor datasets to accelerate commander situational awareness and decision-making. Future AI will transfer most cognitive overhead associated with closed kill-chain operations to autonomous intelligent systems, allowing human commanders to focus on high-level strategic judgment and rapid tactical response. DARPA’s Project MAVEN, an AI-enabled battlefield analytics program, has evolved into the core technical foundation of the U.S. Joint All-Domain Command and Control (JADC2) architecture by 2026.

(3) Advancement of self-organizing communication networks

Self-organizing networks constitute the technical backbone of distributed combat systems. Future distributed architectures will feature exponentially expanding network nodes and combat assets, driving surging demand for network bandwidth and transmission latency performance. Military IoT and advanced combat cloud networks will witness rapid development, transmitting not only raw sensor data but also shared machine-to-machine autonomous decision information to form a wireless combat cloud for real-time inter-node collaborative judgment^[9].

3. Analysis of EW challenges and developmental directions

3.1. Core operational value of electronic warfare

As analyzed above, unmanned surface, subsurface, and aerial vehicles represent the foundational nodes of future distributed combat systems, carrying integrated sensors, strike weapons, and EW payloads^[10]. Via self-organizing networks and AI, they form integrated manned-unmanned combat clusters to enhance battlefield lethality and survivability. Electronic warfare delivers decisive capability degradation by blinding sensors and severing communication links, emerging as the primary countermeasure against distributed combat architectures.

(1) Sensor neutralization tool

EW jams radar, electro-optical, infrared, and navigation sensors, depriving distributed combat clusters of real-time, accurate target metadata, including position, velocity, range, and target classification, completely breaking their situational awareness foundation.

(2) Communication link disruption tool

Inter-platform communication networks form the prerequisite for coordinated distributed combat. EW interference severs data exchange between dispersed nodes, dismantling system-level coordinated combat capacity.

(3) Command-and-control disruption tool

Accurate decision-making relies on credible sensor intelligence, low-latency communication links, and stable autonomous decision algorithms. EW distorts sensor data via deception jamming, delays or infiltrates communication channels, and fundamentally corrupts adversary command and tactical judgment cycles.

3.2. Inherent dilemmas confronting EW countermeasures

While EW can degrade distributed sensor and communication performance, advances in anti-jamming sensing, communication, and signal processing technologies have significantly mitigated traditional jamming effectiveness. Distributed combat architectures adopt multi-modal heterogeneous sensors and redundant communication backbones, diluting single-source EW suppression; in addition, distributed platforms may carry dedicated counter-EW hard-kill weapons targeting EW assets.

(1) Limitations of sensor jamming

Distributed platforms deploy heterogeneous multi-modal sensors with diversified waveform modulation and omnidirectional detection apertures, supported by high-precision multi-source data fusion algorithms. This creates critical EW challenges, including adaptive multi-modal jamming coverage, regional jammer deployment optimization, and real-time jamming effectiveness assessment. Advanced fusion filters can suppress external jamming signals, rendering conventional blanket jamming ineffective or marginally impactful.

(2) Limitations of communication network jamming

Distributed combat relies on self-organizing IP-based mesh networks interconnected via multi-layer routing protocols, complemented by directional datalink systems such as the Integrated Flight Data Link (IFDL) to enhance anti-jamming resilience. EW assets cannot simultaneously blanket all heterogeneous communication links, failing to fully sever inter-platform coordination and collapse distributed combat clusters.

3.3. Future development directions of electronic warfare

Analysis of EW operational utility and inherent limitations confirms that conventional standalone EW systems lack decisive, system-level suppression effects against distributed combat architectures, mandating comprehensive technical iteration and capability upgrades along four core pathways:

(1) Distributed EW system architecture technology

To counter heterogeneous multi-modal sensors and IP mesh communication networks of adversary distributed combat systems, EW must adopt a system-of-systems countermeasure paradigm to deliver equivalent distributed suppression capacity. A modular distributed EW framework paired with agile EW command-and-control algorithms enables multi-domain, multi-modal integrated jamming to dismantle enemy inter-platform coordination.

(2) Cooperative deception jamming technology

Deception jamming represents the dominant evolutionary trend for modern sensor countermeasures and the optimal pathway for communication network infiltration and disruption. Tailored to multi-platform distributed coordination architectures, deception jamming generates false target tracks and phantom contacts to exhaust adversary intercept ordnance and decision-making resources. Multi-source cooperative deception injects inconsistent radar echo data into separate enemy nodes, corrupting multi-sensor data fusion consistency.

(3) Integrated cyber-electronic attack technology

Conventional standalone EW is limited by line-of-sight range constraints and cannot engage remote distributed assets. Integrated cyber-electronic attack technology leverages wireless communication access and sensor terminal vulnerabilities to implant malware and gain remote control over large-scale unmanned swarms, realizing “single-point penetration, wide-area paralysis” and extending EW operational range to deliver decisive system-level disruption of distributed combat architectures.

(4) Cooperative EW-fire strike integration technology

A core objective of counter-distributed operations is disrupting adversary command-and-control and slowing their OODA (Observe-Orient-Decide-Act) decision cycles. Integrated EW-fire strike technology combines active/passive cooperative detection, multi-source heterogeneous data fusion, and communication node geolocation to rapidly identify high-value core command nodes within distributed combat clusters. EW suppression opens penetration corridors for kinetic strike assets, enabling hard-kill elimination of adversary system hubs to deliver fatal collapse of distributed combat superiority.

4. Conclusion

Driven by advances in electronic information technology and large-scale deployment of unmanned platforms, distributed operations will evolve into a mainstream warfare paradigm characterized by modular multi-weapon, multi-platform integrated combat systems, imposing severe response challenges for opposing forces. Electronic warfare serves as the decisive countermeasure against distributed combat architectures. Accelerated development of distributed EW systems, cooperative deception jamming, integrated cyber-electronic attack, and EW-kinetic joint strike technologies will effectively neutralize distributed combat advantages and drastically improve overall combat cost-effectiveness ratios.

Disclosure statement

The authors declare no conflict of interest.

References

- [1] U.S. Department of Defense, 2016, Remarks by Deputy Secretary Work on Third Offset Strategy. U.S. Department of Defense: 5–8.
- [2] Zhou X, Xu Q, He Z, 2020, Preliminary Research on Future Maritime and Air Distributed Combat Concepts and Force Employment. 2020(15): 856–860.
- [3] Clark B, Cropsey S, Walton T, 2020, Sustaining the Undersea Advantage: Disrupting Anti-Submarine Warfare Using Autonomous Systems. Hudson Institute.
- [4] Qin R, 2022, Research on Future Naval Distributed Combat Equipment Development. Journal of Naval Electronic Engineering, 42(07): 14–19.
- [5] Yuan C, Mu Z, 2020, Analysis of U.S. Distributed Air Combat Concepts. Aeronautical Weapons, 2020(27): 11–16.
- [6] Tang F, 2023, Analysis of U.S. Air Force Distributed Combat Progress Based on SoSITE Program. Avionics Technology, 54(04): 9–15.
- [7] DARPA, 2023, DARPA Budget, <https://www.darpa.mil/about-us/budget>

- [8] Li C, Sun Y, Fan Z, et al., 2023, Development Status and Trend of Unmanned Combat Platforms. *Journal of China Academy of Electronics and Information Technology*, 18(03): 274–279.
- [9] Liu G, Zhao T, Mao Y, 2025, Enlightenment from U.S. Communication Network Development for Distributed Air Combat. *Telecommunication Engineering*, 65(12): 1957–1964.
- [10] Yang K, Wang C, Wang B, 2025, Distributed Operations: Concepts, Development and Future Trends. *China Military to Civilian Conversion*, 2025(10): 23–24.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Integrated Services Platform of International Scientific Cooperation

Innoscience Research (Malaysia), which is global market oriented, was founded in 2016. Innoscience Research focuses on services based on scientific research. By cooperating with universities and scientific institutes all over the world, it performs medical researches to benefit human beings and promotes the interdisciplinary and international exchanges among researchers.

Innoscience Research covers biology, chemistry, physics and many other disciplines. It mainly focuses on the improvement of human health. It aims to promote the cooperation, exploration and exchange among researchers from different countries. By establishing platforms, Innoscience integrates the demands from different fields to realize the combination of clinical research and basic research and to accelerate and deepen the international scientific cooperation.

Cooperation Mode



Clinical Workers



In-service Doctors



Foreign Researchers



Hospital



University



Scientific institutions

OUR JOURNALS



The *Journal of Architectural Research and Development* is an international peer-reviewed and open access journal which is devoted to establish a bridge between theory and practice in the fields of architectural and design research, urban planning and built environment research.

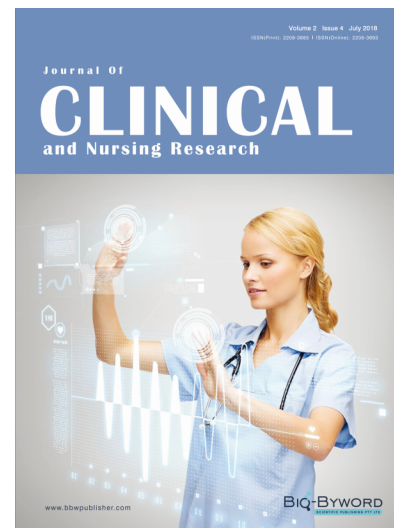
Topics covered but not limited to:

- Architectural design
- Architectural technology, including new technologies and energy saving technologies
- Architectural practice
- Urban planning
- Impacts of architecture on environment

Journal of Clinical and Nursing Research (JCNR) is an international, peer reviewed and open access journal that seeks to promote the development and exchange of knowledge which is directly relevant to all clinical and nursing research and practice. Articles which explore the meaning, prevention, treatment, outcome and impact of a high standard clinical and nursing practice and discipline are encouraged to be submitted as original article, review, case report, short communication and letters.

Topics covered by not limited to:

- Development of clinical and nursing research, evaluation, evidence-based practice and scientific enquiry
- Patients and family experiences of health care
- Clinical and nursing research to enhance patient safety and reduce harm to patients
- Ethics
- Clinical and Nursing history
- Medicine



Journal of Electronic Research and Application is an international, peer-reviewed and open access journal which publishes original articles, reviews, short communications, case studies and letters in the field of electronic research and application.

Topics covered but not limited to:

- Automation
- Circuit Analysis and Application
- Electric and Electronic Measurement Systems
- Electrical Engineering
- Electronic Materials
- Electronics and Communications Engineering
- Power Systems and Power Electronics
- Signal Processing
- Telecommunications Engineering
- Wireless and Mobile Communication

