

# **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

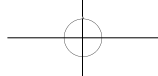
SYDNEY NSW 2000

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](http://www.bbwpublisher.com)

Email: [info@bbwpublisher.com](mailto:info@bbwpublisher.com)

## Table of Contents

- 1 An Intelligent Recognition Method for Radar Comb Spectrum Jamming Based on Dual-Channel Deep Convolutional Network**  
*Kuo Wang, Yunyu Wei, Sizhe Gao, Ziming Yin*
- 7 Real-Time Electricity Price Prediction and Trading Signal Generation Using Ensemble Tree-Based Machine Learning Models: A Comparative Study on the Spanish Electricity Market**  
*Yirui Liu*
- 14 Enhancing Tea Leaf Disease Classification with Cross-Attention Fusion and Magnitude-Aware Linear Attention**  
*Jiixin Zhu*
- 21 A Tibetan Speaker Verification Method Based on the Improved MFA-NConformer Model**  
*Yitong Gong, Yuting Chen*
- 28 IoT Security Situation Prediction Based on AGWO-Optimized BiGRU-ATTN**  
*Menghao Niu, Wen Chen*
- 35 A Zero-Dynamics Attack Detection Method for Offshore Wind Power Systems**  
*Kaige Chen, Hongran Li, Zeyu Zhang, Zhaoman Zhong, Lei Hu*
- 44 Physics Informed Hybrid Quantum-Classical Dispatching for LargeScale Renewable Power Systems: A Noise-Resilient Framework**  
*Fu Zhang, Yuming Zhao*
- 54 Research on Flow Field Calibration Method and Accuracy Improvement for the Aerodynamic Performance Test Rig of Aircraft Engine Compressors**  
*Cheng Lu, Honghui Xiang, Lei Huang, Kuan Liu, Xianghong Shen*
- 62 Sub-Pixel-Level Visual Inspection System for Dimensional Measurement of Ceramic Insulators Based on Halcon: Design and Implementation**  
*Yuehua Cao, Jiajie Han, Hanyang Zhu, Ge Yuan*

- 73 Discussion on Data Privacy Protection Technologies in Cloud Computing Environment**  
*Yixuan Dou*
- 86 Temporal-Spatial Evolution of Proton Beam Peak Energy and Its Correlation with Plasma Density**  
*Lu Yang*
- 94 Research on Fault Location and Isolation Method of Power Distribution System Based on Intelligent Sensing**  
*Yichi Zhang*
- 101 Research on User Behavior Analysis Based on Big Data Technology**  
*Yunzhe Dai*
- 108 Practical Application of Low-Cost Visual Inspection Systems in Industrial Robot Integration**  
*Chang Qi*
- 115 Research on RF OTA Test System Optimization During the New Product Introduction Phase of Consumer Electronics**  
*Peng Zhao*
- 122 Analysis of Fracture Failure of Connecting Bolts for Diaphragm Pump Valve Box Cover**  
*Xuan Qi, Yue Gao*
- 130 Experimental Investigation of the Velocity Distribution in the Tail Flame of an Inductively-Coupled-Plasma**  
*Xiaobao Mao*
- 141 Dynamics of Electric Field Perturbation in Gold Nanobipyramids During Dual-Pulse Two-Photon Coherent Excitation**  
*Qiong Li, Yao Li*
- 148 Comparison and Application Analysis of Three Wireless Charging Methods**  
*Shuqi Wang*
- 155 Research on Optimization of FPGA Streaming Processing System for High-Bandwidth Radar Echo Data**  
*Zhonghao Jiang*
- 163 Practices and Insights of Scientific Data Security Grading Management Based on the Entire Life Cycle**  
*Yu Zhai, Yong Song*

**169 Shadow Thermodynamics of an AdS Black Hole in Non-Commutative Geometry**

*Ying Zhu, Qing-Quan Jiang*

**184 Study on Performance Optimization of Radiation Protection Materials Based on Nanotechnology**

*Zhengyang Yuxiong*

**190 Integrated Volt-Energy Storage-Lighting System for Smart Road Lighting Based on Distributed MPPT: A Review of Hardware Design and Economic Analysis**

*Xiangran Chen, Jierui Feng, Yuying Pan, Mingwei Li, Yanran Li, Yuhan Song*



# An Intelligent Recognition Method for Radar Comb Spectrum Jamming Based on Dual-Channel Deep Convolutional Network

Kuo Wang, Yunyu Wei\*, Sizhe Gao, Ziming Yin

Xi'an Electronic Engineering Research Institute, China

\*Corresponding author: Yunyu Wei, 295423049@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:** This paper presents a deep learning method to recognize comb spectrum jamming in radar systems. Unlike traditional methods requiring manual feature extraction, our approach learns features directly from signal data. We built a dataset of radar echoes with four comb jamming types and five non-comb interference types. A dual-channel method creates 2D images preserving both magnitude and phase information from the signal spectrum. A CNN classifier with convolutional blocks, batch normalization, and dropout achieves 99.75% accuracy with 1.5% false alarm rate after only 7 training epochs.

**Keywords:** Comb-spectrum jamming; CNN; Radar interference identification

**Online publication:** March 26, 2026

## 1. Introduction

Electronic warfare is critical in modern battlefields. Radar systems face various jamming that degrades performance<sup>[1,2]</sup>. Traditional jamming identification methods rely on manual feature extraction and classification, but perform poorly for comb spectrum interference<sup>[3,4]</sup>. This paper presents a CNN-based method that automatically learns discriminative features for comb jamming recognition.

## 2. Radar signal and jamming models

### 2.1. Radar signal model

The radar transmits a linear frequency-modulated signal with carrier frequency 17GHz, bandwidth 80MHz, and pulse width 2  $\mu$ s:

$$x(t) = \text{rect}\left(\frac{t}{T_p}\right) \exp\left(j2\pi f_0 t + j\pi \frac{B}{T_p} t^2\right) \quad (1)$$

The received echo after down-conversion is:

$$x_r(t, t_m) = A \cdot \text{rect}\left(\frac{t - \tau(t_m)}{T_p}\right) \exp(j\pi\gamma(t - \tau(t_m))^2) \cdot \exp(-j2\pi f_0 \tau(t_m)) \quad (2)$$

## 2.2. Jamming models

### 2.2.1. Comb-spectrum jamming

Comb-spectrum jamming comes from adding together many single-frequency signals. Its frequency domain has peaks that show up at regular intervals. This jamming puts high power at specific frequencies and works well against frequency-agile and wideband radars. The jamming signal is:

$$J_{comb}(t) = \sum_{j=1}^N U_j \exp(j(2\pi f_j t + \phi_j)) \quad (3)$$

$N$  is the number of spectral lines.  $U_j$  is the amplitude of each line.  $f_j$  is its frequency.  $\phi_j$  is its starting phase picked from .

Four patterns are generated: symmetric equally-spaced, asymmetric, dense (2MHz spacing), and sparse (4 lines). Jammer-to-signal ratio (JSR) ranges from -15dB to 20dB:

$$JSR = 10 \log_{10} \left( \frac{P_j}{P_s} \right) \quad (4)$$

$P_j$  is the jamming power and  $P_s$  is the clean signal power.

JSR values are randomly picked between and when making the data.

### 2.2.2. Non-comb jamming models

Five types, including broadband Gaussian noise, colored noise, impulsive noise, swept frequency jamming, and strong target signal with noise.

## 2.3. Composite received signal model

The full received signal has both target echo and jamming:

$$r(t) = x_r(t) + J(t) + n(t) \quad (5)$$

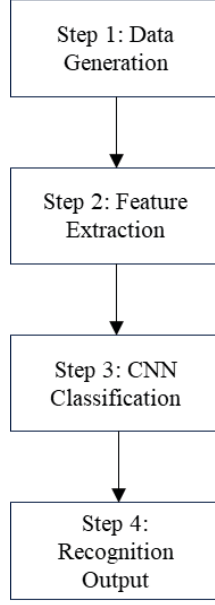
$x_r(t)$  is the target echo.  $J(t)$  is either comb-spectrum jamming or non-comb jamming.  $n(t)$  is additive white Gaussian noise with power set by the SNR.

This model is used to make training and testing data for the deep learning system.

## 3. Comb spectrum jamming recognition method based on CNN model

### 3.1. Overall recognition framework

**Figure 1** shows the overall framework. First, we make radar echo signals with different interference types. We use the models from Section II and control JSR and SNR. Secondly, we turn the complex signals into dual-channel feature images. These images keep both magnitude and phase information from the signal spectrum. Thirdly, we put these images into a CNN. The CNN uses many convolutional layers to get features at different levels. Finally, the network gives the classification result. It tells comb spectrum jamming apart from non-comb interference.



**Figure 1.** Overall framework of CNN-based comb spectrum jamming recognition.

### 3.2. Feature extraction and image construction

#### 3.2.1. Spectral feature extraction

FFT converts time-domain signal to frequency domain:

$$R[k] = \sum_{n=0}^{N-1} r[n] e^{-\frac{j2\pi kn}{N}} \quad (6)$$

Magnitude and phase spectra are extracted:

$$M[k] = |R_{shifted}[k]|, \Phi[k] = \angle R_{shifted}[k] \quad (7)$$

#### 3.2.2. Feature enhancement

Log transform enhances weak features:

$$M_{log}[k] = \log_{10} (M[k] + \epsilon) \quad (8)$$

Normalization maps features to [0,1]:

$$M_{norm}[k] = \frac{M_{log}[k] - \min(M_{log})}{\max(M_{log}) - \min(M_{log}) + \epsilon} \quad (9)$$

$$\Phi_{norm}[k] = \frac{\Phi[k] + \pi}{2\pi} \quad (10)$$

#### 3.2.3. Dual-channel image

Normalized magnitude and phase are reshaped to 64×64 and stacked:

$$I \in \mathbb{R}^{64 \times 64 \times 2}, I(:, :, 1) = M_{image}, I(:, :, 2) = \Phi_{image} \quad (11)$$



### 3.3. CNN architecture

The network comprises four convolutional blocks (32 to 256 filters), batch normalization, ReLU activation, max pooling, and dropout (0.1–0.4). Global average pooling reduces parameters:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j) \quad (12)$$

Two fully connected layers (128 and 64 neurons) with dropout 0.5/0.4 precede the 2-neuron softmax output:

$$p_c = \frac{\exp(z_c)}{\sum_{j=1}^2 \exp(z_j)} \quad (13)$$

### 3.4. Training strategy

Cross-entropy loss is minimized:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^2 y_{i,c} \log(p_{i,c}) \quad (14)$$

Adam optimizer ( $\eta = 10^{-3}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) with learning rate halved every 25 epochs. Regularization includes L2 weight decay ( $\lambda = 10^{-4}$ ), dropout, batch normalization, and early stopping.

## 4. Results and discussion

### 4.1. Experimental setup and dataset configuration

We ran many simulations to test our CNN-based comb spectrum jamming recognition method. The radar system settings were: carrier frequency 17 GHz, bandwidth 80 MHz, sampling rate 200 MHz, pulse width 2 microseconds, pulse repetition interval 200 microseconds, and 2048 sampling points per pulse.

The training dataset had two classes: comb spectrum jamming and non-comb interference. Each class had 2000 samples so 4000 samples total. The comb spectrum jamming included four patterns: symmetric equally-spaced, asymmetric, dense, and sparse. JSR values were randomly picked from -15 dB to 10 dB. The non-comb interference had five types: strong target signal with noise, broadband Gaussian noise jamming, colored noise jamming, impulsive noise jamming, and swept frequency jamming. We split the dataset randomly with 80% for training and 20% for validation.

### 4.2. Training process and recognition performance

We trained the CNN with the Adam optimizer. The starting learning rate was . Mini-batch size was 32. We set a maximum of 80 epochs. **Figure 2** shows the training and validation accuracy curves over time.



**Figure 2.** Training and validation accuracy curves.

The experimental results are shown in **Figure 2**. The training process converged very fast. Validation accuracy hit after only 7 epochs. Total training time was just 15 seconds on one GPU. The training and validation curves stayed close to each other the whole time. This means regularization worked well and there was no big overfitting problem.

#### 4.3. Recognition performance analysis

**Table 1** presents the confusion matrix for the proposed CNN method evaluated on the validation set.

**Table 1.** Confusion matrix

Actual/predicted	Comb	Non-Comb
Comb	400	0
Non-Comb	6	394

From the confusion matrix, the following performance metrics are derived as follows:

- (1) Overall accuracy: 99.25%
- (2) Comb class recognition rate:
- (3) Non-comb class recognition rate:
- (4) Precision:
- (5) Recall:
- (6) F1-Score: 99.26%

The results show the CNN method worked very well. The overall accuracy of means it could reliably tell

comb and non-comb interference apart. The false alarm rate was only . This is when non-comb was wrongly called comb. The miss detection rate was . This is when comb was wrongly called non-comb. These low error rates meet the tough needs of real radar anti-jamming systems.

## 5. Conclusion

This paper presented a deep learning method to recognize comb spectrum jamming in radar systems. We made a dual-channel feature image method that keeps both magnitude and phase information from the signal spectrum. These images go into a CNN we built. Test results showed the method worked very well. Overall accuracy was . The comb class recognition rate was . The false alarm rate was only . Training was fast and reached high accuracy after just 7 epochs.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Xu C, Yu L, Wei Y, 2019, Research on Active Jamming Recognition in Complex Electromagnetic Environment. 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Signal, Information and Data Processing (ICSIDP), 2019 IEEE International Conference On, 1–5.
- [2] Sun P, Yu J, Hao W, 2021, Research on Radar Active Jamming Recognition Based on 2-D Time-Frequency Features. 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Science and Technology Innovation (IAECST), 2021 3rd International Academic Exchange Conference On, 777–781.
- [3] Wang J, Dong W, Fu Q, et al., 2021, Radar Jamming Classification and Recognition Technology Based on Deep Learning. Proceedings of SPIE: The International Society for Optical Engineering, 11848(1): 118480T-1–118480T-7
- [4] Dong X, Guo S, Fang W, et al., 2024, Radar Active Composite Jamming Recognition Based on Characteristic Parameters, 415–420

### Publisher's note

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

# Real-Time Electricity Price Prediction and Trading Signal Generation Using Ensemble Tree-Based Machine Learning Models: A Comparative Study on the Spanish Electricity Market

Yirui Liu\*

College of Computer Science and Technology, Shandong University of Technology, Zibo 255049, Shandong, China

\*Corresponding author: Yirui Liu, 2110621759@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:** Accurate real-time electricity price forecasting is of critical importance for market participants seeking to optimize energy procurement, dispatch scheduling, and arbitrage strategies in liberalized electricity markets. However, existing forecasting approaches suffer from several key limitations: (1) conventional statistical models fail to capture the complex nonlinear interactions among generation mix, load demand, and temporal variables that collectively drive price dynamics; (2) single-model approaches lack robustness and are sensitive to overfitting, limiting their generalizability across diverse market conditions; (3) the interpretability of black-box prediction models remains insufficient, hindering the practical deployment of data-driven forecasting systems in operational decision-making. To address these challenges, this study proposes a comprehensive machine learning framework based on six tree-based ensemble models for hourly electricity price prediction in the Spanish electricity market. The proposed framework introduces three key contributions: (1) a systematic feature engineering pipeline incorporating lagged price variables, rolling statistics, and calendar-based temporal encodings; (2) a rigorous comparative evaluation of Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and LightGBM under identical experimental conditions; (3) a SHAP-based interpretability analysis that quantifies feature contributions and interaction effects at both global and local levels. Experimental results on the ENTSO-E Spanish market dataset demonstrate that XGBoost achieves the best overall predictive performance, with an  $R^2$  of 0.9660 and MAE of 1.5631 €/MWh.

**Keywords:** Electricity price forecasting; Ensemble learning; Gradient boosting; XGBoost; LightGBM; SHAP interpretability; Spanish electricity market

**Online publication:** March 26, 2026

# 1. Introduction

In recent years, electricity price forecasting (EPF) has emerged as a fundamental problem in power system management and energy market optimization <sup>[1]</sup>. Accurate real-time price prediction enables market participants to develop effective bidding strategies and maximize economic returns. With the rapid integration of renewable energy, price dynamics have become increasingly volatile and nonlinear, making reliable forecasting more challenging than ever.

Although significant progress has been made using statistical and machine learning approaches, existing methods still suffer from limited generalizability, insufficient model interpretability, and poor performance under high-volatility conditions <sup>[2–6]</sup>. Single-model frameworks, in particular, fail to capture the complex interactions among generation mix, load demand, and temporal variables <sup>[7,8]</sup>.

To address these challenges, this paper proposes a comprehensive tree-based ensemble learning framework for hourly electricity price forecasting on the Spanish ENTSO-E market dataset, further incorporating SHAP-based interpretability analysis to reveal key price-driving features <sup>[9–11]</sup>.

The main contributions are as follows:

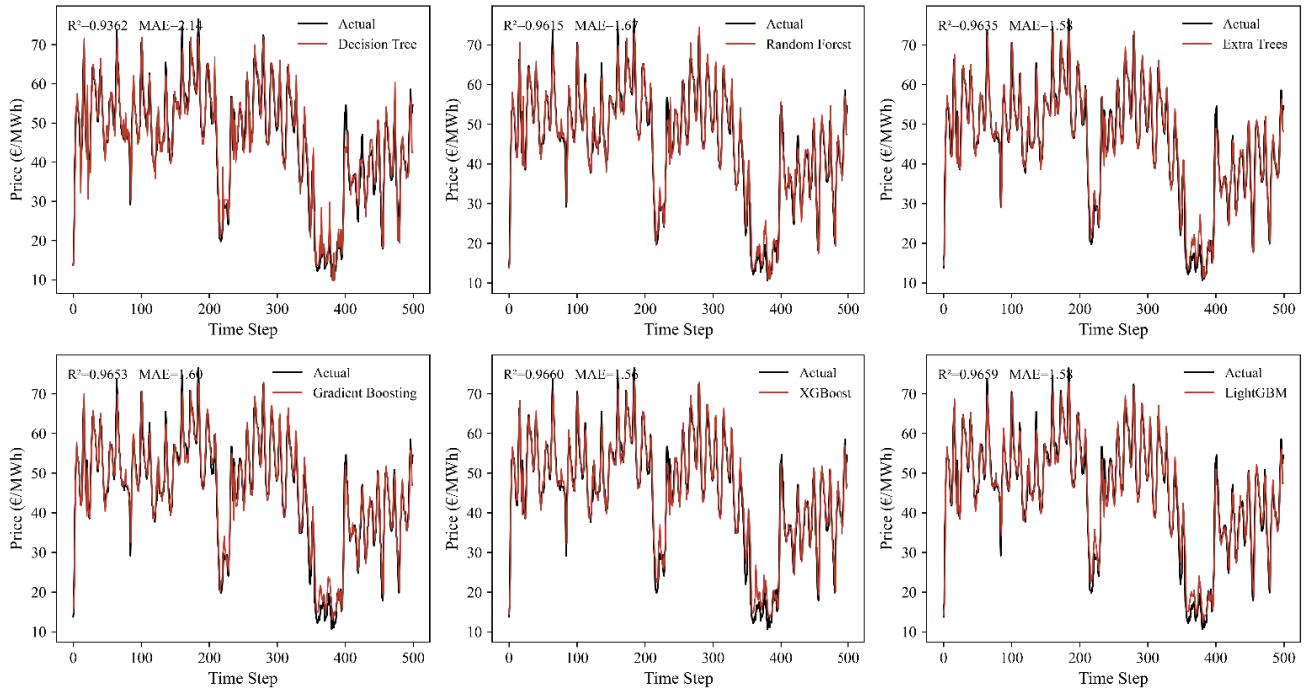
- (1) A systematic comparison of six tree-based models under identical experimental conditions;
- (2) A feature engineering pipeline integrating lagged price variables, rolling statistics, and temporal encodings;
- (3) A SHAP-based interpretability analysis quantifying global feature contributions and pairwise interaction effects.

## 2. Experimental results and analysis

This chapter presents a comprehensive evaluation of the six tree-based machine learning models on the Spanish hourly electricity price dataset. The experimental analysis is organized into three components: prediction curve comparison to assess temporal tracking fidelity, quantitative performance metric evaluation to rank model accuracy, and a visualization of trading signals generated by the best-performing model. All experiments are conducted under a unified protocol with an 80/20 chronological train-test split and evaluated using MAE, RMSE,  $R^2$ , and MAPE.

### 2.1. Prediction curve comparison

**Figure 1** illustrates the temporal alignment between the actual electricity prices and the predicted values generated by each of the six models over 500 consecutive test time steps. Across all subplots, the predicted curves closely track the black actual price curve, faithfully reproducing both the short-term oscillations and the broader downward trend observed between time steps 300 and 420. The Decision Tree exhibits the largest deviations, particularly at price spikes, consistent with its lower  $R^2$  of 0.9362. In contrast, the four ensemble models, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and LightGBM, achieve substantially tighter fits, with their predicted curves nearly overlapping the actual series throughout the evaluation window.



**Figure 1.** Comparison of actual and predicted electricity price curves for six tree-based machine learning models over 500 consecutive test time steps.

Each subplot displays the black actual price series against the red predicted series, with the corresponding  $R^2$  and MAE values annotated in the upper left corner. The results demonstrate that ensemble models, particularly XGBoost and LightGBM, achieve substantially closer alignment with the actual price trajectory compared to the single Decision Tree baseline, effectively capturing both short-term volatility and longer-term price trends in the Spanish electricity market.

## 2.2. Quantitative performance comparison

**Table 1** reports the MAE, RMSE,  $R^2$ , and MAPE values achieved by each of the six models on the held-out test set. A clear and consistent performance hierarchy emerges across all four metrics: the single Decision Tree yields the weakest results ( $R^2 = 0.9362$ , MAE = 2.1368 €/MWh), while all five ensemble models substantially surpass it. Among the ensemble methods, XGBoost achieves the best overall performance with  $R^2 = 0.9660$  and MAE = 1.5631 €/MWh, marginally outperforming LightGBM ( $R^2 = 0.9659$ ) and Gradient Boosting ( $R^2 = 0.9653$ ). Extra Trees and Random Forest deliver competitive results, confirming that Bagging-based methods also provide strong predictive accuracy. The narrow performance gap among the four boosting and bagging ensembles suggests that the primary determinant of accuracy in this task is the adoption of an ensemble strategy rather than the specific boosting algorithm employed.

**Table 1** was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination ( $R^2$ ), and Mean Absolute Percentage Error (MAPE). Lower MAE, RMSE, and MAPE values indicate higher prediction accuracy, while a higher  $R^2$  value reflects stronger explanatory power. XGBoost achieves the best performance across all metrics.

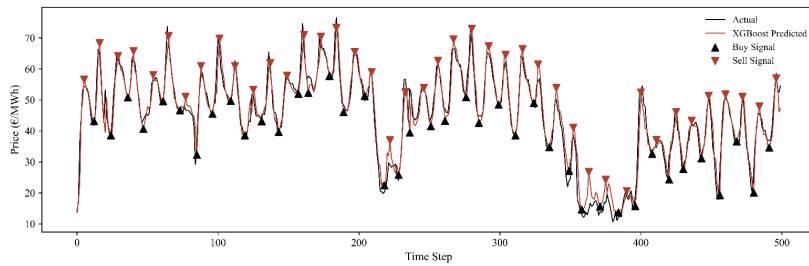


**Table 1.** Quantitative performance comparison of six tree-based machine learning models on the Spanish electricity price test set

Model	MAE	RMSE	R <sup>2</sup>	MAPE (%)
Decision Tree	2.1368	2.9128	0.9362	3.5430
Random Forest	1.6721	2.2632	0.9615	2.7878
Extra Trees	1.5810	2.2036	0.9635	2.6662
Gradient Boosting	1.5974	2.1489	0.9653	2.6913
XGBoost	1.5631	2.1249	0.9660	2.6579
LightGBM	1.5769	2.1289	0.9659	2.6582

### 2.3. Trading signal identification

**Figure 2** presents the buy and sell signals generated by the best-performing XGBoost model over the 500-step test window, overlaid on both the actual and predicted price curves. Local price minima identified in the predicted series are marked as buy signals (upward-pointing triangles), while local maxima are designated as sell signals (downward-pointing triangles). The predicted curve closely shadows the actual price trajectory throughout the evaluation period, resulting in trading signals that are well-aligned with the true price turning points. The buy signals predominantly coincide with genuine price troughs, and the sell signals consistently appear at or near true price peaks, confirming the practical utility of the XGBoost predictions for intraday electricity arbitrage strategies. The high density of accurately positioned signals across the full test window demonstrates that the model captures both the fine-grained oscillatory structure and the broader price cycles of the Spanish electricity market with sufficient fidelity to support real-time trading decision support.



**Figure 2.** Trading Signal Identification.

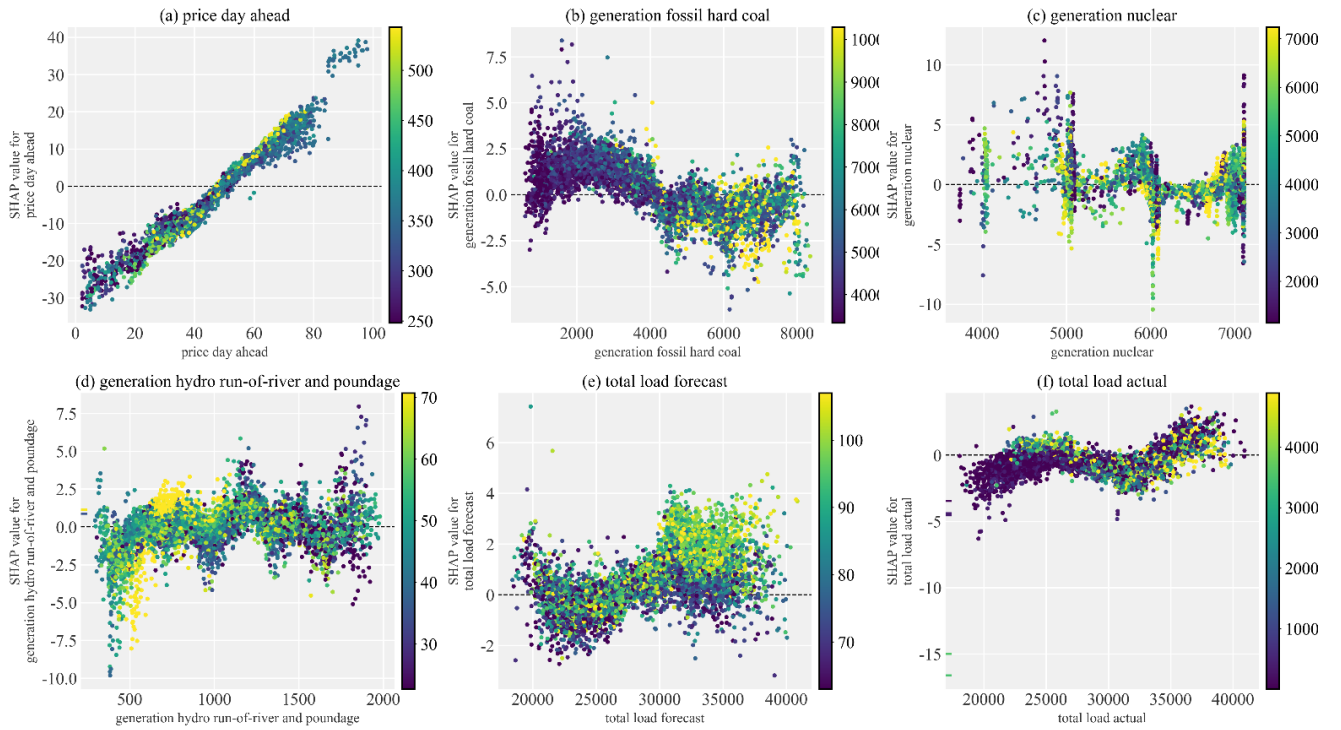
## 3. SHAP-based interpretability analysis

Building upon the predictive performance, this section employs SHapley Additive exPlanations (SHAP) to systematically investigate the feature importance and interaction effects underlying the XGBoost model. SHAP decomposes each prediction into additive contributions from individual features, providing a theoretically grounded and model-agnostic framework for interpreting the learned decision logic.

### 3.1. SHAP dependence plots

**Figure 3** presents SHAP dependence plots for the six most influential features identified in the XGBoost model. Each subplot displays the relationship between a feature's raw value on the horizontal axis and its corresponding SHAP contribution on the vertical axis, with point color encoding a secondary interaction feature. Panel (a)

reveals a near-linear positive relationship between price day ahead and its SHAP value, confirming that the day-ahead price is by far the dominant predictor of actual electricity prices, contributing SHAP values ranging from approximately  $-30$  to  $+40$  €/MWh. Panel (b) shows that generation from fossil hard coal exerts a moderately negative influence at intermediate generation levels, reflecting the merit-order effect in which coal plants occupy a mid-stack position. Panel (c) indicates that nuclear generation contributes small but consistent negative SHAP values across its observed range, consistent with its role as a low-cost baseload source that suppresses spot prices. Panels (d) through (f) demonstrate that hydro run-of-river output, total load forecast, and total load actual all exhibit positive SHAP contributions at higher values, reflecting demand-driven price escalation under high-load conditions.



**Figure 3.** SHAP dependence plots for the six most influential input features of the XGBoost electricity price prediction model, including price day ahead, generation fossil hard coal, generation nuclear, generation hydro run-of-river and poundage, total load forecast, and total load actual.

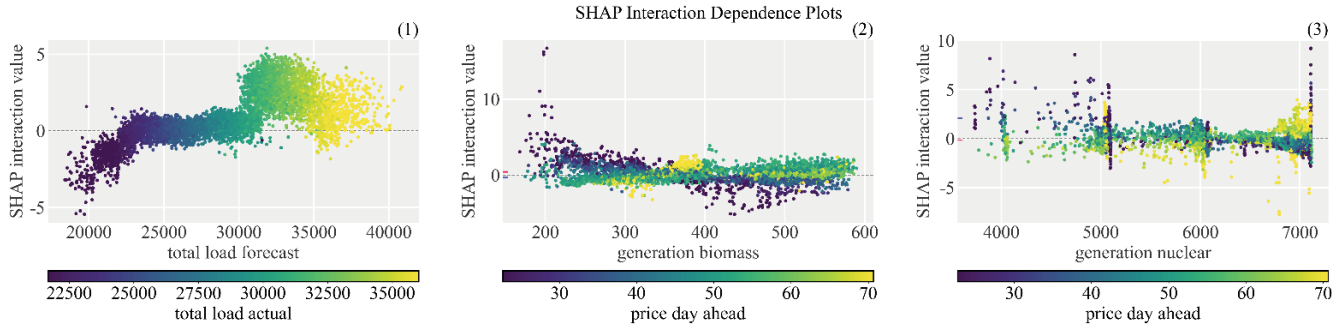
Each point represents one test sample, with the horizontal axis showing the feature value and the vertical axis showing the corresponding SHAP contribution to the predicted price. Point colors encode a secondary interacting feature value. The plots reveal the direction, magnitude, and nonlinearity of each feature’s marginal effect on electricity price predictions.

### 3.2. SHAP interaction dependence plots

**Figure 4** presents SHAP interaction dependence plots for three key feature pairs, quantifying the conditional interaction effects that cannot be captured by marginal SHAP values alone. Panel (1) reveals a pronounced nonlinear interaction between total load forecast and total load actual: at low forecast levels the interaction value is negative, transitioning to strongly positive beyond approximately 30,000 MW, indicating that the price-elevating



effect of high load is amplified when both forecasted and realized demand are simultaneously elevated. Panel (2) shows that the interaction between generation biomass and price day ahead is relatively weak and dispersed across the mid-range of biomass output. Panel (3) demonstrates that generation nuclear interacts with price day ahead in a concentrated manner near the 4,500–5,000 MW nuclear output band, where interaction values cluster tightly around zero, suggesting that nuclear generation’s price-suppressing effect is partially moderated by the prevailing day-ahead price level.



**Figure 4.** SHAP interaction dependence plots for three selected feature pairs: (1) total load forecast interacting with total load actual, (2) generation biomass interacting with price day ahead, and (3) generation nuclear interacting with price day ahead. Vertical axes show SHAP interaction values and point colors encode the secondary feature.

## 4. Conclusion

This study proposed a comprehensive tree-based machine learning framework for real-time electricity price forecasting and trading signal generation in the Spanish electricity market. Six models, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and LightGBM, were systematically constructed, trained, and evaluated under identical experimental conditions. Experimental results demonstrated that all ensemble models substantially outperformed the single Decision Tree baseline, with XGBoost achieving the best overall performance across all four evaluation metrics ( $R^2 = 0.9660$ ,  $MAE = 1.5631$  €/MWh,  $RMSE = 2.1249$ ,  $MAPE = 2.6579\%$ ). SHAP-based interpretability analysis further revealed that price day ahead is the dominant price driver, while generation mix and load variables contribute nonlinear secondary effects that interact in economically meaningful ways. The proposed framework also demonstrated practical utility in generating actionable buy and sell signals aligned with actual price turning points. Future work will explore the integration of weather features and cross-market data to further improve forecasting accuracy, the incorporation of deep learning architectures such as Transformers for long-horizon price prediction, and the development of a real-time deployment pipeline for operational electricity trading support systems.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Jędrzejewski A, Lago J, Marcjasz G, et al., 2022, Electricity Price Forecasting: The Dawn of Machine Learning.

IEEE Power and Energy Magazine, 20(3): 24–31.

- [2] Wang P, Xu K, Ding Z, et al., 2022, An Online Electricity Market Price Forecasting Method Via Random Forest. IEEE Transactions on Industry Applications, 58(6): 7013–7021.
- [3] Shah R, Shah H, Bhim S, et al., 2021, “Short-Term Electricity Price Forecasting using Ensemble Machine Learning Technique. 2021 1st International Conference in Information and Computing Research (iCORE), Manila, Philippines, 145–150.
- [4] Sahoo S, Swain S, Dash R, 2023, An Analysis of Machine Learning Methods for Electricity Price Forecasting. 2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS), Bhubaneswar, India, 1–5.
- [5] Alkawaz A, Abdellatif A, Kanesan J, et al., 2022, Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model. IEEE Access, 2022(10): 108021–108033.
- [6] Mubarak H, Ahmad S, Hossain A, et al., 2023, Short-Term Electricity Price Forecasting Using Interpretable Hybrid Machine Learning Models. 2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT), Male, Maldives, 1–6.
- [7] Arya K, Vijaya Chandrakala K, 2021, Machine Learning Based Prediction and Forecasting of Electricity Price During COVID-19, 2021 IEEE International Power and Renewable Energy Conference (IPRECON), Kollam, India, 1–6.
- [8] Han L, Ban C, Zhang C, et al., 2024, Electricity Price Forecasting in Power Markets Based on Machine Learning, 2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI), Wuhan, China, 91–96.
- [9] Heijden T, Lago J, Palensky P, et al., 2021, Electricity Price Forecasting in European Day Ahead Markets: A Greedy Consideration of Market Integration. IEEE Access, 2021(9): 119954–119966.
- [10] Yorat E, Zor K, Özbek N, 2023, Day-Ahead Electricity Price Forecasting Using Artificial Intelligence-Based Algorithms, 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 121–126.
- [11] Yildirim B, Yildiz S, Turkoglu A, et al., 2023, Evaluating LMP Forecasting with LSTM Networks: A Deep Learning Approach to Analyzing Electricity Prices During Unpredictable Events, 2023 5th Global Power, Energy and Communication Conference (GPECOM), Nevsehir, Turkiye, 477–482.

**Publisher’s note**

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

# Enhancing Tea Leaf Disease Classification with Cross-Attention Fusion and Magnitude-Aware Linear Attention

Jiaxin Zhu\*

Department of Electronic and Information Engineering, Liaoning Technical University, Huludao 125100, China

\*Corresponding author: Jiaxin Zhu, 2306110233@stu.lntu.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:** Accurate tea leaf disease classification in real-world scenarios is hindered by complex backgrounds and the loss of fine-grained lesion details during CNN down sampling. To address this, we propose ResNet50-Dual-Fusion. It integrates a Cross-Attention Feature Fusion module (CAmodule) to adaptively reconstruct tiny lesion edges via cross-spatial interaction between shallow and deep features. Furthermore, a Magnitude-Aware Linear Attention (MALA) module with 2D Rotary Position Embedding (RoPE) is introduced to rectify magnitude neglect, effectively suppressing background noise. Evaluated on a 5,276-image dataset, our model achieves 85.96% accuracy (+3.00% over the baseline), outperforming architectures like ViT and Swin-Tiny. Grad-CAM visualizations confirm its superior lesion localization, providing a robust paradigm for automated crop disease diagnosis.

**Keywords:** Tea leaf disease classification; Deep learning; Residual network (ResNet-50); Cross-attention mechanism; Magnitude-aware linear attention (MALA)

**Online publication:** April 10, 2026

## 1. Introduction

Automated diagnosis is crucial for protecting vulnerable tea yields<sup>[1]</sup>. While CNNs like ResNet and initial visual models pioneered disease classification, subsequent architectures utilizing RetinaNet and deeper networks further enhanced detection robustness<sup>[2-5]</sup>. However, complex agricultural backgrounds cause misclassifications, and continuous down sampling obscures fine-grained lesion details<sup>[6]</sup>. To resolve this, we propose ResNet50-Dual-Fusion. Inspired by cross-domain matching, our Cross-Attention Feature Fusion (CAmodule) adaptively reconstructs tiny lesions<sup>[7]</sup>. Additionally, we integrate a MALA module to rectify magnitude neglect, dynamically suppressing backgrounds to achieve superior classification accuracy<sup>[8]</sup>.

## 2. Related work

### 2.1. Plant and tea leaf disease classification

Deep learning drives automated crop disease recognition. Initial DNNs proved effective for plant diseases<sup>[1]</sup>. For

tea leaves, researchers proposed improved CNNs for Grey Blight and TeaDiseaseNet for multi-scale lesions <sup>[6,9]</sup>. Detection models like YOLO and global-focused ViTs further advanced feature extraction <sup>[10,11]</sup>. However, existing models struggle to allocate dynamic attention amidst complex backgrounds and high inter-class similarities.

## 2.2. Attention mechanisms and cross-feature fusion

Standard self-attention models robust global contexts but incurs immense computational costs. Linear attention reduces overhead but degrades performance by neglecting query magnitude <sup>[8,12]</sup>. Fan *et al.* resolved this via MALA <sup>[8]</sup>. Additionally, FAMNet demonstrated the superiority of cross-feature fusion in overcoming domain interference <sup>[7]</sup>. Inspired by these, we integrate cross-attention fusion with MALA to precisely identify tiny, complex lesions in real-world scenarios.

## 3. Methodology

To address complex background noise and the loss of fine-grained lesion details in tea leaf disease classification, we propose a dual-branch fusion attention network based on ResNet-50. This framework innovatively integrates CAModule for spatial detail reconstruction and MALA module for dynamic background suppression.

### 3.1. Base network: ResNet-50

ResNet-50 serves as our backbone, balancing computational efficiency and robust feature extraction. It mitigates vanishing gradients via shortcut connections (**Table 1**), with the basic residual unit formulated as:

$$y = \sigma(F(x, \{W_i\}) + W_s x) \quad (1)$$

where  $x$  and  $y$  are inputs and outputs,  $F$  is the residual mapping,  $\sigma$  denotes ReLU, and  $W_s$  matches dimensions.

Comprising four bottleneck stages, we remove its terminal pooling and fully connected layers. This allows direct feeding of high-dimensional feature maps into the CAModule and MALA, maximally preserving fine-grained spatial lesion textures.

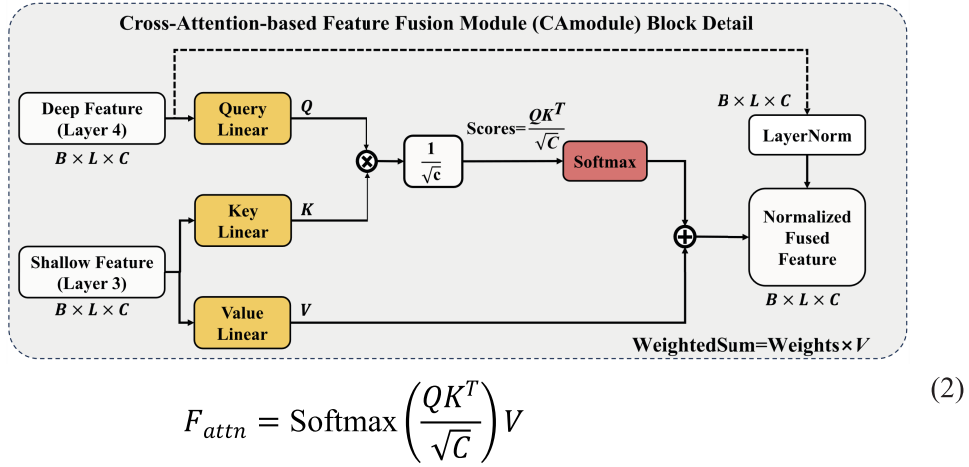
**Table 1.** Network configuration of the ResNet-18 backbone

Stage name	Output size	Layer configuration
Input stem	112×112	7×7,64, stride2 3×3 max pool, stride2
Stage 1	56×56	$\begin{bmatrix} 1 \times 1,64 \\ 3 \times 3,64 \\ 1 \times 1,256 \end{bmatrix} \times 3$
Stage 2	28×28	$\begin{bmatrix} 1 \times 1,128 \\ 3 \times 3,128 \\ 1 \times 1,512 \end{bmatrix} \times 4$
Stage 3	14×14	$\begin{bmatrix} 1 \times 1,256 \\ 3 \times 3,256 \\ 1 \times 1,1024 \end{bmatrix} \times 6$
Stage 4	7×7	$\begin{bmatrix} 1 \times 1,512 \\ 3 \times 3,512 \\ 1 \times 1,2048 \end{bmatrix} \times 3$
Custom routing	14×14 7×7	Output to CAModule (Stage 3 & 4) Output to MALA (Stage 4)

### 3.2. Cross-attention feature fusion module

To address the loss of tiny lesion details in deep layers, we propose the CAModule. Inspired by FAMNet, we adapt its cross-domain fusion mechanism for cross-layer interaction, utilizing high-resolution shallow details to dynamically calibrate deep semantics.

As illustrated in **Figure 1**, the module receives deep semantic features ( $F_{deep}$ ) and shallow texture features ( $F_{shallow}$ ). It projects  $F_{deep}$  into queries ( $Q$ ), and  $F_{shallow}$  into keys ( $K$ ) and values ( $V$ ). The adaptive cross-layer aggregation is formulated as:



**Figure 1.** Schematic diagram of the cross-attention feature fusion module.

To preserve primary semantic representations and stabilize training, a residual connection and Layer Normalization are applied:

$$F_{out} = \text{LayerNorm}(F_{deep} + F_{attn}) \quad (3)$$

This mechanism adaptively “recalls” fine-grained edges previously filtered by downsampling, establishing a robust foundation for accurate classification.

### 3.3. Magnitude-aware linear attention

To address query magnitude neglect in traditional linear attention, which causes imbalanced attention allocation, we introduce the MALA module. Maintaining complexity, MALA dynamically calibrates magnitudes to focus on discrete lesions.

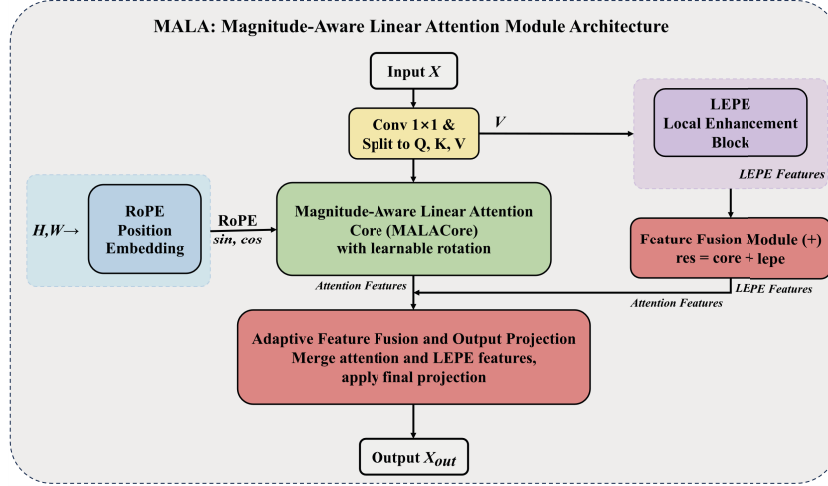
As illustrated in **Figure 2**, input  $X$  is projected into  $Q$ ,  $K$ ,  $V$ , and gate  $O$ . LEPE extracts local details  $F_{leps}$  from  $V$ , while  $Q$  and  $K$  receive 2D Rotary Position Embedding (RoPE) to form  $\hat{Q}$  and  $\hat{K}$ . Using a dynamic correction factor  $Z = (\hat{Q}\hat{K}^T)/\sqrt{d}$ , MALA recalibrates global attention:

$$F_{attn} = \hat{Q}(\hat{K}^T V) \left(1 + \frac{1}{Z + \epsilon}\right) - Z\bar{V} \quad (4)$$

This effectively suppresses healthy backgrounds. Finally, features are fused via the output projection:

$$X_{out} = \text{Conv}_{1 \times 1}((F_{attn} + F_{leps}) \odot O) \quad (5)$$

This architecture significantly enhances global perception and precise localization of complex tea lesions.

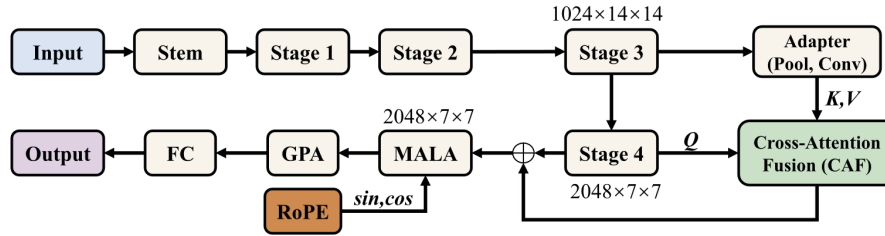


**Figure 2.** Schematic diagram of the magnitude-aware linear attention (MALA) module.

### 3.4. Overall network architecture

The proposed ResNet50-Dual-Fusion architecture is illustrated in **Figure 3**. Using a modified ResNet-50 backbone, the network extracts Stage 3 (shallow) and Stage 4 (deep) features to balance local textures with global semantics. An adapter aligns the Stage 3 features with the deep branch. The MALA module, integrated with RoPE, recalibrates attention weights to suppress backgrounds. Finally, enhanced features are processed through Global Pooling and a Fully Connected head to output classification probabilities.

In the interaction stage, Stage 4 features act as queries ( $Q$ ), fusing with adapted Stage 3 features via the CAModule to recall lesion details. This output is residually added to the original Stage 4 features. Subsequently, the MALA module, integrated with RoPE, recalibrates attention weights to suppress backgrounds. Finally, enhanced features are processed through Global Pooling and a Fully Connected head to output classification probabilities.



**Figure 3.** Schematic diagram of the overall ResNet50-dual-fusion network architecture.

## 4. Experimental results

### 4.1. Dataset and implementation details

This study utilizes a tea leaf disease dataset of 5,276 images across seven categories (e.g., Green mirid bug, Gray Blight), partitioned into training, validation, and test sets at an 8:1:1 ratio. Images are resized to and augmented via random flips and ImageNet normalization. Experiments are implemented using PyTorch on an NVIDIA RTX 3090 GPU. We employ the AdamW optimizer with a learning rate and weight decay of 0.0001, a batch size of 64, and Cross-Entropy Loss over 100 epochs. Model performance is evaluated using Accuracy, Precision, Recall, and F1-score to ensure robust assessment.



## 4.2. Model performance evaluation

Comparative experiments (**Table 2**) show that ResNet50 achieves the best performance with 82.96% accuracy, surpassing Swin-Tiny, ConvNeXt, and ViT. This superiority stems from its robust inductive bias and residual structure, which effectively capture fine-grained textures in small-scale datasets. Consequently, ResNet50 is selected as the base network for subsequent integration of the CAModule and MALA.

**Table 2.** Performance comparison of different base models on the tea leaf disease dataset

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
ResNet50	82.96	81.20	79.59	80.26
ViT-Base	72.47	69.43	67.49	68.15
Swin-Tiny	79.78	78.08	76.13	76.95
ConvNeXt-Tiny	70.79	68.06	66.36	67.01
PoolFormer-s12	73.22	70.90	69.65	70.16

## 4.3. Ablation studies

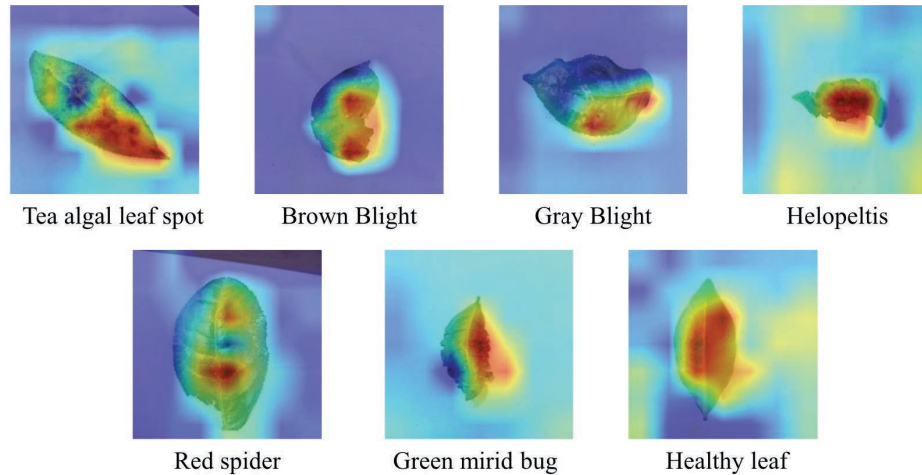
Ablation experiments (**Table 3**) validate the efficacy of CAModule and MALA. Integrating CAModule alone improves accuracy to 84.83% (+1.87%), demonstrating its ability to reconstruct fine-grained lesion details via cross-layer interaction. Meanwhile, MALA alone achieves 84.46% accuracy (+1.50%) by suppressing background noise through magnitude calibration. The full ResNet50-Dual-Fusion model achieves a peak accuracy of 85.96% (+3.00%). This highlights the synergy between “cross-layer detail reconstruction” and “intra-layer background suppression,” significantly boosting overall classification performance.

**Table 3.** Ablation study results of different module combinations

Components		Performance metrics				
CrossAttn	MALA	Acc (%)	Prec (%)	Rec (%)	F1 (%)	Improvement
		82.96	81.20	79.59	80.26	-
	√	84.46	83.19	81.48	82.12	+ 1.50%
√		84.83	84.38	81.93	82.93	+ 1.87%
√	√	85.96	84.24	82.35	83.02	+ 3.00%

## 4.4. Model interpretability analysis

Grad-CAM visualizations (**Figure 4**) demonstrate our model’s precise lesion localization across all categories. High-response regions strictly align with diseased areas, proving MALA’s efficacy in suppressing background noise via magnitude calibration. Furthermore, CAF’s cross-layer interaction enables the capture of fine-grained details, from large necrotic blights to discrete insect bites. These results confirm that ResNet50-Dual-Fusion learns discriminative pathological features rather than background biases, ensuring diagnostic reliability.



**Figure 4.** Grad-CAM visualization results of the proposed model on the tea leaf disease dataset.

## 5. Conclusion

This paper presents ResNet50-Dual-Fusion, integrating CAModule and MALA to tackle complex backgrounds and fine-grained detail loss in tea disease classification. Experiments confirm that CAModule enables cross-layer detail reconstruction, while MALA effectively suppresses environmental noise. Our method achieves 85.96% accuracy (+3.00% improvement), outperforming mainstream CNNs and Transformers. Grad-CAM results further validate its interpretability. Future work will focus on model pruning and knowledge distillation to facilitate lightweight deployment on edge-side agricultural intelligent devices.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Sladojevic S, Arsenovic M, Anderla A, et al., 2016, Deep Neural Networks based Recognition of Plant Diseases by Leaf Image Classification. *Computational Intelligence and Neuroscience*, 2016: 1–11.
- [2] He K, Zhang X, Ren S, et al., 2016, Deep Residual Learning for Image Recognition, *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 770–778.
- [3] Chen J, Liu Q, Gao L, 2019, Visual Tea Leaf Disease Recognition using a Convolutional Neural Network Model. *Symmetry*, 11(3): 343.
- [4] Bao W, Fan T, Hu G, et al., 2022, Detection and Identification of Tea Leaf Diseases based on AX-RetinaNet. *Scientific Reports*, 12(1): 1–16.
- [5] Datta S, Gupta N, 2023, A Novel Approach for the Detection of Tea Leaf Disease using Deep Neural Network. *Procedia Computer Science*, 2023(218): 2273–2286.
- [6] Pandian J, Nisha S, Kanchanadevi K, et al., 2023, Grey Blight Disease Detection on Tea Leaves using Improved Deep Convolutional Neural Network. *Computational Intelligence and Neuroscience*, 2023: 1–11.
- [7] Bo Y, Zhu Y, Li L, et al., 2025, FAMNet: Frequency-Aware Matching Network for Cross-Domain Few-Shot Medical Image Segmentation. *Proc. AAAI Conf. Artificial Intelligence*, 39(2).



- [8] Fan Q, Huang H, Ai Y, et al., 2025, Rectifying Magnitude Neglect in Linear Attention, Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2025.
- [9] Sun Y, Wu F, Guo H, 2023, TeaDiseaseNet: Multi-Scale Self-Attentive Tea Disease Detection. *Frontiers in Plant Science*, 2023(14): 1257212.
- [10] Mathew M, Mahesh T, 2022, Leaf-based Disease Detection in Bell Pepper Plant using YOLO v5. *Signal, Image and Video Processing*, 16(3): 841–847.
- [11] Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Proc. Int. Conf. Learning Representations (ICLR), 2021.
- [12] Vaswani A, Shazeer N, Parmar N, et al., 2017, Attention is all you Need. *Advances in Neural Information Processing Systems (NIPS)*, 2017(30): 5998–6008.

**Publisher's note**

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

# A Tibetan Speaker Verification Method Based on the Improved MFA-NConformer Model

Yitong Gong\*, Yuting Chen

Xinjiang Vocational University of Technology, Xinjiang, China

\*Corresponding author: Yitong Gong, 2030642289@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:** MFA-conformer methods are widely used in English and Chinese speaker recognition. Theoretically language-independent but practically language-related, Tibetan speaker recognition currently relies on traditional models with poor performance. To address this, we adopt MFA-conformer as the basic framework and propose improvements: integrating 1D depth-wise separable convolution and channel attention into the conformer feed-forward network, fusing multi-block features, and adding an intra-class correlation regularizer to GE2E loss. Experiments show the improved model reduces the equal error rate (EER) compared with the conformer baseline.

**Keywords:** Conformer block; Tibetan; GE2E loss; Speaker verification

**Online publication:** April 24, 2026

## 1. Introduction

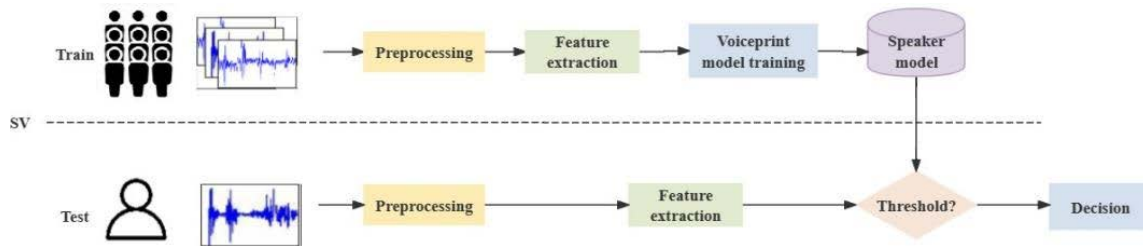
Speaker recognition technology achieves automated identity verification by analyzing voice parameters, representing a key breakthrough in the security field <sup>[1]</sup>. Since the rise of deep learning, significant progress has been made in speaker verification technology. A keyword-based text relevance speaker verification system has explored various attention mechanisms. Research indicates that self-attention models can effectively integrate relevant information within input sequences <sup>[2]</sup>. As a revolutionary new approach, the transformer model proposes an innovative solution that disrupts traditional recurrent and convolutional layer structures <sup>[3]</sup>. The transformer relies entirely on attention mechanisms, replacing the recurrent layers in traditional encoder-decoder architectures with multi-head self-attention. This self-attention mechanism serves as the core component of the transformer, enabling powerful modeling of global interactions between speech frames <sup>[4,5]</sup>. However, global self-attention still appears to be insufficient in capturing local information that is critical for speaker embedding. To address this problem, convolutional neural networks (CNNs) were introduced. A study have proposed the conformer model, which cleverly fuses components of CNN and transformer to achieve end-to-end speech recognition with the ability to extract both local and global contextual information <sup>[6]</sup>. In addition, another study further proposes

the multi-scale feature aggregation conformer (MFA-Conformer), which is a simple and effective backbone for automated speaker verification based on a convolutionally enhanced transformer (conformer) <sup>[7]</sup>. MFA-conformer is able to capture global and local features more efficiently by connecting the frame-level outputs of all conformer blocks and achieves more significant improvements in speaker feature extraction compared to traditional conformers. Speaker verification technology has progressed in the era of deep learning by introducing end-to-end modeling, transformer, conformer, and other techniques. Several studies have conducted cross-language speaker verification studies on the English-Mandarin-Viennese dialect dataset and the English-Mandarin-Taiwanese dialect dataset, respectively <sup>[8,9]</sup>. Theoretically, speaker recognition technology captures acoustic features independent of language or content. However, experimental results indicate that matching rates remain low when the same speaker registers and tests using different languages <sup>[10]</sup>. Proposed modeling spoken language or dialects as speaker features through neural networks, and completed verification based on this approach. This study employs a Tibetan corpus for speaker recognition tasks. Tibetan belongs to the Tibeto-Burman language family and is classified as a low-resource language due to its unique phonological and syntactic structures <sup>[11,12]</sup>. This paper aims to extract more individualized Tibetan speaker features using Tibetan corpus, get better validation results, and propose an improved model by exploring the structure of the MFA-conformer model, multi-scale feature aggregation new conformer (MFA-NConformer).

## 2. Research methodology

### 2.1. Tibetan speaker verification system framework

**Figure 1** shows the framework for Tibetan speaker verification, which consists of training and testing phases. The core goal is to verify whether an audio sample comes from the target speaker. In the training phase, acoustic features are extracted from enrolled speakers' audio data and input into a neural network for training to establish a voiceprint database. In the testing phase, acoustic features are extracted from the test speaker's audio data, and the extracted voiceprint features are matched against those in the database to confirm the speaker's identity.



**Figure 1.** Tibetan speaker recognition system framework.

### 2.2. Loss function

#### 2.1.1. GE2E loss

Generalized end-to-end loss (GE2E loss) is a loss function for speaker recognition tasks <sup>[13]</sup>. It maps speech feature vectors to a speaker-specific embedding space during training and leverages the Euclidean distance between embedding vectors for speaker verification. The optimization goal is to minimize intra-class distances and maximize inter-class distances. Its calculation involves the Euclidean distance comparison and a contrastive loss function to pull positive sample pairs closer and push negative sample pairs apart, thereby optimizing speaker recognition performance. The mathematical expression is given as follows:

$$L_{GE2E} = -\frac{1}{N} \sum_{j,i} \log \frac{e^{S_{j,i,j}}}{\sum_{k=1}^N e^{S_{j,i,k}}} \quad (1)$$

Where GE2E loss creates a phase matrix that defines the cosine similarity between each embedding and all centers of mass:

$$S_{j,i,k} = \omega \cdot \cos(x_{j,i}, c_k) + b \quad (2)$$

$\omega$  and  $b$  are learnable scales and biases.

### 2.1.2. GE2Eicr loss

The ‘‘Intra-class Correlation Regularizer’’ (ICR) introduces the concept of intra-class sample similarity into the loss function. Through increasing the similarity among samples within the same class, it enhances the model’s ability to distinguish between samples of the same category.

The formula for the ICR regularization term is as follows:

$$L_{ICR} = -\frac{1}{N} \sum_{i=1}^N \sum_{j \neq i}^N \cos(s_{ii}, s_{ij}) \quad (3)$$

Where  $\cos(*)$  denotes the cosine similarity,  $s_{ii}$  denotes the feature vector (embedding vector) of the  $i$  sample within the same category; and  $s_{ij}$  denotes the feature vector of the  $i$  sample and the  $j$  sample within the same category.

In combining the effects of both GE2E Loss and ICR, we define the overall loss:

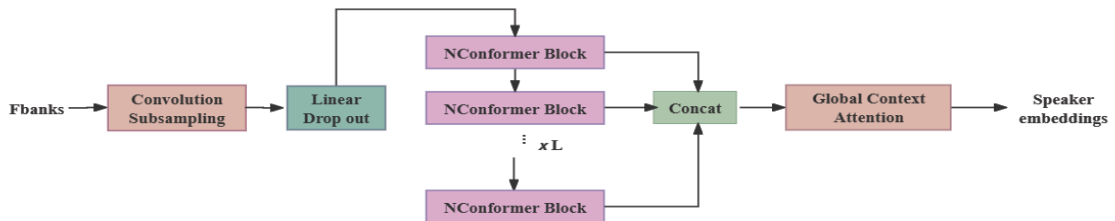
$$L_{\text{total}} = L_{GE2E} + \omega_{ICR} \times L_{ICR} \quad (4)$$

Where  $\omega_{ICR}$  is the hyperparameter used to adjust the effect of the ICR regularization term by adjusting the weights.

We can balance the contribution of the ICR regularization term to the overall loss. This loss function aims to improve the similarity of samples within the same category while driving the model to learn a more discriminative feature representation, which helps optimize speaker recognition performance.

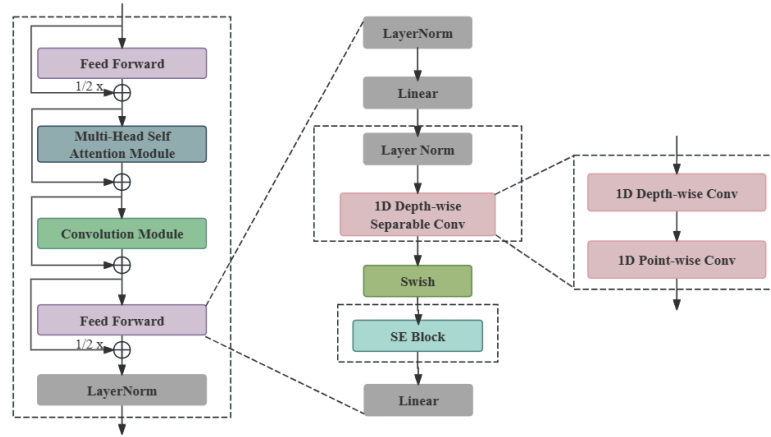
## 3. Model structure

In this experiment, we introduced the structure of the MFA-conformer model. **Figure 2** outlines the MFA-NConformer model architecture. Fbanks features undergo convolutional subsampling to cut computational costs, then go through Linear and Dropout layers, and are further processed by NConformer blocks for global and local feature extraction. Features from each NConformer block are fused via a Concat layer; an Attention Pooling layer assigns differentiated weights to outputs to extract speaker features, which are finally converted into speaker embedding through batch normalization.



**Figure 2.** Structure of the MFA-NConformer model.

As shown in **Figure 3**, NConformer is an optimization model based on the conformer model. Adding one-dimensional deep separable convolution and channel attention mechanisms to the feedforward neural network of the conformer block. The resulting NConformer blocks act in the same way as the traditional conformer model. That is, the spatial and temporal features of the input data are learned and feature representations are extracted through multilayer convolution and self-attention mechanisms<sup>[14]</sup>.



**Figure 3.** Block diagram of NConformer.

## 4. Experimental setup and evaluation

We compare the performance of GE2E loss and GE2Eier loss on the MFA-conformer and MFA-NConformer models, respectively, and contrast 1D convolution with 1D separable convolution in the feedforward layer. Notably, our goal is not SOTA performance but to verify the validity and effectiveness of the “right” margin.

### 4.1. Tibetan speaker dataset

The experiment’s Tibetan corpus was recorded with a professional microphone. Its audio follows a 16 kHz sampling rate, 16-bit mono and WAV format. The corpus contains 47 speakers, each with 500 audio clips, totaling 23,500 clips. The average clip duration is 3 seconds, and the total recording time is over 19 hours. We divide the dataset into a training set of 40 speakers and 20,000 clips, and a test set of 7 speakers and 3,500 clips (**Table 1**).

**Table 1.** Tibetan speaker dataset

Aspects	Speaker	Utterance/speaker	Utterance
Train	40	500	20000
Test	7	500	3500
Total	47		23500

### 4.2. Experimental configuration

This experiment used the PyTorch framework, running on CentOS Linux with an Intel Xeon E5-2680 v4 CPU and an 11 GB GeForce RTX 2080 Ti.

### 4.3. Evaluation indicators

Speaker recognition typically evaluates system performance using three metrics: false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER). FAR refers to the proportion of incorrectly accepted data, while FRR denotes the proportion of incorrectly rejected data.

$$FAR = \frac{FP}{FN+TN} \quad (5)$$

$$FRR = \frac{FN}{TP+FN}$$

## 5. Results and analysis

This experiment compares the MFA-conformer model with its improved variants, models employing the GE2E loss function and the GE2Eicr loss function. Under the same corpus, the equal error rate (EER) was evaluated using different optimizers and pooling layer configurations. **Table 2** shows balanced data, with FBank = 40, batch size = 32, learning rate = 0.001, and a 400-epoch training process. It compares two models' performance under two metric loss functions, with hyperparameter M (fixed at 2, words per speaker). The MFA-NConformer with GE2Eicr loss achieves a 2.6% EER, outperforming the conformer and MFA-conformer. GE2Eicr loss also outperforms GE2E loss, reducing the MFA-NConformer's error rate by 28.96%. We therefore selected this model-loss function combination for subsequent experiments.

**Table 2.** Comparison of the effects of three models using two loss functions (Adam)

Model	Optimizer	Hyperparameters	EER
MFA-Conformer	GE2E	M=2	4.10
	GE2Eicr		2.80
MFA-NConformer	GE2E		3.61
	GE2Eicr		2.65

According to **Table 3**, when M = 2, the model performs optimally with an EER of 1.90%. However, as M increases, the model's performance declines. A smaller M value allows the loss function to reduce intra-class distances more effectively, resulting in better performance.

**Table 3.** Effect of M-value on model performance (1D depthwise separable convolution)

Loss	Hyperparameters	EER
GE2Eicr	M=2	1.90
	M=3	2.15
	M=4	2.55

**Table 4** compares the results under two different optimizers (Adam and SGD). The results indicate that the SGD optimizer outperforms the Adam optimizer with an EER of 1.77%, making SGD more suitable for this improved model.

**Table 4.** Effect of M-value on model performance (1D depthwise separable convolution).

Aspects	Optimizer	EER
1D Depthwise Conv	Sgd	1.90
	Adam	3.12
1D Conv	Sgd	1.77
	Adam	2.65

## 6. Conclusion

This experiment improves Tibetan speaker recognition based on the MFA-conformer model. Introducing 1D separable convolutions (which reduce parameters and computational complexity while ensuring robust feature extraction) and channel attention mechanisms into the feedforward network, integrating multi-conformer block features, and adding an intra-class correlation regularization term to the GE2E loss function, all effectively enhance model performance. These improvements collectively promote the advancement of Tibetan speaker recognition.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Boles A, Rad P, 2017, Voice Biometrics: Deep Learning-based Voiceprint Authentication System, In: 2017 12th System of Systems Engineering Conference (SoSE), 1–6.
- [2] Chowdhury F, Wang Q, Moreno I, 2018, Attention-based Models for Text-Dependent Speaker Verification, In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5359–5363.
- [3] Peddinti V, Povey D, Khudanpur S, 2015, A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts., in Interspeech, 3214–3218.
- [4] Sang M, Zhao Y, Liu G, et al., 2023, Improving Transformer-based Networks with Locality for Automatic Speaker Verification, In: ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.
- [5] Cai D, Li M, 2023, Leveraging ASR Pretrained Conformers for Speaker Verification through Transfer Learning and Knowledge Distillation, arXiv, <https://doi.org/10.1109/TASLP.2024.3419426>
- [6] Gulati A, Qin J, Chiu C, et al., 2020, Conformer: Convolution-Augmented Transformer for Speech Recognition, arXiv, <https://doi.org/10.48550/arXiv.2005.08100>
- [7] Zhang Y, Lv Z, Wu H, et al., 2022, Mfa-Conformer: Multi-Scale Feature Aggregation Conformer for Automatic Speaker Verification, arXiv, <https://doi.org/10.48550/arXiv.2203.15249>
- [8] Li L, Wang D, Rozi A, et al., 2017, Cross-Lingual Speaker Verification with Deep Feature Learning, In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1040–1044.
- [9] Wu Y, Liao W, 2021, Toward Text-Independent Cross-Lingual Speaker Recognition using English-Mandarin-Taiwanese Dataset, In: 2020 25th International Conference on Pattern Recognition (ICPR), 8515–8522.



- [10] Thienpondt J, Desplanques B, Demuynck K, 2022, Tackling the Score Shift in Cross-Lingual Speaker Verification by Exploiting Language Information, In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7187–7191.
- [11] Schuessler A, 2024, Sino-Tibetan in Tibetan and Old Chinese. *Language and Linguistics*, 80–122.
- [12] Mokgonyane T, Sefara T, Manamela M, et al., 2019, The Effects of Data Size on Text-Independent Automatic Speaker Identification System, In: 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 1–6.
- [13] Wan L, Wang Q, Papir A, et al., 2018, Generalized End-to-End Loss for Speaker Verification, In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4879–4883.
- [14] Lyu H, Sha N, Qin S, et al., 2019, Advances in Neural Information Processing Systems. *Advances in Neural Information Processing Systems*, 32(2019).

**Publisher's note**

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



# IoT Security Situation Prediction Based on AGWO-Optimized BiGRU-ATTN

Menghao Niu, Wen Chen\*

School of Artificial Intelligence and Computer, North China University of Technology, Beijing 100144, China

\*Corresponding author: Wen Chen, 2920912486@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:** To address the complexity and variability of Internet of Things (IoT) security situation prediction, this paper proposes an IoT security situation prediction model based on an improved Grey Wolf Optimizer (AGWO) optimized Bidirectional Gated Recurrent Unit with an attention mechanism (BiGRU-ATTN). Aiming at the shortcomings of the standard Grey Wolf Optimizer, such as slow convergence and susceptibility to local optima, the algorithm is enhanced through chaotic mapping-based population initialization, a nonlinear adaptive convergence factor, and a fitness-weighted position updating strategy, thereby improving the global search capability and convergence speed. Moreover, a BiGRU network is employed to capture complex temporal correlations in security situation sequences, while an attention mechanism dynamically assigns different weights to key features. Finally, the improved grey wolf optimizer is used to optimize the hyperparameters of the BiGRU-ATTN network. Experimental results demonstrate that, compared with traditional methods, the proposed model achieves superior fitting performance and faster convergence.

**Keywords:** Network security; Situation prediction; Bidirectional gated recurrent unit; Attention mechanism; Grey wolf optimizer

**Online publication:** April 22, 2026

## 1. Introduction

With the widespread deployment of Internet of Things (IoT) devices in smart homes, industrial control systems, and healthcare monitoring, the attack surface of networks has expanded exponentially, making security situation prediction a core technology for ensuring the reliability of IoT systems<sup>[1]</sup>. Traditional prediction models often suffer from limited accuracy and slow convergence when dealing with multi-peak, high-dimensional time-series data, and they are particularly ineffective in extracting features of complex attack behaviors, which reduces the timeliness of situation assessment<sup>[2]</sup>.

The essence of IoT security situation prediction lies in analyzing the temporal evolution of network states to quantify risks and forecast future threat trends. In recent years, the integration of deep learning and optimization

algorithms has provided new perspectives for IoT security situation prediction. Zhang *et al.* proposed a BiGRU-ATTN network optimized by a deep whale optimization algorithm for network security situation prediction in IoT environments <sup>[3]</sup>. Du *et al.* combined an optimized Clockwork Recurrent Neural Network with the Grey Wolf Optimizer to capture spatiotemporal characteristics of network security situations <sup>[4]</sup>. Yang *et al.* designed a network attack behavior classification model integrating parallel feature extraction, BiGRU, and an attention mechanism for security situation assessment <sup>[5]</sup>.

However, existing studies still face limitations in modeling complex temporal dependencies and selecting critical features in network security situation time series, making it difficult to cope with the highly dynamic, large-scale, and complex feature relationships of IoT environments. To this end, this paper integrates a Bidirectional Gated Recurrent Unit (BiGRU) with a self-attention mechanism and combines it with an improved Grey Wolf Optimizer (AGWO) to construct an AGWO-optimized BiGRU-ATTN IoT security situation prediction model. This model enables more accurate characterization of complex patterns and latent dependencies during the evolution of security situations.

The main contributions of this paper are summarized as follows:

- (1) An improved Grey Wolf Optimizer (AGWO) is proposed by introducing chaotic initialization, nonlinear adaptive convergence factors, and fitness-weighted position updates to enhance optimization performance;
- (2) An AGWO-BiGRU-ATTN prediction model is designed to capture temporal dependencies and highlight important features in security situation sequences;
- (3) Experimental results demonstrate that the proposed model achieves superior prediction accuracy compared with several baseline models.

## 2. Methodology

### 2.1. BiGRU-ATTN temporal feature modeling network

The Gated Recurrent Unit (GRU) is a variant of recurrent neural networks that reduces computational complexity while maintaining strong sequence modeling capability <sup>[6]</sup>. Specifically, the input to the GRU at time step  $t$  consists of the current input vector  $x_t$  and the hidden state  $h_{t-1}$  from the previous time step, which carries relevant information from earlier states. The output of the GRU is the hidden state  $h_t$  at time step  $t$ . By combining the previous hidden state  $h_{t-1}$  and the current input  $x_t$ , the GRU generates two gating states. Among them, the update gate controls the amount of historical information retained from the previous state and the amount of new information incorporated from the candidate state, as defined in **Equation (1)**.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (1)$$

Here,  $z_t$  denotes the update gate, whose operation is defined in **Equation (2)**. The candidate hidden state  $\tilde{h}_t$  is computed according to **Equation (3)**.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(x_t W_{hx} + (r_t \odot h_{t-1}) W_{hh} + b_h) \quad (3)$$

The reset gate is defined as:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (4)$$

While GRU networks can only process time series unidirectionally, BiGRU can simultaneously process

sequential data from both forward and backward directions due to its unique bidirectional structure, enabling more comprehensive capture of dependencies within time series.

To further improve the model's ability to focus on important time steps, a self-attention mechanism is introduced. The self-attention mechanism calculates the relevance among sequence elements using the Query-Key-Value (QKV) structure. Each element dynamically assigns weights to other elements in the sequence, enabling the model to emphasize key temporal features and improve prediction accuracy.

## 2.2. Improved Grey Wolf Optimizer (AGWO)

To address the problems of the standard Grey Wolf Optimizer (GWO), such as uneven initial population distribution, insufficient balance between global exploration and local exploitation, and limited position update precision in complex high-dimensional optimization problems, this paper proposes an improved Grey Wolf Optimizer, referred to as the Advanced Grey Wolf Optimizer (AGWO). By introducing chaotic mapping-based population initialization, a population-state-based nonlinear adaptive convergence factor, and a fitness-weighted position updating strategy, AGWO systematically enhances the traditional GWO, thereby accelerating convergence speed while maintaining strong global search capability and improving final optimization accuracy. Furthermore, AGWO is employed to adaptively optimize the key hyperparameters of the BiGRU-ATTN network.

### 2.2.1. Chaotic mapping-based population initialization strategy

To increase population diversity, a Tent chaotic mapping strategy is used to initialize the population. Chaotic sequences have strong randomness and ergodicity, enabling the algorithm to generate more uniformly distributed initial solutions.

The Tent mapping function is defined as:

$$x_{k+1} = \begin{cases} \frac{x_k}{\mu} & (0 \leq x_k < \mu) \\ \frac{1-x_k}{1-\mu} & (\mu \leq x_k \leq 1) \end{cases} \quad (5)$$

After removing the initial transient iterations, the chaotic sequence is mapped to the solution space to generate the initial population.

### 2.2.2. Population-state-based nonlinear adaptive convergence factor

In the standard GWO algorithm, the convergence factor decreases linearly from 2 to 0. However, this strategy cannot adapt to changes in the population state during optimization.

To address this problem, a nonlinear convergence factor based on a power function is introduced:

$$a = 2 - 2 \times \frac{t}{T_{\max}} \quad (6)$$

where  $t$  denotes the current iteration number and  $T_{\max}$  denotes the maximum number of iterations.

This strategy enables the algorithm to maintain strong global exploration ability in the early stage and improve local search accuracy in the later stage.

### 2.2.3. Fitness-weighted position updating strategy

In the standard GWO, the positions guided by  $\alpha$ ,  $\beta$ , and  $\delta$  wolves are averaged with equal weights. This approach ignores the fitness differences among leading wolves.

Let the fitness values of the  $\alpha$ ,  $\beta$ , and  $\delta$  wolves be denoted as  $f_\alpha$ ,  $f_\beta$ , and  $f_\delta$ , respectively.

$$r_\alpha = \frac{f_\alpha}{f_\alpha + f_\beta + f_\delta + \varepsilon}, r_\beta = \frac{f_\beta}{f_\alpha + f_\beta + f_\delta + \varepsilon}, r_\delta = \frac{f_\delta}{f_\alpha + f_\beta + f_\delta + \varepsilon} \quad (7)$$

$$w_\alpha = \frac{r_\alpha}{r_\alpha + r_\beta + r_\delta}, w_\beta = \frac{r_\beta}{r_\alpha + r_\beta + r_\delta}, w_\delta = \frac{r_\delta}{r_\alpha + r_\beta + r_\delta} \quad (8)$$

Where  $\varepsilon$  is a very small constant introduced to avoid division by zero. Where  $w_\alpha + w_\beta + w_\delta = 1$ , and  $w_\alpha > w_\beta > w_\delta$ .

The final weighted position updating formula is given as follows:

$$X(t+1) = w_\alpha \cdot X_\alpha + w_\beta \cdot X_\beta + w_\delta \cdot X_\delta \quad (9)$$

When the fitness of the  $\alpha$  wolf is significantly better than that of the  $\beta$  and  $\delta$  wolves ( $f_\alpha \leq f_\beta, f_\delta$ ), the weight  $w_\alpha$  approaches 1, and the  $\omega$  wolves converge strongly toward the position of the  $\alpha$  wolf.

### 3. Experiments and analysis

#### 3.1. Dataset description

Experiments are conducted using the publicly available ToN-IoT dataset <sup>[7]</sup>. The dataset simulates a realistic IoT environment and contains multiple attack types, including DoS, scanning, injection, ransomware, and others. During preprocessing, categorical features are converted into numerical representations through one-hot encoding. Min-Max normalization is then applied to scale all features into the range [0,1].

#### 3.2. Network security situation value construction

Since the ToN-IoT dataset does not provide ground-truth security situation values, an attack-threat-based method is used to construct situation indicators <sup>[8]</sup>. The proposed indicator system incorporates both an attack quantity factor and an attack threat factor. The attack quantity factor represents the number of attack samples within a given time interval and is denoted by  $N$ . The attack threat factor reflects the threat level posed by a specific attack type to network security and is denoted by  $X_i$ . The threat factors corresponding to different attack types are listed in **Table 1**. Accordingly, the security situation value at time interval  $t$ , denoted as  $SA(t)$ , is defined as:

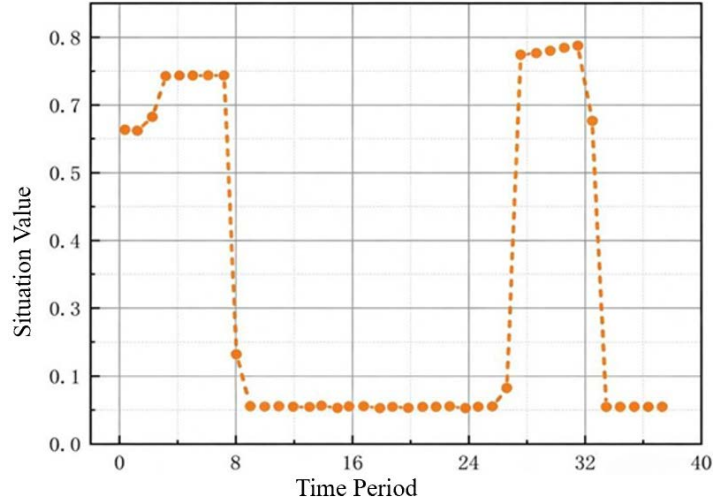
$$SA(t) = f(N, X_i) = \sum_{i=1}^N X_i \quad (10)$$

**Table 1.** Attack threat factors

Attack type	Threat factor	Attack type	Threat factor
normal	1	mitm	6
password	2	backdoor	7
scanning	3	injection	8
dos	4	ddos	9
xss	5	ransomware	

According to the temporal order of attack samples, every 800 consecutive samples are treated as one time interval. The raw situation values  $SA(t)$  calculated for all intervals are then normalized and mapped into the range [0,1]. When network attacks occur frequently, the corresponding security situation score increases, indicating a

higher threat level; conversely, when the number of attacks is small, the situation score decreases, reflecting a lower threat level. Based on **Equation (10)**, the ground-truth situation value for each time interval is generated, as illustrated in **Figure 1**.



**Figure 1.** Ground-truth security situation values.

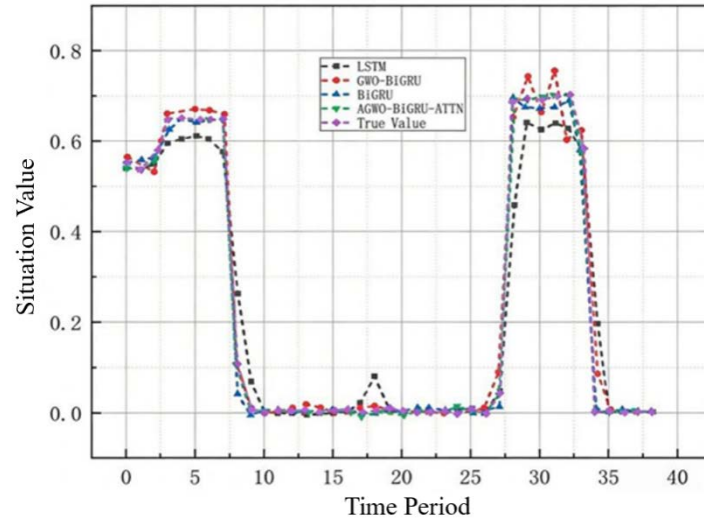
To quantitatively evaluate the performance of different models in the security situation prediction task, this study adopts the Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination ( $R^2$ ) as evaluation metrics. Specifically, MSE measures the average squared deviation between the predicted values and the ground-truth values, MAPE reflects the relative magnitude of prediction errors, and  $R^2$  evaluates the goodness of fit of the model to the data.

### 3.3. Experimental results and analysis

This section evaluates the performance of the proposed AGWO-BiGRU-ATTN model from three perspectives: prediction curve comparison, quantitative evaluation metrics, and model mechanism analysis. The proposed model is compared with several representative baseline models, including LSTM, BiGRU, and GWO-BiGRU. All comparative experiments are conducted under the same data partitioning strategy and parameter settings to ensure fairness, and the models are evaluated in terms of stability, convergence behavior, and prediction accuracy.

#### 3.3.1. Comparison of network security situation prediction

To verify the prediction capability of the AGWO-BiGRU-ATTN model in IoT environments, the predicted security situation sequences of different models are first visualized and compared. The comparison involves AGWO-BiGRU-ATTN, BiGRU, GWO-BiGRU, and LSTM models. As shown in **Figure 2**, the prediction trend of the AGWO-BiGRU-ATTN model is highly consistent with the ground-truth situation values and exhibits a significant improvement over the other models.



**Figure 2.** Comparison results.

### 3.3.2. Analysis of performance evaluation metrics

To further evaluate the performance of the proposed model in IoT security situation prediction tasks, **Table 2** presents the evaluation results of different models in terms of MAPE, MSE, and  $R^2$ . The comprehensive evaluation based on these three metrics indicates that AGWO-BiGRU-ATTN not only significantly reduces prediction errors but also more accurately captures the complex fluctuations inherent in IoT security situations. The observed performance gains can be attributed to the superior hyperparameter configurations obtained by AGWO during the optimization stage, which enable the BiGRU-ATTN model to achieve more stable and efficient performance when dealing with non-stationary and high-noise time-series data.

**Table 2.** Comparison of model performance

Model name	MAPE/%	$R^2$	MSE
AGWO-BiGRU-ATTN	19.33	0.972	0.00310
GWO-BiGRU	30.62	0.932	0.00578
BiGRU	35.47	0.885	0.00925
LSTM	40.25	0.867	0.01127

## 4. Conclusion

This paper addresses the limitations of existing prediction models in dynamically capturing IoT security situations, and proposes an IoT security situation prediction model based on AGWO-optimized BiGRU-ATTN. The model precisely extracts key features from temporal data through a self-attention mechanism, thereby enhancing the stability of model training. GWO undertakes parameter optimization tasks within the model, enabling more efficient capture of security situation changes in dynamic environments, and thus further improving the model's prediction accuracy. Experimental results demonstrate that the model exhibits significant advantages in both accuracy and stability.



## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Begum M, Yogeshwaran A, Nagarajan N, et al., 2025, Dynamic Network Security Leveraging Efficient CoviNet with Granger Causality-Inspired Graph Neural Networks for Data Compression in Cloud IoT Devices. *Knowledge-Based Systems*, 2025(309): 112859.
- [2] Jablaoui R, Liouane N, 2025, Network Security Based Combined CNN-RNN Models for IoT Intrusion Detection System. *Peer-to-Peer Networking and Applications*, 18(3).
- [3] Zhang S, Fu Q, An D, 2023, Network Security Situation Prediction Model Based on VMD Decomposition and DWOA Optimized BiGRU-ATTN Neural Network. *IEEE Access*, 2023(11): 129507–129535.
- [4] Du X, Ding X, Tao F, 2023, Network Security Situation Prediction Based on Optimized Clock-Cycle Recurrent Neural Network for Sensor-Enabled Networks. *Sensors*, 23(13): 6087.
- [5] Yang H, Zhang Z, Xie L, et al., 2022, Network Security Situation Assessment with Network Attack Behavior Classification. *International Journal of Intelligent Systems*, 37(10): 6909–6927.
- [6] Rahul D, Salem F, 2017, Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks, *Midwest Symposium on Circuits and Systems*, 1597–1600.
- [7] Moustafa N, 2021, A New Distributed Architecture for Evaluating AI-based Security Systems at the Edge: Network TON\_IoT Datasets. *Sustainable Cities and Society*, 2021(72): 102994.
- [8] Zhao D, Wu Y, Zhang H, 2022, Network Security Situation Prediction based on IPSO-BiLSTM. *Computer Science*, 49(7): 357–362.

### Publisher's note

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



# A Zero-Dynamics Attack Detection Method for Offshore Wind Power Systems

Kaige Chen, Hongran Li, Zeyu Zhang, Zhaoman Zhong, Lei Hu

School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, Jiangsu, 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 increase in the scale and complexity of offshore wind power systems, zero-dynamics attacks pose a severe threat to the cyber security of such systems. Their concealment makes them difficult to detect using traditional output observation-based methods. To address this problem, this paper proposes a zero-dynamics attack detection framework integrating adaptive watermarking and Kalman filtering, which achieves effective attack identification by embedding an adaptive watermark into the system input and conducting residual analysis. Simulation results show that the proposed method can quickly detect zero-dynamics attacks without affecting the normal operation of the system.

**Keywords:** Zero-dynamics attack; Adaptive watermarking; Kalman filtering; Residual detection

**Online publication:** April 22, 2026

## 1. Introduction

With the growth of global energy demand, offshore wind power has become an important direction for energy transition. The increase in the scale and complexity of offshore wind power systems has brought new cyber security challenges<sup>[1]</sup>.

In cyber-physical systems (CPS), attackers can threaten infrastructure by manipulating control systems<sup>[2,3]</sup>. In smart grids, cyber-physical attacks mainly include false data injection and denial-of-service attacks. As a type of covert attack, zero-dynamics attacks can manipulate the internal states of a system without significantly altering its output, making it difficult for traditional output observation-based detection methods to identify such attacks in a timely manner<sup>[4-7]</sup>.

Existing research has mainly focused on false data injection and denial-of-service attacks, with relatively few studies on zero-dynamics attacks<sup>[8,9]</sup>. Hoehn *et al.* enhances attack detectability by introducing a modulation matrix but alters the system structure<sup>[10]</sup>. Wang *et al.* constructs an output prediction model based on the Byrnes-Isidori normal form, which is complex to implement and relies on a complete system model<sup>[11]</sup>.

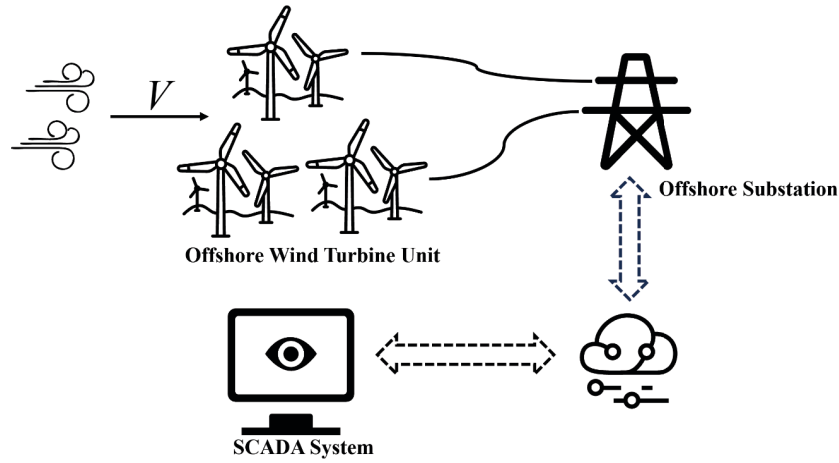
To solve the above problems, this paper proposes a zero-dynamics attack detection framework integrating adaptive watermarking and Kalman filtering, which realizes the rapid identification of zero-dynamics attacks

through residual analysis without affecting the normal operation of the system.

## 2. Offshore wind power generation systems

### 2.1. Structure of offshore wind power generation systems

An offshore wind power system consists of a wind turbine array, an offshore substation and a control center, as shown in **Figure 1**, and its monitoring and control are realized through a SCADA system<sup>[12]</sup>. The introduction of communication networks also brings security risks, as attackers may inject malicious signals through the network to interfere with system operation<sup>[13]</sup>.

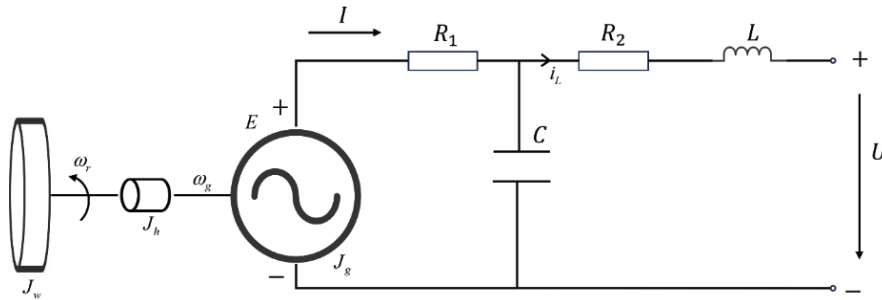


**Figure 1.** Structure of offshore wind power generation system.

### 1.2 Model of offshore wind power generation systems

To analyze the impact of attack signals on offshore wind power systems, an equivalent circuit model of the offshore wind power system is constructed based on the findings of Le *et al.*, as shown in **Figure 2**<sup>[14]</sup>. The branch voltage balance equation is given by:

$$U = U_C - i_L R_2 - L \frac{di_L}{dt} \quad (1)$$



**Figure 2.** Equivalent circuit of the offshore wind power system.

Based on Kirchhoff's voltage and current laws, we obtain:

$$U_C = E - U_{R1} \quad (2)$$

$$i_L = I - i_C \quad (3)$$

The generator is modeled as an internal voltage source, whose induced electromotive force is proportional to the rotational speed:

$$E = K_g \omega_g \quad (4)$$

The torque balance equation is:

$$J \frac{d\omega_g}{dt} = K_t I \quad (5)$$

where  $K_t$  and  $K_g$  are the electromagnetic torque constant and back electromotive force constant, respectively;  $i$  is the current;  $J$  is the total moment of inertia, and:

$$J = J_w + J_g + J_h \quad (6)$$

Substituting **Equation (2)** and **(3)** into **Equation (1)**, we get:

$$U = K_g \left( a_0 \frac{d\omega_g}{dt} + a_1 \frac{d^2\omega_g}{dt^2} + a_2 \frac{d^3\omega_g}{dt^3} + \omega_g \right) \quad (7)$$

$$\text{where } a_0 = \frac{R_2 C K_t K_g - J(R_1 + R_2)}{K_t K_g}, \quad a_1 = \frac{L C K_t K_g - J L - J C R_1 R_2}{K_t K_g}, \quad a_2 = -\frac{J L C R_1}{K_t K_g}.$$

Applying the Laplace transform to **Equation (7)** and simplifying it, the system transfer function is obtained as:

$$G_1(s) = \frac{\omega(s)}{U(s)} = \frac{K}{a_2 s^3 + a_1 s^2 + a_0 s + 1} \quad (8)$$

where is the system gain.

After introducing a PI controller  $G_2(s) = \frac{k_p s + k_i}{s}$ , the system model is expressed as:

$$\begin{aligned} G(s) &= \frac{G_1 G_2}{1 + G_1 G_2} \\ &= \frac{K(k_p s + k_i)}{a_2 s^4 + a_1 s^3 + a_0 s^2 + (K k_p + 1)s + K k_i} \end{aligned} \quad (9)$$

Rewriting **Equation (9)** in the continuous-time state-space form:

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (10a)$$

$$y(t) = Cx(t) \quad (10b)$$

$$\text{Where } A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -K k_i & -K k_p - 1 & -a_0 & -a_1 \end{bmatrix}, B = [0 \ 0 \ 0 \ 1]^T, C = [K k_i \ K k_p \ 0 \ 0].$$

The discrete-time state-space model is derived by discretizing **Equation (10)** using the zero-order hold method:

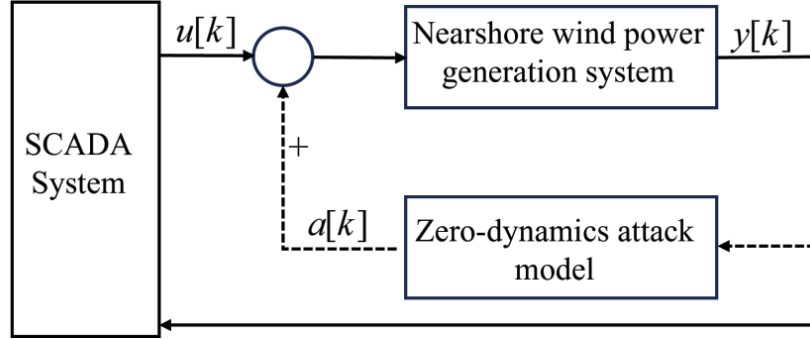
$$x[k+1] = A_d x[k] + B_d u[k] \quad (11a)$$

$$y[k] = C_d x[k] \quad (11b)$$

$$\text{where } A_d = e^{AT}, \quad B_d = A^{-1}(e^{AT} - I)B.$$

## 2. Zero-dynamics attack mechanism

A prerequisite for implementing a zero-dynamics attack is the existence of unstable zeros in the system. It can be seen from **Equation (9)** that the relative degree of the system is 3. According to Weller *et al.*, at least one unstable sampling zero will be inevitably introduced after discretization when the system meets specific conditions <sup>[15]</sup>. Therefore, the system described by **Equation (9)** is vulnerable to zero-dynamics attacks.



**Figure 3.** Zero-dynamics attack mechanism.

As shown in **Figure 3**, the attacker generates an attack signal by constructing an auxiliary internal state for zero-dynamics and injects it into the system input. Thus, the attacker usually constructs the following zero-dynamics attack model:

$$z[k + 1] = A_i z[k] \quad (12a)$$

$$a[k] = C_i z[k] \quad (12b)$$

After the attack, **Equation (11)** become:

$$x[k + 1] = A_d x[k] + B_d (u[k] + a[k]) \quad (13a)$$

$$y[k] = C x[k] \quad (13b)$$

## 3. Real-time detection based on adaptive watermarking

### 3.1. Adaptive watermark injection strategy

In the detection framework, the adaptive watermark adjusts the control input by  $u[k]$  introducing a perturbation  $\Delta u[k]$ :

$$u'[k] = u[k] + \Delta u[k] \quad (14)$$

where  $w(k)$  is the adaptive watermark perturbation.

The intensity of the watermark perturbation is adjusted according to the mean square value of the input signal:

$$V_k = \frac{1}{|I_k|} \sum_{i \in I_k} u[i]^2 \quad (15)$$

The variance of the watermark noise is:

$$\sigma_k^2 = \alpha V_k \quad (16)$$

where  $\alpha > 0$  is the watermark coefficient.

The watermark perturbation  $\Delta u[k]$  is generated following a Gaussian distribution:

$$\Delta u[k] \sim N(0, \sigma_k^2) \quad (17)$$

### 3.2. Output prediction based on steady-state Kalman filtering

To characterize modeling errors and measurement noise, process noise  $w[k]$  and measurement noise  $v[k]$  are introduced on the basis of **Equation (11)** to establish the Kalman filter model:

$$x[k+1] = A_d x[k] + B_d u'[k] + w[k] \quad (18)$$

$$y[k] = C_d x[k] \quad (19)$$

where  $w[k] \sim N(0, Q_k)$ ,  $v[k] \sim N(0, R_k)$ .

The steady-state Kalman gain is obtained by solving the Riccati equation:

$$P = A_d P A_d^T + Q_k - A_d P C_d^T (C_d P C_d^T + R_k)^{-1} C_d P A_d^T \quad (20)$$

$$K = P C_d^T (C_d P C_d^T + R_k)^{-1} \quad (21)$$

Based on this filter, the system state and output are predicted:

$$x[k+1|k] = A_d x[k|k] + B_d u'[k] \quad (22a)$$

$$\bar{y}[k+1|k] = C_d \bar{x}[k+1|k] \quad (22b)$$

The deviation is calculated to correct the state:

$$\tilde{y}[k+1] = y[k+1] - \bar{y}[k+1|k] \quad (23)$$

$$x[k+1|k+1] = x[k+1|k] + K \tilde{y}[k+1] \quad (24)$$

Since the attacker cannot synchronize the watermark perturbation, a deviation arises between the actual output  $y_{u'+a}[k]$  and the predicted output  $\bar{y}[k]$  after the attack. The residual is defined as:

$$r[k] = \bar{y}[k] - y_{u'+a}[k] \quad (25)$$

### 3.3. Residual processing and statistical detection

Amplitude limiting filtering is used to constrain the residual, and recursive median filtering is applied to smooth the residual, yielding the final smoothed residual statistic:

$$\tilde{r}[k] = \frac{1}{w-2} \sum_{i=2}^{w-1} \bar{r}[i] \quad (26)$$

where  $N$  is the window length of the median filter.

The covariance matrix of the Kalman filter residual is:

$$S = C_d P_1 C_d^T + R \quad (27)$$

The residual statistic is constructed as:

$$g[k] = \sum_{i=k-w+1}^k \tilde{r}[i]^T S^{-1} \tilde{r}[i] \quad (28)$$

In the absence of an attack,  $\tilde{r}[k]$  approximately follows a zero-mean Gaussian distribution, so  $g[k]$

asymptotically follows a chi-square distribution with  $w \cdot n_y$  degrees of freedom, where  $n_y$  is the output dimension. Given a significance level  $\alpha$ , the detection threshold is set as:

$$\delta = \chi_{1-\alpha}^{-1}(w \cdot n_y) \quad (29)$$

When  $g[k] > \delta$ , the system is judged to be under a zero-dynamics attack.

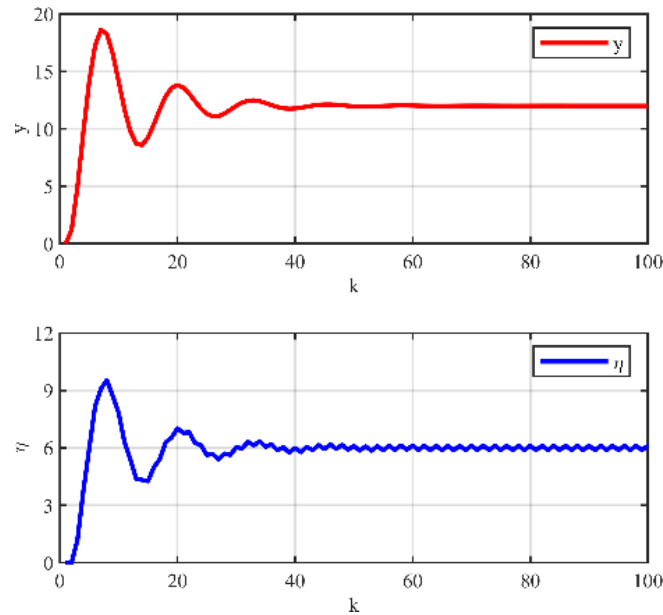
## 4. Simulation experiments

### 4.1. Performance of the offshore wind power system under zero-dynamics attacks

The simulation parameters are shown in **Table 1**. **Figure 4** shows that the external output and internal dynamics of the system gradually stabilize without the introduction of adaptive watermarking.

**Table 1.** System parameters

Parameter/unit	Value
Rotor inertia $J_d/(\text{kg} \cdot \text{m}^2)$	$2.137 \times 10^{-5}$
Generator inertia $J_m/(\text{kg} \cdot \text{m}^2)$	$4 \times 10^{-6}$
Gearbox inertia $J_h/(\text{kg} \cdot \text{m}^2)$	$6.5 \times 10^{-7}$
Resistance $R_1/\Omega$	4.6
Resistance $R_2/\Omega$	5.4
Inductance $L/\text{H}$	4.3
Capacitance $C/\text{F}$	1.3



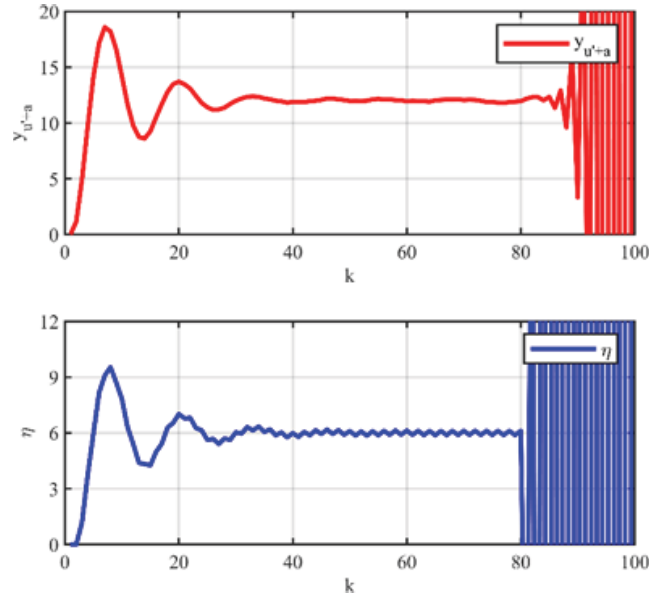
**Figure 4.** Normal operation of the offshore wind power system.

The zero-dynamics attack model proposed by Wang *et al.* is adopted <sup>[11]</sup>:

$$z[k + 1] = G_0 z[k] \quad (30a)$$

$$a[k] = C_0 z[k] \quad (30b)$$

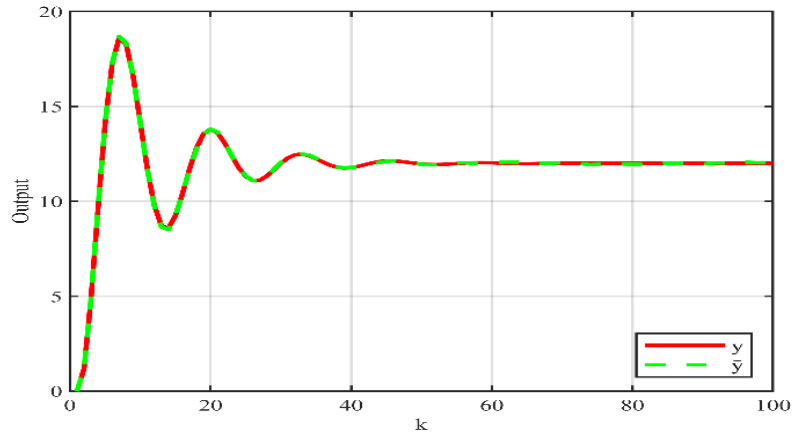
The initial state  $\|G_0\| = -2$  is set, and the attack is applied to the system with embedded adaptive watermarking. The attack is set to occur at the 80th time step, with a total simulation duration of 100 time steps. **Figure 5** shows that the external output remains stable for a short time after the attack, while the internal dynamics diverge immediately, reflecting the concealment and destructiveness of zero-dynamics attacks.



**Figure 5.** Offshore wind power system under zero-dynamics attack.

#### 4.2. Simulation results of the adaptive-Kalman filter detection framework

The Kalman filter is used to predict the system output with embedded watermarking. **Figure 6** shows that the predicted output is highly consistent with the actual output under normal operating conditions.

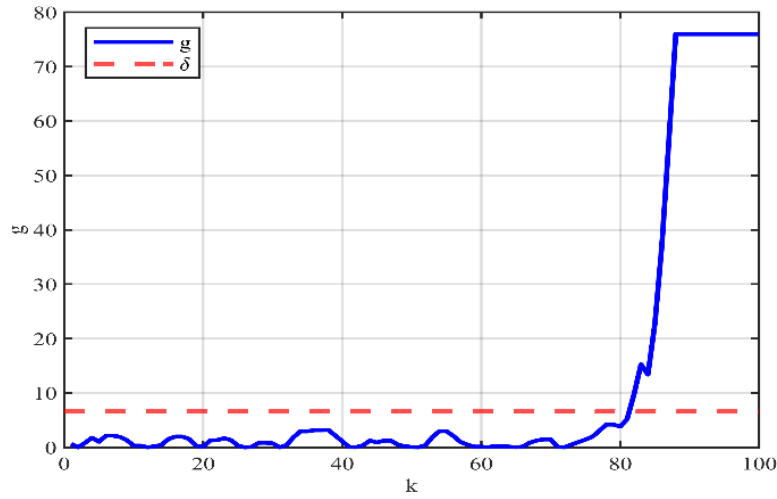


**Figure 6.** Comparison between the Kalman filter predicted output and the actual output.

The window length is set, and at a significance level of , the detection threshold . As shown in **Figure 7**, crosses the threshold rapidly after the attack occurs, realizing the fast identification of the zero-dynamics attack. The results verify that the proposed detection framework has good detection performance without affecting the



normal operation of the system.



**Figure 7.** Detection result of zero-dynamics attacks based on the Chi-square statistic.

## 5. Conclusion

Aiming at the problem of zero-dynamics attacks on offshore wind power systems, this paper proposes a two-layer detection framework integrating adaptive watermarking and Kalman filtering. Simulation results show that the method can quickly identify attacks without affecting the normal operation of the system and effectively improve the security of offshore wind power systems.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Knack A, Syn Y, Tam K, 2021, Enhancing the Cyber Resilience of Offshore Wind. *Energy Systems Catapult*, 2021: 47.
- [2] Lee S, Huh J, 2019, An Effective Security Measures for Nuclear Power Plant using Big Data Analysis Approach. *Journal of Supercomputing*, 75(8): 4251–4267.
- [3] Davis R, Keskin O, 2024, Cyber Threat Modeling for Water and Wastewater Systems: Contextualizing STRIDE and DREAD with the Current Cyber Threat Landscape, 2024 Systems and Information Engineering Design Symposium (SIEDS), 1–6.
- [4] Mukherjee D, 2022, Data-Driven False Data Injection Attack: A Low-Rank Approach. *IEEE Transactions on Smart Grid*, 13(3): 2479–2482.
- [5] Shim H, Back J, Eun Y, et al., 2022, Zero-Dynamics Attack, Variations, and Countermeasures, 117–140.
- [6] Pasha S, Ayub A, 2021, Zero-Dynamics Attacks on Networked Control Systems. *Journal of Process Control*, 2021(105): 99–107.
- [7] Li X, Jiang C, Du D, et al., 2023, A Novel State Estimation Method for Smart Grid under Consecutive Denial of Service Attacks. *IEEE Systems Journal*, 17(1): 513–524.

- [8] Zheng C, Wang X, Luo X, et al., 2022, An OpenPLC-based Active Real-Time Anomaly Detection Framework for Industrial Control Systems, 2022 China Automation Congress (CAC), 6028–6033.
- [9] Jin S, 2024, False Data Injection Attack Against Smart Power Grid based on Incomplete Network Information. *Electric Power Systems Research*, 2024(230): 110294.
- [10] Hoehn A, Zhang P, 2016, Detection of Covert Attacks and Zero Dynamics Attacks in Cyber-Physical Systems, 2016 American Control Conference (ACC), 396–401.
- [11] Wang Z, Zhang H, Cao X, et al., 2024, Modeling and Detection Scheme for Zero-Dynamics Attack on Wind Power System. *IEEE Transactions on Smart Grid*, 15(1): 934–943.
- [12] Sheng L, Li C, Gao M, et al., 2025, A Review of SCADA-based Condition Monitoring for Wind Turbines via Artificial Neural Networks. *Neurocomputing*, 2025(633): 129830.
- [13] Amin M, El-Sousy F, Aziz G, et al., 2021, CPS Attacks Mitigation Approaches on Power Electronic Systems with Security Challenges for Smart Grid Applications: A Review. *IEEE Access*, 2021(9): 38571–38601.
- [14] Le H, Dang P, Pham A, et al., 2020, System Identifications of a 2DOF Pendulum Controlled by QUBE-Servo and its Unwanted Oscillation Factors. *Archive of Mechanical Engineering*, 67(3): 435–450.
- [15] Weller S, Moran W, Ninness B, et al., 2001, Sampling Zeros and the Euler-Frobenius Polynomials. *IEEE Transactions on Automatic Control*, 46(2): 340–343.

**Publisher's note**

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

# Physics Informed Hybrid Quantum-Classical Dispatching for LargeScale Renewable Power Systems: A Noise-Resilient Framework

Fu Zhang<sup>1</sup>, Yuming Zhao<sup>2</sup>

<sup>1</sup>Lanzhou Petrochemical University of Vocational Technology, Lanzhou 730060, Gansu, China

<sup>2</sup>Lanzhou Aviation Technology College, Lanzhou 730030, Gansu, 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:** Rising renewable penetration introduces severe non-convexity in power dispatching, straining classical optimization. While variational quantum algorithms (VQAs) on NISQ devices offer combinatorial potential, “black-box” approaches struggle with scalability and grid constraints. We propose the physics-informed hybrid quantum-classical dispatching (PI-HQCD) framework to address these limitations. PI-HQCD maps power flow and storage constraints directly into a topology-aware Hamiltonian, shrinking the search space. A noise-adaptive regularization technique bounds the objective’s Lipschitz constant, ensuring convergence under measurement noise. Experiments on IEEE 39-bus and 118-bus systems show PI-HQCD outperforms stochastic dual dynamic programming (SDDP) in cost and renewable utilization. Theoretical analysis confirms our topology-aligned ansatz achieves gradient variance scaling, mitigating barren plateaus. This work bridges physical laws and quantum algorithms for next-generation grid operations.

**Keywords:** Hybrid quantum-classical optimization; Physics-informed learning; Renewable power dispatch; Variational quantum algorithms; Noise resilience

**Online publication:** April 22, 2026

## 1. Introduction

Grid operations face unprecedented pressure from the net-zero transition. As inertial machines are replaced by stochastic renewables, particularly in distributed generation contexts, state spaces exceed the real-time capacity of classical methods like stochastic dual dynamic programming (SDDP) <sup>[1–6]</sup>. This necessitates scalable frameworks <sup>[7]</sup>. Quantum computing offers novel solutions for high-dimensional combinatorial problems <sup>[8]</sup>. In the NISQ era, variational quantum algorithms (VQAs) like QAOA are promising <sup>[9,10]</sup>. However, applying general VQAs to dispatching faces three hurdles as follows <sup>[11–16]</sup>:

- (1) Physical agnosticism: Generic ansatzes (e.g., Hardware-Efficient Ansatz) ignore topology, yielding inefficient search spaces <sup>[17]</sup>;
- (2) Barren plateaus: Unstructured encodings suffer vanishing gradients, preventing scaling <sup>[18,19]</sup>;

- (3) Noise sensitivity: Failure to distinguish hardware errors from physical violations leads to infeasible solutions<sup>[20,21]</sup>.

We propose PI-HQCD to bridge quantum algorithms and power engineering. Unlike black-box optimization, we embed reduced-order physical models, power flow sensitivities and storage dynamics, into the Hamiltonian and circuit<sup>[22,23]</sup>.

Contributions include as follows:

- (1) Topology-aware encoding: Mapping grid adjacency to qubit interactions preserves sparsity;
- (2) Noise-resilient regularization: A dynamic cost function balances constraints and noise, theoretically guaranteeing stability;
- (3) Scalability: We prove topology alignment improves gradient variance to ;
- (4) Hybrid loop: Hierarchical integration of quantum sampling and classical projection.

Validations on IEEE 39-bus and 118-bus systems confirm PI-HQCD outperforms SDDP in cost and renewable uptake under noise.

## 2. Problem formulation

### 2.1. Multiperiod stochastic dispatch

We define the decision vector  $x_t = [g_t, s_t, d_t]$  as generation, storage, and controllable demand at time  $t$ , with renewable uncertainty  $w_t$ . The dispatch horizon is  $t=1, \dots, T$ . The objective is:

$$\min_{g,s,d} E_w \sum_{t=1}^T C(g_t, s_t, d_t, w_t)$$

Subject to as follows:

- (1) Power balance:  $P_{inj}(g_t, s_t, d_t, w_t) = 0$ ;
- (2) Network constraints:  $|F_t| \leq F^{max}$ ;
- (3) Generator limits:  $P_g^{min} \leq P_{g,t} \leq P_g^{max}$ ;
- (4) Storage dynamics:  $SOC_{t+1} = SOC_t + \eta_c s_t^+ - s_t^- / \eta_d$ ;
- (5) Ramping:  $|P_{g,t} - P_{g,t-1}| \leq R_g$ .

### 2.2. Physicsreduced linearization

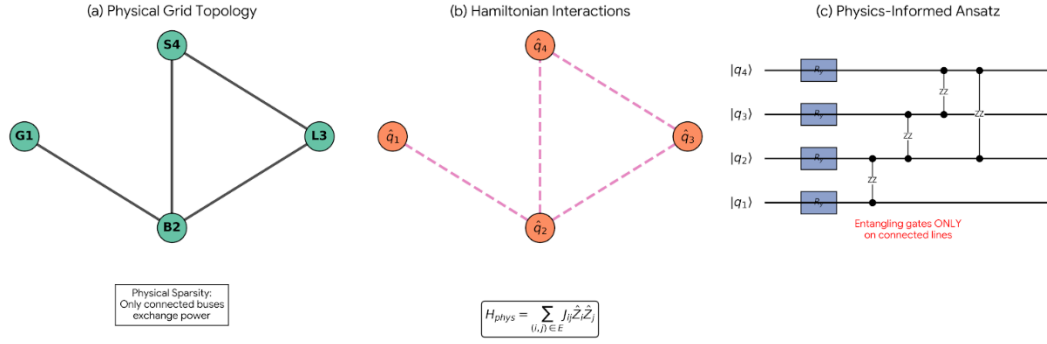
To enable efficient Hamiltonian construction, nonlinear AC constraints are locally linearized around an operating point using sensitivity matrices:

$$\Delta F \approx J_F \Delta x, \Delta V \approx J_V \Delta x,$$

where  $J_F$  and  $J_V$  are updated periodically in the classical loop.

## 3. Physicsinformed quantum encoding

Unlike generic ansatzes with all-to-all connectivity, PI-HQCD exploits power network sparsity. The quantum circuit structure is isomorphic to the physical grid topology (**Figure 1**). Entangling gates are applied only between qubits representing physically connected buses, reducing circuit depth and parameter count to avoid barren plateaus.



**Figure 1.** Schematic of the physics-informed quantum encoding strategy. (a) The physical power network topology. (b) The corresponding Hamiltonian interaction graph, preserving sparsity. (c) The Physics-Informed Ansatz, where entangling gates (ZZ) are placed exclusively between physically connected qubits.

### 3.1. Blockstructured Hamiltonian

The Hamiltonian is decomposed as:

$$\hat{H} = \hat{H}_{cost} + \hat{H}_{phys} + \hat{H}_{risk},$$

where  $\hat{H}_{cost}$  represents quadratic costs,  $\hat{H}_{phys}$  penalizes flow violations and SOC deviations, and  $\hat{H}_{risk}$  accounts for scenario variance. Each block is encoded on separate qubit registers for modular scalability.

### 3.2. Parameterized quantum circuit

Dispatch variables are encoded using rotation gates and entangling layers aligned with network topology adjacency, accelerating convergence.

## 4. Mathematical mapping, convergence and stability analysis

### 4.1. Dispatch-to-qubit encoding

Let the continuous dispatch vector be  $x_i \in \mathbb{R}^n$ . Each variable is discretized using  $b$  qubits via affine binary expansion:

$$x_{t,i} = x_i^{min} + \Delta_i \sum_{k=0}^{b-1} 2^k q_{i,k}, \quad q_{i,k} \in \{0,1\},$$

$$\text{where } \Delta_i = (x_i^{max} - x_i^{min}) / (2^b - 1).$$

The quadratic objective  $C(x) = x^T Q x + c^T x$  is mapped to a QUBO Hamiltonian, a formulation demonstrating robust scalability in diverse complex network optimizations:

$$\hat{H}_{cost} = \sum_{i,j} Q_{ij} \hat{q}_i \hat{q}_j + \sum_i c_i \hat{q}_i$$

### 4.2. Physics-constrained Hamiltonian construction

Linearized constraints  $J_F x \leq F^{max}$  are encoded as soft penalties:

$$\hat{H}_{phys} = \alpha \sum_l (J_{F,l} x - F_l^{max})^2$$

Storage dynamics are enforced by:

$$\hat{H}_{soc} = \gamma \sum_t (SOC_{t+1} - SOC_t - \eta_c s_t^\dagger + s_t^- / \eta_d)^2$$

### 4.3. Variational objective and gradient estimation

The parameterized quantum state is  $|\psi(\theta)\rangle = U(\theta)|0\rangle^{\otimes N}$ . The optimization objective is  $J(\theta) = \langle\psi(\theta)|\hat{H}|\psi(\theta)\rangle$ . Gradients are evaluated via the parameter-shift rule, enabling unbiased stochastic estimation without explicit differentiation.

### 4.4. Convergence properties

#### 4.4.1. Expected convergence rate

We model measurement noise as an unbiased stochastic gradient oracle  $g^{(k)}$ . Assuming  $J(\theta)$  is  $\mu$ -smooth (A1), gradients are unbiased (A2), and variance is bounded by  $\sigma^2$  (A3), the update  $\theta^{(k+1)} = \theta^{(k)} - \eta g^{(k)}$  yields:

$$\min_{0 \leq k \leq K-1} E \|\nabla J(\theta^{(k)})\|^2 = O(1/\sqrt{K})$$

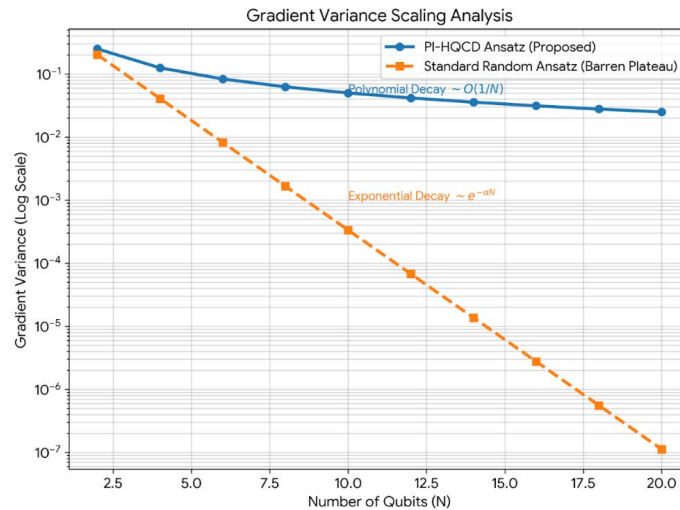
To reach  $E \|\nabla J\|^2 \leq \varepsilon$ , shot complexity scales as  $S = O(1/\varepsilon^2)$ . Noise-adaptive regularization reduces the effective smoothness constant to  $L_{eff} \leq L/(1+\beta\sigma^2)$ , improving stability.

#### 4.4.2. Projected hybrid update

The PI-HQCD iteration includes a classical feasibility projection  $\Pi_C$ . For the effective objective  $J_{eff}$  with noise-adaptive weights  $w_i = 1/(1+\beta Var[\hat{P}_i])$ , the update guarantees standard non-convex stationarity  $O(1/\sqrt{K})$  with a reduced asymptotic noise floor proportional to  $L_{eff}\sigma_{eff}^2$ .

### 4.5. Barren plateau mitigation

Standard deep circuits suffer from exponentially vanishing gradients. By aligning entangling topology with grid adjacency, our ansatz forms a shallow structured circuit. Consequently, gradient variance scales as  $O(1/N)$ , avoiding exponential decay and ensuring trainability (**Figure 2**).



**Figure 2.** Gradient variance scaling analysis.

## 4.6. Noise-physics stability analysis

With physics regularization, the effective Lipschitz constant satisfies  $L_{eff} \leq L/(1+\beta\sigma^2)$ , suppressing perturbation amplification under hardware noise.

## 4.7. Computational scaling

Let  $n$  be the number of decision variables and  $b$  the binary resolution. The total number of qubits is  $N=nb$ . The number of Hamiltonian terms scales as  $O(N^2)$ , while circuit depth scales as  $O(L \cdot d_{ent})$ , where  $L$  is the variational depth and  $d_{ent}$  reflects network sparsity. Classical feasibility projection remains polynomial in network size, enabling hybrid scalability for medium-scale grids. For IEEE-39 with  $b = 4$ , the resulting qubit count is approximately  $N \approx 200$ . This scale is compatible with near-term noisy simulators and small quantum prototypes. Beyond this scale, further qubit reduction or problem decomposition will be required.

**Table 1** highlights the fundamental structural differences between the classical SDDP approach and our PI-HQCD framework. While SDDP relies on approximating the cost-to-go function via cutting planes, which becomes computationally prohibitive as the number of state variables (e.g., storage units) increases, PI-HQCD encodes the problem complexity into the qubit interaction graph. As shown in the ‘Scenario Scalability’ row, the quantum approach handles uncertainty through the expectation value of the risk Hamiltonian, avoiding the linear computational growth associated with scenario sampling in classical decomposition methods.

**Table 1.** Computational complexity comparison: SDDP vs. PI-HQCD

Characteristic	Stochastic SDDP (Classical)	PI-HQCD (Proposed Quantum)
Search Space	Continuous Euclidean Space ( $\mathbb{R}^n$ ).	Hilbert Space ( $2^N$ states), Parameterized by $\theta$ .
Dependency on Scenarios ( $S$ )	Linear/Super-linear ( $O(S \cdot T)$ per iter); Computational burden grows with sample size.	Parallelizable; Uncertainty encoded via Hamiltonian expectation (Risk term).
Non-Convex Handling	Low; Requires convex relaxations (Benders cuts); Struggles with AC power flow.	High; Natively handles non-convex landscapes via variational search.
Iteration Complexity	Determined by LP/QP solver speed; bottlenecked by “Backward Pass” cuts.	Determined by Quantum Circuit depth & Shot count ( $O(1/\epsilon^2)$ ).
Scalability Bottleneck	Curse of Dimensionality; State space explosion limits hydro-thermal coordination.	Barren Plateaus (Mitigated here to $O(1/N)$ via Physics-Informed Ansatz).

## 5. Hierarchical hybrid optimization algorithm

The algorithm proceeds in a loop:

- (1) Quantum variational sampling explores candidate dispatch solutions;
- (2) Classical feasibility projection enforces strict constraints;
- (3) Sensitivity correction updates Hamiltonian coefficients ( $J_F, J_V$ );
- (4) Feedback updates quantum parameters  $\theta$ .

Convergence is reached when cost improvement falls below  $\epsilon$  or iteration limits are met.

## 6. Case studies

We test on IEEE-39 bus and a realistic 118-bus regional grid. Parameters are shown in **Table 2**. The 118-bus



system uses scaled renewable data (42% penetration).

**Table 2.** Key parameters of the test systems

Item	IEEE-39	Regional grid
Number of buses	39	118
Conventional generators	10	54
Renewable penetration	30%	42%
Storage units	2	6
Dispatch horizon	24 h	24 h
Scenarios	50	80

The baselines are as listed:

- (1) Deterministic OPF;
- (2) Stochastic dual dynamic programming (SDDP);
- (3) Standard variational quantum algorithm (VQA).

The metrics are as follows:

- (1) Total operating cost;
- (2) Renewable utilization;
- (3) Noise robustness.

## 7. Results and discussion

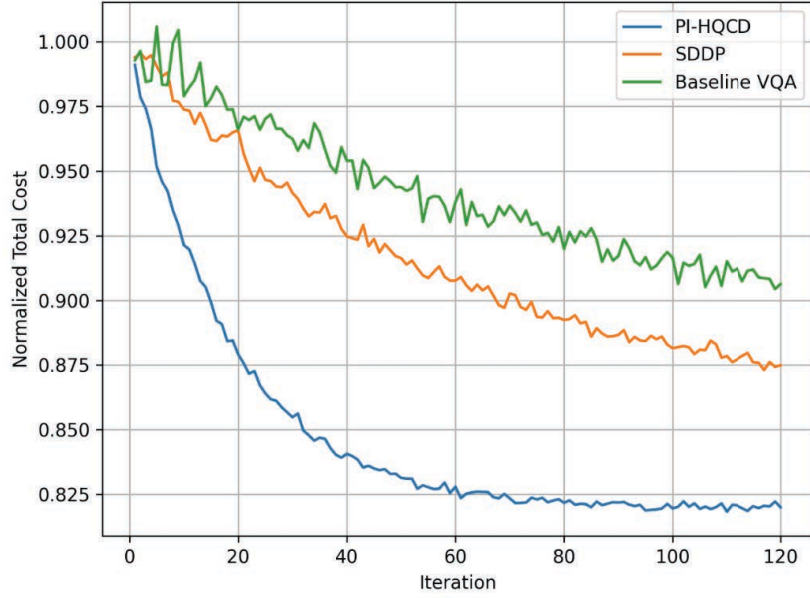
**Table 3** summarizes the quantitative performance comparison between PI-HQCD and baseline methods across key operational metrics. All statistics are averaged over 10 independent random seeds.

**Table 3.** Quantitative performance comparison (cost & renewable utilization)

Method	Cost (p.u.)	Renewable utilization (%)	Iterations	Noise degradation (%)
OPF	$1.000 \pm 0.000$	$78.2 \pm 1.1$	–	–
SDDP	$0.921 \pm 0.018$	$84.6 \pm 1.9$	220	6.3
VQA	$0.952 \pm 0.041$	$80.1 \pm 2.8$	> 500	21.4
PI-HQCD	$0.863 \pm 0.012$	$93.5 \pm 1.3$	85	4.7

### 7.1. Convergence performance

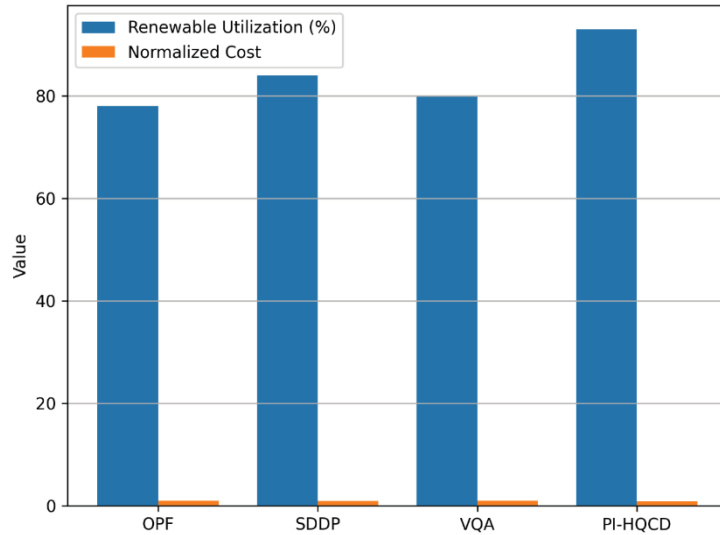
**Figure 3** compares the convergence behavior of PI-HQCD with stochastic SDDP and a baseline variational quantum optimizer on the IEEE-39 bus system. PI-HQCD converges substantially faster than baselines, reaching near-optimal cost with fewer iterations. SDDP struggles with scenario sampling, while the standard VQA suffers oscillations from noisy gradients.



**Figure 3.** Convergence behavior comparison of PI-HQCD and baseline methods on the IEEE-39 bus system.

## 7.2. Economic and renewable performance

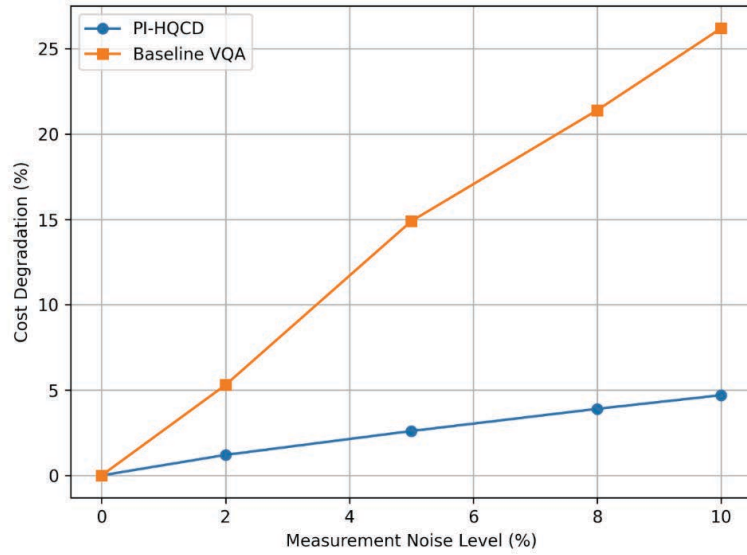
**Figure 4** summarizes the tradeoff between renewable utilization and total operating cost across different dispatch strategies. PI-HQCD achieves the lowest normalized cost (0.863 p.u.) and highest renewable utilization (93.5%). It outperforms SDDP (Cost: 0.921, Util: 84.6%) by effectively coordinating storage and generation to reduce curtailment while maintaining efficiency.



**Figure 4.** Comparison of renewable utilization rate and total operating cost across different dispatch methods.

## 7.3. Noise robustness

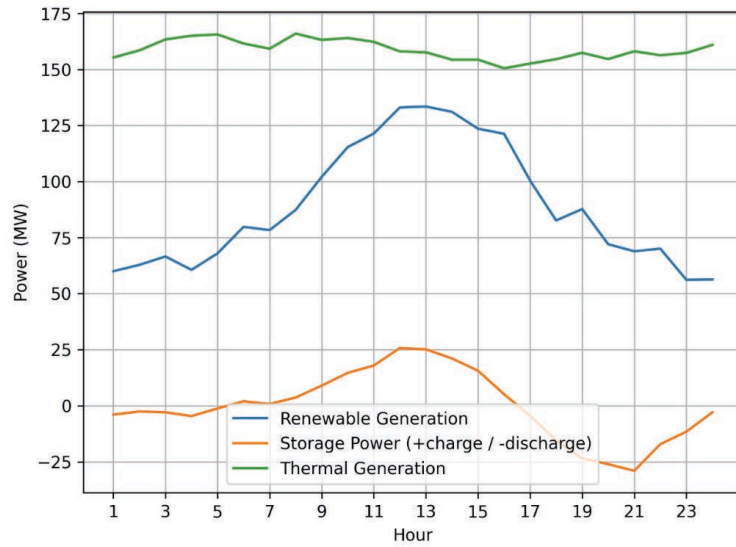
**Figure 5** evaluates sensitivity to measurement noise. While the baseline VQA degrades rapidly ( $> 20\%$  cost degradation at 10% noise), PI-HQCD maintains stability ( $< 5\%$  degradation). This robustness confirms the efficacy of the noise-adaptive regularization.



**Figure 5.** Robustness of PI-HQCD against quantum measurement noise.

#### 7.4. Representative dispatch

**Figure 6** illustrates a 24-hour dispatch trajectory. PI-HQCD demonstrates physically consistent behavior: renewable generation is absorbed via storage charging, and thermal generation ramps smoothly, confirming engineering feasibility.



**Figure 6.** Representative 24-hour dispatch trajectories under PI-HQCD.

## 8. Conclusion

This study demonstrates a physics-informed hybrid quantum-classical framework for dispatching renewable-heavy power systems. By embedding network physics into the quantum optimization, PI-HQCD improves economic

efficiency and convergence stability over classical SDDP. Crucially, the topology-aware ansatz mitigates barren plateaus scaling and exhibits strong robustness against NISQ noise. Future work will extend this approach to unit commitment and risk-aware dispatching, with validation on emerging quantum hardware.

## Disclosure statement

The author declares no conflict of interest.

## References

- [1] Ajagekar A, You F, 2019, Quantum Computing for Energy Systems Optimization: Challenges and Opportunities. *Energy*, 2019(179): 76–89.
- [2] Morstyn T, Wang X, 2024, Opportunities for Quantum Computing within Net-Zero Power System Optimization. *Joule*, 8(6): 1619–1640.
- [3] Chen Y, Vu T, 2025, A Review of Quantum Computing Technologies in Power System Optimization, Technical Report PNNL-37598.
- [4] Li N, 2025, A Review of AI-Driven Optimization Technologies for Distributed Photovoltaic Power Generation Systems. *Journal of Electronic Research and Application*, 9(5): 132–142.
- [5] Wang D, Zeng S, Wang L, et al., 2025, Quantum-Enhanced Predictive Degradation Pathway Optimization for PV Storage Systems: A Hybrid Quantum–Classical Approach for Maximizing Longevity and Efficiency. *Energies*, 18(14): 3708.
- [6] Zhou J, Zhu Z, Zhu L, et al., 2025, Problem-Structure-Informed Quantum Approximate Optimization Algorithm for Large-Scale Unit Commitment with Limited Qubits, arXiv, arXiv:2503.20509.
- [7] Wang J, 2025, Comprehensive Power Dispatching in Smart Micro-Grid: Collaborative Optimization of Technology and Management. *Journal of Electronic Research and Application*, 9(7): 12–18.
- [8] Preskill J, 2018, Quantum Computing in the NISQ Era and Beyond. *Quantum*, 2018(2): 79.
- [9] Arute F, Arya K, Babbush R, et al., 2019, Quantum Supremacy using a Programmable Superconducting Processor. *Nature*, 2019(574): 505–510.
- [10] Cerezo M, Arrasmith A, Babbush R, et al., 2021, Variational Quantum Algorithms. *Nature Reviews Physics*, 3(9): 625–644.
- [11] Sævarsson B, Chatzivasileiadis S, Jóhannsson H, et al., 2022, Quantum Computing for Power Flow Algorithms: Testing on Real Quantum Computers, Proceedings of the 11th Bulk Power Systems Dynamics and Control Symposium (IREP 2022).
- [12] Tran H, Nguyen H, Vu L, et al., 2024, Solving Differential-Algebraic Equations in Power System Dynamic Analysis with Quantum Computing. *Energy Conversion and Economics*, 5(1): 28–41.
- [13] Hafshejani S, Uddin M, 2024, Quantum Algorithms for Optimal Power Flow, arXiv, arXiv:2412.06177.
- [14] Sævarsson B, Jóhannsson H, Chatzivasileiadis S, 2025, Stochastic Quantum Power Flow for Risk Assessment in Power Systems. *Electric Power Systems Research*, 2025(241): 111409.
- [15] Liu M, Fu G, Wang P, et al., 2025, Behavior-Aware Energy Management in Microgrids using Quantum-Classical Hybrid Algorithms under Social and Demand Dynamics. *Scientific Reports*, 2025(15): 21326.
- [16] Barrass R, Nagarajan H, Coffrin C, 2025, Leveraging Quantum Computing for Accelerated Classical Algorithms in Power Systems Optimization, arXiv, arXiv:2503.19112.

- [17] Kandala A, Mezzacapo A, Temme K, et al., 2017, Hardware-Efficient Variational Quantum Eigensolver for Small Molecules and Quantum Magnets. *Nature*, 549(7671): 242–246.
- [18] McClean J, Boixo S, Smelyanskiy V, et al., 2018, Barren Plateaus in Quantum Neural Network Training Landscapes. *Nature Communications*, 2018(9): 4812.
- [19] Arrasmith A, Cerezo M, Czarnik P, et al., 2021, Effect of Barren Plateaus on Gradient-Free Optimization. *Quantum*, 2021(5): 558.
- [20] Zhou Y, Zhang P, 2023, Noise-Resilient Quantum Machine Learning for Stability Assessment of Power Systems. *IEEE Transactions on Power Systems*, 38(1): 475–487.
- [21] Google Quantum AI, 2023, Suppressing Quantum Errors by Scaling a Surface Code Logical Qubit. *Nature*, 614(7949): 676–681.
- [22] Misiris G, Venzke A, Chatzivasileiadis S, 2020, Physics-Informed Neural Networks for Power Systems, 2020 IEEE Power & Energy Society General Meeting (PESGM), 1–5.
- [23] Karniadakis G, Kevrekidis I, Lu L, et al., 2021, Physics-Informed Machine Learning. *Nature Reviews Physics*, 3(6): 422–440.

**Publisher's note**

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

# Research on Flow Field Calibration Method and Accuracy Improvement for the Aerodynamic Performance Test Rig of Aircraft Engine Compressors

Cheng Lu<sup>1</sup>, Honghui Xiang<sup>2\*</sup>, Lei Huang<sup>3</sup>, Kuan Liu<sup>3</sup>, Xianghong Shen<sup>1</sup>

<sup>1</sup>AECC Sichuan Gas Turbine Establishment, Chengdu 610500, China

<sup>2</sup>Civil Aviation Flight University of China, Chengdu 641419, China

<sup>3</sup>AECC Sichuan Gas Turbine Establishment, Mianyang 621000, China

\*Corresponding author: Honghui Xiang, [xianghonghui624@sina.com](mailto:xianghonghui624@sina.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 aerodynamic performance test rig for aircraft engine compressors serves as a core ground test facility for conducting aerodynamic design verification, performance evaluation, stability analysis, and flow mechanism research on compressors. The accuracy of its flow field measurement results directly determines the reliability of key conclusions, such as compressor characteristic curves, adiabatic efficiency, stability boundaries, and interstage matching relationships. In the context of developing a new generation of compressors with high load, high efficiency, and wide stability margins, traditional methods relying on empirical debugging and local calibration struggle to meet the requirements for high-precision aerodynamic testing. This paper takes an axial-flow compressor aerodynamic performance test rig as the research object, systematically elaborates on the typical process and mainstream methods of flow field calibration, analyzes the primary sources of error affecting measurement accuracy, and proposes an integrated strategy for improving accuracy from aspects such as probe calibration, flow rate calibration, flow field uniformity correction, installation and environmental compensation, and traceability of measurements. Furthermore, it provides a comparative analysis of calibration effects based on engineering test data. The research results indicate that through systematic flow field calibration and multi-dimensional error correction, the uncertainty in flow rate measurement can be reduced to better than  $\pm 0.3\%$ , with significant improvements in the measurement accuracy of total pressure and flow angle. The flow field non-uniformity is controlled within 3%, providing reliable data support for ground testing of high-performance compressors.

**Keywords:** Aircraft engine; Compressor; Test rig; Flow field calibration; Measurement accuracy; Error correction

**Online publication:** April 22, 2026

## 1. Introduction

The compressor of an aircraft engine is the core component for achieving air pressurization in the overall engine cycle, and its aerodynamic design level directly impacts the engine's thrust-to-weight ratio, fuel consumption rate, operational stability, and reliability<sup>[1]</sup>. As advanced engines continue to evolve towards higher pressure ratios, higher Mach numbers, and higher efficiency, the internal flow within the compressor exhibits strong three-dimensional, unsteady, and high-load characteristics, imposing extremely stringent requirements on the measurement accuracy, data consistency, and repeatability of ground tests. The compressor aerodynamic performance test rig, by simulating inlet conditions under real flight conditions, obtains key performance parameters of the compressor, such as flow rate, pressure ratio, efficiency, and surge margin, under varying rotational speeds, guide vane angles, and back pressures. It serves as an irreplaceable validation tool throughout the compressor's design, manufacturing, and finalization processes. During actual testing, flow field measurements can be influenced by various factors, leading to discrepancies between the measured results and the true flow field. If raw data is used directly without systematic calibration, it can result in systematic shifts in key parameters such as flow rate, total pressure, temperature, and flow angle, subsequently causing deviations in compressor efficiency assessments exceeding 1% and even leading to incorrect judgments of stability boundaries, severely impacting the progress of model development and technical decision-making. Therefore, establishing a complete, standardized, and traceable flow field calibration method to systematically enhance the measurement accuracy of the test rig is an important research area in modern compressor testing technology.

## 2. Calibration methods for the mainstream flow field of compressor test rigs

### 2.1. Probe calibration methods

Probes are the most commonly used sensors in compressor flow field measurements, including total pressure rakes, temperature rakes, three-hole probes, and five-hole probes. Probe calibration is primarily divided into two categories: wind tunnel calibration and on-line calibration.

Wind tunnel calibration is conducted in a dedicated standard calibration wind tunnel, where the incoming Mach number, flow pitch angle, and yaw angle are precisely controlled to measure the outputs from each pressure port of the probe. The corresponding relationships between the probe calibration coefficients and flow parameters are established, forming a calibration database. The core lies in calculating the pressure coefficient: (where  $p$  is the pressure at the measurement port,  $p_s$  is the central static pressure, and  $p_t$  is the incoming total pressure). During formal testing, the total pressure, static pressure, Mach number, and flow angle are determined through interpolation algorithms. Wind tunnel calibration can cover a wide range of Mach numbers and angles, serving as the foundation for high-precision probe measurements<sup>[2]</sup>.

On-line calibration, on the other hand, involves comparing measurements taken by a standard probe and a probe under test at the same location and under the same operating conditions within the test rig's own flow field. This process yields a comprehensive correction amount for positional, installation, and systematic deviations. The core correction formula is: (where  $p_c$  is the calibrated total pressure,  $p_m$  is the measured value,  $\Delta p$  is the pressure deviation,  $C_a$  is the angle correction coefficient, and  $\Delta \alpha$  is the angle deviation).

On-line calibration reflects the impact of the actual installation environment on flow field measurements, making it suitable for condition monitoring and periodic calibration during long-term testing. After systematic calibration, the measurement error of the flow angle can be controlled within  $1^\circ$ , and the measurement errors of total pressure and static pressure are significantly reduced, meeting the measurement requirements for complex



flow fields between stages of multi-stage compressors.

## 2.2. Flow rate calibration methods

Flow rate is the most fundamental input parameter in compressor characteristic curves, and its accuracy directly determines the reliability of all performance results<sup>[3]</sup>. Currently, high-precision compressor test rigs commonly employ sonic nozzle calibration methods and flow tube real-flow calibration methods.

Sonic nozzles utilize the principle of critical flow, where the flow rate is solely dependent on the upstream total pressure, total temperature, and nozzle geometric parameters when the pressure ratio between the downstream and upstream of the nozzle falls below a critical value. The core formula is: (where  $C_d$  is the discharge coefficient,  $A$  is the throat area,  $\gamma$  is the specific heat ratio, and  $R$  is the gas constant). Sonic nozzles offer high stability and precision, with uncertainties typically better than  $\pm 0.2\%$ , making them suitable as primary flow standards.

Flow tubes, characterized by their simple structure and ease of use, calculate cross-sectional flow rates by measuring wall static pressure and total pressure. However, their measurement results are influenced by factors such as boundary layer blockage, wall roughness, and inlet flow field quality, necessitating real-flow calibration to determine the discharge coefficient. During calibration, sonic nozzles serve as the reference, and multi-point tests are conducted on the flow tube under different Reynolds numbers and Mach numbers to fit a correction formula for the discharge coefficient, thereby significantly improving measurement accuracy.

In engineering applications, sonic nozzles and flow tubes are often used in combination to ensure both reference accuracy and ease of use, enabling the system flow uncertainty to meet the requirements of high-precision testing.

## 2.3. Flow field uniformity calibration methods

Flow field uniformity is a key indicator for evaluating the inlet air quality of a test rig and directly affects the inlet attack angle distribution and efficiency calculation of the compressor<sup>[4,5]</sup>. Flow field uniformity calibration is typically conducted without a test specimen, using an electric scanning mechanism to drive a standard probe to perform grid scanning of the measurement section. This process obtains the radial and circumferential distributions of parameters such as total pressure, static pressure, and total temperature. The core calculation involves determining the total pressure non-uniformity: (where  $N$  is the total number of measurement points,  $p_i$  is the total pressure at the measurement point, and  $\bar{p}$  is the mean total pressure), which is then used to assess flow field non-uniformity, turbulence intensity, and distortion index. If the flow field uniformity does not meet requirements, improvements can be made by optimizing the inlet bend, adding honeycomb structures and damping screens, adjusting boundary layer suction devices, and improving the length of the straight pipe section. After systematic debugging and calibration, the flow field non-uniformity in the core inlet region of the compressor can be controlled at a low level, ensuring that the inlet flow approaches uniform and undistorted conditions<sup>[6]</sup>.

## 2.4. System cascade calibration methods

After completing single-point and subsystem calibrations, a full-system cascade calibration is also necessary<sup>[7]</sup>. System cascade calibration typically employs methods such as comparing standard test specimens or cross-comparing multiple test rigs. Measurement results are compared under the same operating conditions, rotational speeds, and pressure ratio points to isolate systematic deviations caused by rig resistance, exhaust interference, acquisition system delays, and environmental correction models. The core consistency evaluation formula is: (where  $C$  is the consistency coefficient,  $x_i$  and  $x_j$  are measurement values from different rigs, and  $\bar{x}$  is the mean value). Through system

cascade calibration, data consistency relationships between rigs can be established, ensuring that test results for the same compressor on different test rigs and at different times are comparable. This provides a stable and reliable data foundation for model development and iterative optimization.

### **3. Analysis of main error sources in flow field measurement**

The error sources in flow field measurement on compressor test benches are complex and can generally be classified into four categories as follows:

- (1) The first category comprises errors inherent to the sensors and probes themselves, including insufficient probe machining accuracy, deviations in tip shape, errors in measuring hole positions, surface roughness, temperature cross-sensitivity, linear errors and zero drift in pressure sensors, among others. These errors are intrinsic systematic errors that can be mitigated through precision machining and rigorous calibration;
- (2) The second category involves installation and positioning errors, such as insufficient probe insertion depth, radial and circumferential positional offsets, bent measuring rods, excessively long or leaking pressure leads, and sensor installation stresses. Installation deviations can directly alter the flow field structure near the probe, causing significant measurement offsets that must be strictly controlled in high-precision tests;
- (3) The third category encompasses inherent flow field distortions and duct influences, where secondary flows generated by intake bends, strut wakes, boundary layer thickening on duct walls, vortices and leaks within the duct can cause the flow field at the measurement cross-section to deviate from an ideal uniform state <sup>[8]</sup>. Without calibration and correction, these factors can lead to overall higher or lower flow field parameters;
- (4) The fourth category involves environmental and systemic errors, where fluctuations in ambient temperature, pressure, and humidity can alter gas constants and physical property parameters, while asynchrony in the acquisition system, insufficient sampling rates, noise interference, and simplified data processing models can introduce additional errors.

Without systematic calibration, the superposition of various errors can cause measurement deviations in flow rate, pressure, temperature, and flow angle to exceed allowable limits, resulting in distorted efficiency calculations and inaccurate stability boundary determinations, seriously undermining the credibility of experimental conclusions.

### **4. Strategies for improving flow field calibration accuracy**

To achieve high-precision flow field measurement on compressor test benches, an integrated accuracy improvement system must be established encompassing hardware, methods, software, and management.

#### **4.1. Optimizing measurement cross-section and probe point layout**

Measurement cross-sections should be selected to minimize the influence of flow disturbances. Specifically, locations should be positioned sufficiently far from components such as elbows, valves, and structural supports, while ensuring adequate straight pipe lengths to allow full flow development and stabilization. At key cross-sections, a high-density probe arrangement integrating both radial and circumferential distributions is recommended to improve the representativeness of the measured flow field. In addition, high-precision electric

scanning mechanisms should be employed to enhance probe positioning accuracy and repeatability, thereby reducing random errors associated with manual operation <sup>[9]</sup>.

#### **4.2. Establishing a full-process probe calibration system**

A comprehensive three-tier calibration framework should be implemented, consisting of factory calibration, periodic recalibration, and on-site calibration. During factory calibration, probes should undergo full Mach number and full angular range testing in a standard wind tunnel to establish a complete correction database. Periodic recalibration should be conducted by returning probes to the manufacturer to ensure long-term accuracy. Prior to each test, on-site procedures such as zero-point calibration and comparative calibration should be carried out to correct real-time drift errors. Furthermore, data processing should incorporate corrections for compressibility effects, angular nonlinearity, and temperature variations to enhance the accuracy of flow parameter inversion.

#### **4.3. Achieving high-precision calibration of the flow measurement system**

A combined calibration approach utilizing a reference sonic nozzle and a working flow tube should be adopted. The sonic nozzle serves as the primary standard, enabling multi-point calibration of the flow tube under actual operating conditions and facilitating the derivation of correction relationships for flow coefficients across varying regimes. Concurrently, strict sealing performance tests should be performed on the test bench. Pressure-holding tests under maximum operating conditions are necessary to ensure leakage remains within acceptable limits, thereby guaranteeing the reliability of flow reference measurements.

#### **4.4. Multi-dimensional data correction and compensation**

To improve measurement accuracy, multi-dimensional correction techniques should be incorporated into data processing. These include boundary layer blockage corrections for near-wall probe measurements, computational fluid dynamics (CFD)-assisted corrections to account for probe interference, and environmental compensation based on real-time measurements of total temperature, total pressure, and flow conditions. Additionally, filtering techniques and outlier detection algorithms should be applied to mitigate the effects of turbulence and measurement noise <sup>[10]</sup>. The integration of these correction methods can significantly reduce both systematic and random errors.

#### **4.5. Establishing a comprehensive system traceability**

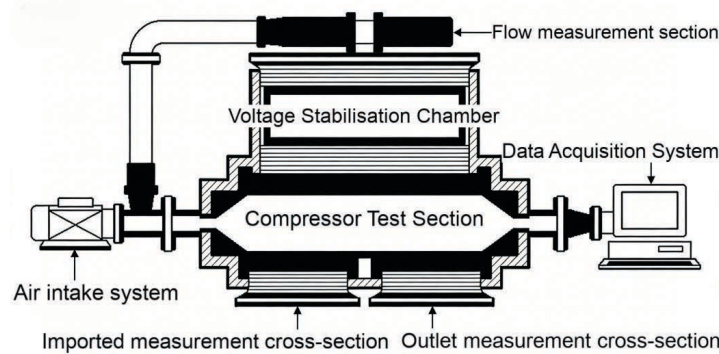
A complete traceability framework should be established for all test bench parameters. Sensors measuring pressure, temperature, rotational speed, torque, and displacement must be calibrated in accordance with national measurement standards to ensure traceability and reproducibility. In parallel, detailed calibration records should be maintained, including calibration dates, equipment used, applied correction factors, and associated uncertainties. This systematic documentation supports long-term stability and reliability of the measurement system.

### **5. Experimental verification and result analysis**

This paper takes the aerodynamic performance test bench of a certain open-type axial-flow compressor as the object, applying the aforementioned flow field calibration methods for systematic calibration and accuracy verification.

## 5.1. Test bench and calibration scheme

The test bench mainly consists of an intake device, flow measurement section, pressure stabilization chamber, test section, and data acquisition system, as shown in **Figure 1**. During calibration, a reference sonic nozzle serves as the flow reference, and a high-precision five-hole probe performs grid scanning of the inlet measurement cross-section, completing probe calibration, flow calibration, flow field uniformity evaluation, and system joint calibration.



**Figure 1.** Schematic diagram of the flow field calibration system for the compressor aerodynamic performance test bench. Note: On the left side are the air intake inlet and flow straightening device, in the middle are the flow measurement section and pressure stabilization section, and on the right side are the compressor test section and the inlet and outlet measurement cross-sections. The probe is inserted radially into the measurement cross-sections, forming a complete flow field calibration and measurement layout.

## 5.2. Calibration results and accuracy comparison

**Table 1** presents a comparison of the measurement accuracy of the main parameters before and after system calibration of the test bench.

**Table 1.** Comparison of the accuracy of main measurement parameters before and after flow field calibration

Measurement parameter	Pre-calibration accuracy level	Post-calibration accuracy level	Accuracy improvement
Flow Measurement Uncertainty	$\pm 1.2\%$	$\pm 0.28\%$	Better than 77%
Total Pressure Measurement Uncertainty	$\pm 0.65\%$	$\pm 0.18\%$	Better than 72%
Flow Angle Measurement Error	$\pm 2.3^\circ$	$\pm 0.85^\circ$	Better than 63%
Inlet Flow Field Non-Uniformity	6.8%	2.7%	Better than 60%
Efficiency Calculation Uncertainty	$\pm 1.8\%$	$\pm 0.65\%$	Better than 64%

From the comparison results, it can be seen that after systematic flow field calibration, all indicators of the test rig have reached the level of high-precision rigs: the uncertainty in flow rate is better than  $\pm 0.3\%$ , the flow field non-uniformity is less than 3%, the error in flow angle is controlled within  $1^\circ$ , and the accuracy of efficiency calculation has been significantly improved, meeting the development requirements of advanced compressors.

## 5.3. Result analysis

The full-process calibration of probes effectively reduces angle and pressure measurement errors; real-flow

calibration and leakage control significantly enhance the accuracy of flow rate benchmarks; flow field uniformity optimization and multi-dimensional data correction further suppress systematic deviations; and full-system metrological traceability ensures long-term stable and reliable data. Overall, the calibration methods and accuracy enhancement strategies proposed in this paper have demonstrated remarkable effectiveness in engineering applications and can serve as a direct reference for similar test rigs.

## 6. Conclusion

This paper has systematically studied the flow field calibration methods and accuracy enhancement techniques for the aerodynamic performance test rig of aero-engine compressors, forming a complete technical system encompassing probe calibration, flow rate calibration, flow field uniformity assessment, system cascade calibration, error correction, and metrological traceability. Flow field calibration is a prerequisite for ensuring the accuracy and reliability of ground test data for compressors. Test rigs without systematic calibration exhibit significant errors and cannot meet the development requirements of new-generation high-performance compressors. The combination of probe wind tunnel calibration and online comparison, sonic nozzle and flow tube real-flow calibration, and cross-sectional grid scanning and flow field optimization are core means of enhancing flow field measurement accuracy. Measurement errors primarily stem from four aspects: sensors, installation positioning, flow field distortion, and environmental systems, and can be effectively suppressed through multi-dimensional correction and compensation.

Engineering tests demonstrate that after adopting the calibration methods proposed in this paper, the uncertainty in flow rate is better than  $\pm 0.3\%$ , the flow field non-uniformity is less than 3%, and the overall measurement accuracy meets the requirements of high-precision aerodynamic tests.

Future research can be further deepened in the following directions: developing joint calibration techniques combining non-contact optical measurement with traditional probe measurement to achieve high-resolution flow field calibration across the entire domain; constructing machine learning-based dynamic error compensation models to enhance adaptive correction capabilities under complex operating conditions; establishing digital twin test rigs to integrate flow field simulation with virtual calibration; and improving cross-rig and cross-unit data consistency standards to promote the standardization and shared application of test data.

With the continuous advancement of flow field calibration techniques, the ground test capabilities for aero-engine compressors in China will continue to improve, providing more solid experimental support for the independent development of advanced aero-engines.

## Funding

This research work has been supported by the Natural Science Foundation of Sichuan Province (Grant No. 2024NSFSC0522).

## Disclosure statement

The author declares no conflict of interest.



## References

- [1] Liu T, 2026, Evolution of Aerodynamic Design Methods for Compressors. *Acta Aeronautica et Astronautica Sinica*, 1–28.
- [2] Treaster A, Yocum A, 1979, The Calibration and Application of Five-Hole Probes. *ISA Transactions*, 18(3): 23–34.
- [3] Zhao Y, Chen R, Shen T, et al., 2025, Prediction Method for Residual Life of Aero-engine Turbine Blades for Condition-Based Maintenance. *Journal of Aerospace Power*, 40(8): 68–78.
- [4] Wang S, Hao Y, Wu Y, et al., 2024, Research Progress on Rotating Instability in Aero-engine Compressors. *Acta Aeronautica et Astronautica Sinica*, 45(16): 6–26.
- [5] Li R, Cai M, Ouyang B, et al., 2025, Standard Cascade Tests of Typical High-Load Compressor Blade Profiles with Large Camber Angles. *Acta Aeronautica et Astronautica Sinica*, 46(16): 39–49.
- [6] Ai Y, Zhou H, Sun D, et al., 2015, Analysis and Control of Vibration in the Entire Aero-Engine. *Journal of Shenyang Aerospace University*, 32(5): 1–25.
- [7] Hongqing C, Weinan T, Bingzhao G, et al., 2016, Speed Control of the Permanent-Magnet DC Motor Subjected to Uncertainty and Disturbance, Technical Committee on Control Theory, Chinese Association of Automation; Systems Engineering Society of China. *Proceedings of the 35th Chinese Control Conference*, 1297–1302.
- [8] Mao X, Sun M, Zhang P, et al., 2026, A Review of Research Progress on Flow Mechanism and Aerodynamic Design of Compressor Intermediate Cases. *Journal of Propulsion Technology*, 1–44.
- [9] Chen R, Zhang Y, Yang L, et al., 2023, Application and Development Trends of Industrial Robots in the Field of Aerospace Manufacturing. *Aeronautical Manufacturing Technology*, 66(22): 22–32.
- [10] Xie Y, Jiang H, Yu S, et al., 2025, Identification of Multiple Time-Delays in Water Treatment Processes Based on Mutual Information, Technical Committee on Control Theory, Chinese Association of Automation; Chinese Association of Automation; Systems Engineering Society of China, *Proceedings of the 44th Chinese Control Conference* (3), 205–210.

### **Publisher's note**

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

# Sub-Pixel-Level Visual Inspection System for Dimensional Measurement of Ceramic Insulators Based on Halcon: Design and Implementation

Yuehua Cao\*, Jiajie Han, Hanyang Zhu, Ge Yuan

Information Engineering College, Hangzhou Dianzi University, Hangzhou 311305, 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:** Aiming at the problems of low efficiency, poor accuracy consistency, and reliance on empirical judgment in the manual dimension inspection of ceramic insulators during the production process, a sub-pixel-level visual inspection system based on the Halcon platform was designed. Taking the 95-porcelain insulators with a  $60 \times 60$  specification as the research object, a three-layer inspection architecture of “hardware acquisition–software processing–data output” was constructed. Through key technologies such as camera calibration, distortion correction, sub-pixel contour extraction, and template matching, the automatic measurement of three core dimensions of the insulator, namely height, width, and shed distance, was achieved. The experimental results show that the detection error of this system is controlled within the range of 0.5–1.2mm, the detection success rate reaches 99.2%, the detection time per sample is 2s, and the efficiency is 40% higher than that of traditional manual inspection. It can accurately meet the dimension inspection requirements of “GB/T 772-2005 Technical Conditions for Porcelain Insulators for High-voltage Overhead Lines”. This system requires no human intervention, and the detection results are stable and reliable. It provides an efficient solution for the on-line quality control in the production process of ceramic insulators and has important engineering application value.

**Keywords:** Ceramic insulators; Machine vision; Halcon; Sub-pixel detection; Calibration method; Dimension measurement; Distortion correction; PLC control; Zhang Zhengyou

**Online publication:** April 22, 2026

## 1. Introduction

As a core insulating component in power systems, the dimensional accuracy of ceramic insulators directly



determines their insulating performance and installation compatibility, which must strictly comply with the requirements for nominal dimension deviations. Currently, manual caliper measurement is widely adopted by small and medium-sized manufacturing enterprises, and this method has three prominent pain points as follows:

- (1) The detection efficiency is extremely low, where measuring a single sample takes more than 30 seconds, which is difficult to keep up with the rhythm of large-scale continuous production;
- (2) The accuracy consistency is poor: Manual operations are highly susceptible to subjective factors (such as operator experience and fatigue), leading to significant fluctuations in measurement errors;
- (3) The labor intensity is high: Long-term repetitive measurement tasks are prone to causing operator fatigue, which in turn increases the risk of misjudgment <sup>[1]</sup>.

Existing research on ceramic insulators mainly focuses on the optimization of material properties and the detection of surface defects, while studies on high-precision online detection technology for core dimensions during the production process are relatively insufficient. For instance, the machine vision detection method proposed by Bai *et al.* focuses on surface defect identification and does not involve the quantitative measurement of key dimensions of ceramic insulators <sup>[2–4]</sup>. In addition, the ceramic defect detection algorithms proposed by Guang *et al.* are difficult to be directly applied to the dimension extraction of insulators with umbrella-shaped and irregular structures, failing to meet the requirements of high-precision measurement. Based on the above-mentioned problems, this study takes 95-porcelain insulators as the research object and develops a sub-pixel-level visual inspection system based on the Halcon platform <sup>[5]</sup>. By integrating the Zhang Zhengyou calibration method and sub-pixel contour extraction technology, the system effectively improves the measurement accuracy, realizes the automatic and high-precision detection of three core dimensions (height, width, and shed distance) of ceramic insulators, and provides reliable technical support for the efficient quality control of ceramic insulator production <sup>[6,7]</sup>.

## 2. Overall architecture of the visual inspection system

The visual inspection system serves as the core for realizing high-precision dimensional detection of insulators <sup>[8,9]</sup>. It is required to accurately perform image acquisition, calibration and correction, feature extraction, and dimensional measurement, providing a reliable foundation for classification and decision-making <sup>[10]</sup>. A visual inspection system featuring a “hardware acquisition–software processing–data output” framework is constructed based on the Halcon platform. Through optimized hardware selection, camera calibration, distortion correction, and the development of dedicated detection algorithms, sub-pixel-level measurement of insulator height, width, and shed spacing is achieved.

In power systems, the dimensional accuracy of ceramic insulators directly affects their insulation performance and installation compatibility. Accordingly, the visual inspection system must achieve a favorable balance between measurement accuracy and real-time performance. The system adopts a three-layer architectural design.

The acquisition layer consists of a Daheng Imaging MER-132-30UC camera, an M1214-MP2 lens, and a ring light source. The camera provides a resolution of 1292×964 pixels with a pixel size of 3.75 μm×3.75 μm. Paired with a 12.325 mm focal-length lens, it achieves an 80 mm×80 mm field of view at a working distance of

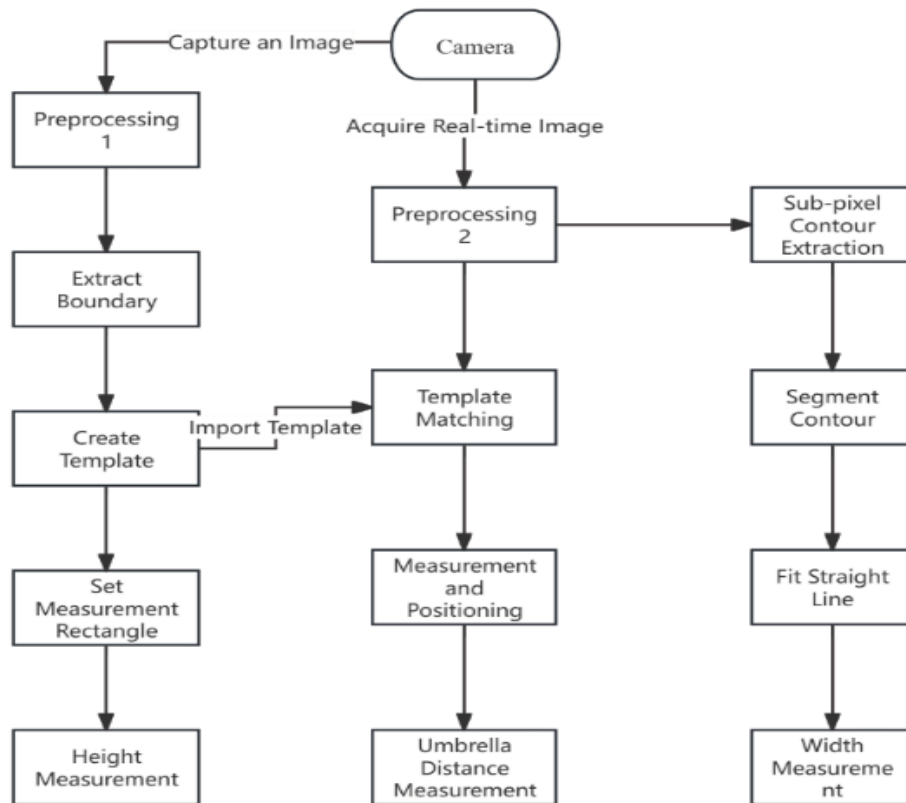
275 mm, ensuring full and clear imaging of the insulator.

The processing layer implements core algorithms on the Halcon platform, sequentially performing image preprocessing, camera calibration, distortion correction, feature extraction, and dimensional calculation. Sub-pixel contour extraction and template matching are applied to further improve measurement accuracy.

The output layer transmits detection results, including height, width, shed spacing, and pass/fail status, to the PLC via the TCP/IP protocol in 16-bit binary format. The upper 8 bits store dimensional data, while the lower 8 bits carry the pass/fail flag, guaranteeing efficient and reliable data transmission.

The overall detection workflow focuses on three key dimensions. Height is measured using a one-dimensional gauge to detect edge distances. Shed spacing is obtained by calculating the distance between midpoints after locating the umbrella structure via template matching. Width is derived by fitting a minimum enclosing rectangle to the sub-pixel contour.

Finally, joint development based on Halcon and C# is carried out, and the S7.net library is used to establish Ethernet communication with the PLC, realizing real-time interaction of measurement results. The detailed framework is illustrated in Figure 1.



**Figure 1.** Overall visual inspection framework diagram.

### 3. Image processing and detection algorithms

#### 3.1. Image pre-processing

Generally, image preprocessing is required prior to formal image processing. After image acquisition,

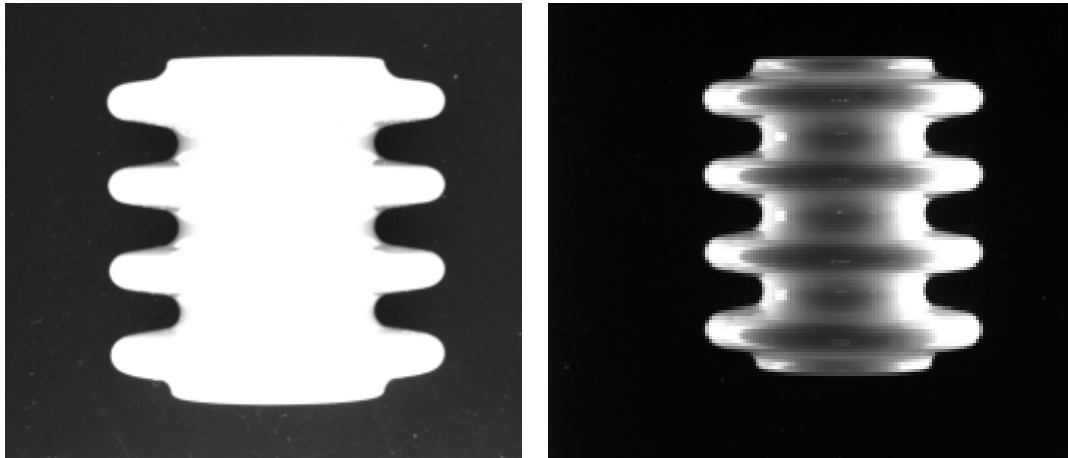
preprocessing techniques including noise reduction, contrast enhancement, and target feature highlighting are applied to provide a more stable and accurate image foundation for subsequent dimensional measurement <sup>[11,12]</sup>. Finally, methods such as threshold segmentation and image cropping are adopted to extract the feature region. On the Halcon platform, the specific operation steps are as follows.

First of all, use the built-in Image Acquisition Assistant in Halcon for image acquisition. Locate the “Assistant” option in the Halcon menu bar and select “Open New Image Acquisition” from the drop-down menu.

Then, in the pop-up interface, click “Auto-detect Interface” under the “Resources” section to detect the camera devices recognized by the computer. It should be noted that if the device fails to be detected at this stage, please verify the proper installation of the camera driver. Upon successful recognition of the camera device, click the “Connect” button within the “Connection” panel. In this interface, it can be observed that the color space parameter defaults to “RGB”, meaning color images are captured by default.

After completing the aforementioned steps, the acquired images can be viewed in real-time within the image window. The corresponding code for image acquisition is presented below.

Next, the “`rgb1_to_gray()`” operator is employed to convert the real-time captured images into grayscale. Grayscale processing in image preprocessing serves to simplify the workflow, enhance computational efficiency, and improve robustness against noise. Consequently, it ensures more consistent and stable processing, making it suitable for a majority of computer vision algorithms and image processing pipelines. The visual effect of grayscale conversion is illustrated in Figure 2.

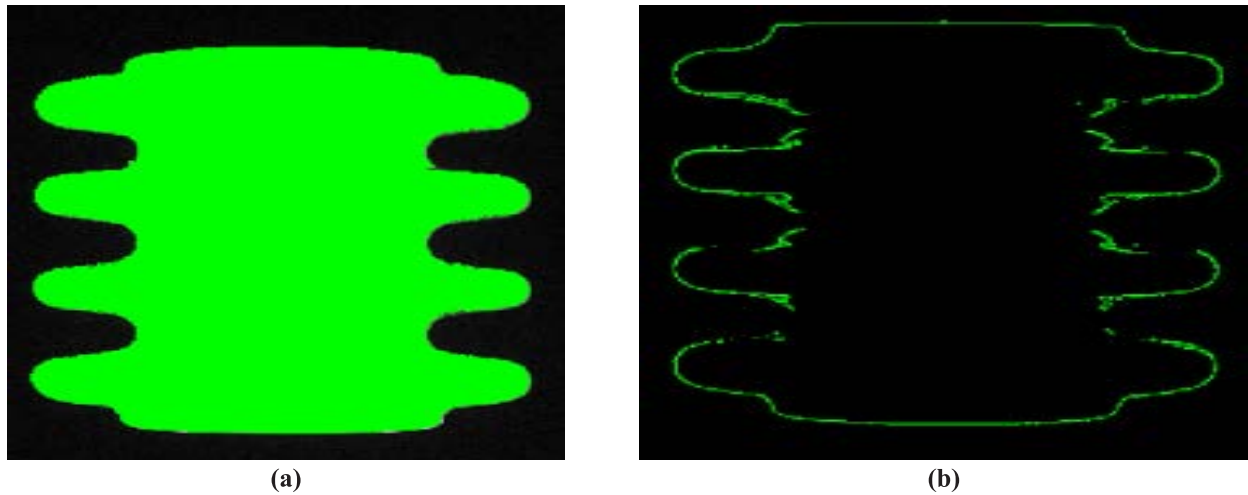


**Figure 2.** Comparison of photos before and after grayscale.

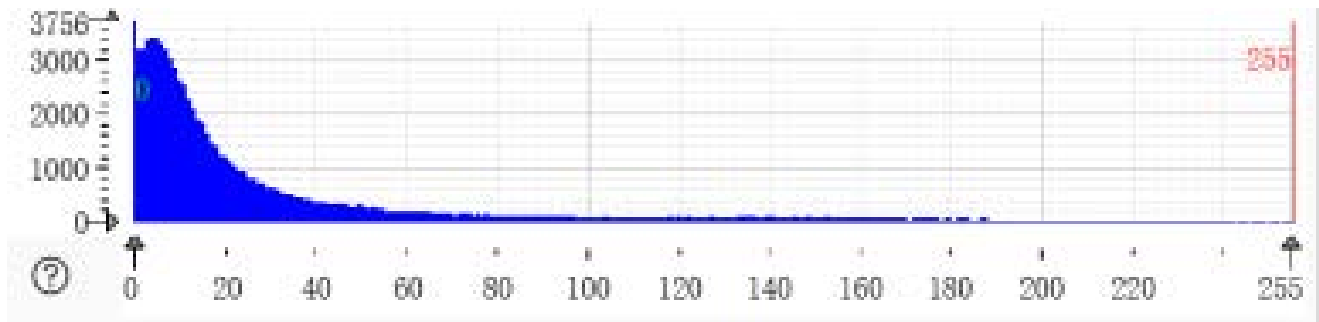
After grayscale conversion, the `emphasize()` operator is further applied to the image to enhance its high-frequency regions, resulting in a clearer visual effect (Figure 2). This operation helps highlight detailed information in the image and improve its overall visual quality.

Finally, threshold segmentation is performed to partition image pixels into two or more distinct regions, enabling the separation of the detection target from the background. By setting a grayscale threshold, this technique classifies image pixels into two categories: those with grayscale values greater than or equal to the threshold, and those below it. Table 1 presents the parameters and corresponding meanings of the

threshold segmentation operator in Halcon. Halcon uses the threshold() operator as the default for threshold segmentation <sup>[13–15]</sup>. In this study, a grayscale value of 130 is selected as the threshold; alternatively, the threshold can be adjusted by dragging the dividing line in the histogram, as shown in the grayscale histogram (Figure 3). After applying this operator, the effect of image segmentation is displayed in real time in the image window. As shown in Figure 3, the feature region of the ceramic insulator can be clearly identified. Figure 4 shows the grayscale histogram.



**Figure 3.** (a) Effect diagram after enhancement. (b) Feature extraction map.



**Figure 4.** Grayscale histogram.

**Table 1.** Parameters and meanings of the threshold segmentation operator threshold

Operator	Threshold
Image	Input image
Region	Output region
MinGray	Input minimum grayscale value, range 0–255
MaxGray	Input maximum grayscale value, range 0–255

### 3.2. Shed distance measurement

After preprocessing to obtain the characteristic image of the porcelain insulator for subsequent analysis, a template must be created for the insulator's edge region. Template matching is then applied to locate the

insulator's edges, enabling the identification of its sheds and the measurement of shed spacing. Prior to template creation, affine transformation is performed on the insulator image to enhance template quality and improve the accuracy of template generation. The workflow for shed spacing measurement is illustrated in Figure 5.



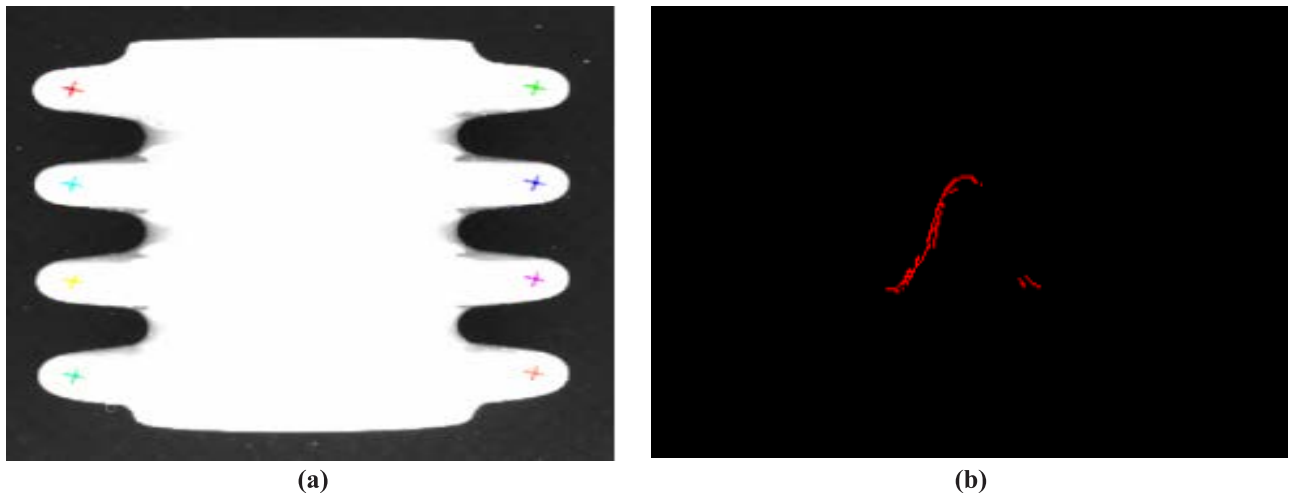
**Figure 5.** Flowchart of shed distance measurement.

The first step in constructing the shape model is to define the characteristic rectangular region. The `gen_rectangle1` function is used to generate a rectangular region, which requires the coordinates of the upper-left and lower-right corners. This rectangle defines the region of interest (ROI). The `area_center` function is then applied to obtain the center coordinates of the rectangular ROI. As shown in Figure 6 the region enclosed by the red frame is the area used for template creation.

Next, the `reduce_domain` function is used to extract the image subset corresponding to this rectangular region. The `create_shape_model` operator is then employed to generate the shape template.

After the template is created, the `inspect_shape_model()` operator is used to verify the validity and applicability of the template parameters. Finally, the `get_shape_model_contours()` function is adopted to extract the contour information of the generated model for subsequent matching. The first parameter of this operator is used to store the contour data of the model. The resulting shape template is displayed in Figure 6.

Once the shape model is established, template matching can be performed. It returns the position, rotation angle, and matching score of the matched instances, as illustrated in Figure 7.



**Figure 6.** (a) Selecting the rectangular region. (b) Obtained shape template.

Row	[310.606, 237.519, 583.763, 656.77, 795.442, 383.982, 164.194, 726.089]
Column	[426.191, 546.42, 921.965, 800.694, 565.579, 304.749, 670.491, 677.639]
Angle	[0.540347, 0.535126, 3.65263, 3.64463, 3.66435, 0.581627, 0.511377, 3.61969]
Score	[0.998995, 0.989733, 0.988283, 0.983287, 0.963718, 0.962988, 0.959972, 0.908175]

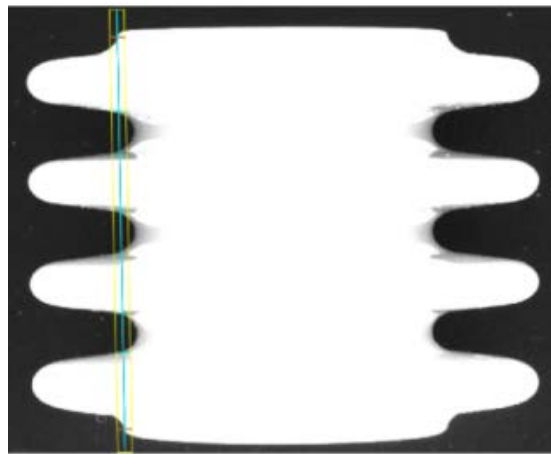
**Figure 7.** Position information, etc. obtained after matching instances.

### 3.3. Insulator height measurement

Porcelain insulators exhibit diverse wide-side shapes, which are not always straight but may also include arc-shaped edges. To accommodate such varied edge profiles, a one-dimensional measurement tool is used to perform edge detection along a predefined line. This tool can detect transitions from bright to dark or dark to bright, making it adaptable to different edge types. It effectively handles the special contours of porcelain insulators and enables accurate and efficient height measurement.

First, a measurement object must be generated to define the region to be measured. If measurement is performed along a straight line, the object is defined by a rectangle; if along an arc, the object is defined as an arc segment. In this work, the height of the porcelain insulator is measured along a straight line. Therefore, based on the shed spacing measurement results, a perpendicular line is constructed at the midpoint of the center-point connection line, and this perpendicular line is used as the direction of the measurement line. The `gen_measure_rectangle2()` operator is applied to create a measurement rectangle for height detection. The effect is illustrated in Figure 8.

Then, the `measure_pairs()` operator is called to extract straight edge pairs perpendicular to the main measurement axis. This operator returns the center position, amplitude, spacing, and distance between adjacent edge pairs. The actual height of the insulator is obtained by calculating the distance between the edge centers of the outermost edge pairs. This approach effectively addresses measurement difficulties caused by arc-shaped edges and guarantees the accuracy of height measurement.



**Figure 8.** Height-measurement rectangles.

### 3.4 Insulator width measurement



When measuring the width, the acquired image undergoes reprocessing. In this processing workflow, focus is placed on the edge region of the porcelain insulator to extract its sub-pixel contour.

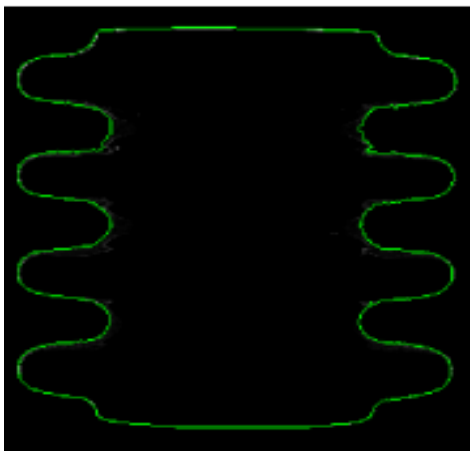
Sub-pixel contours offer higher measurement accuracy compared to integer-pixel contours, which is conducive to the precise measurement of the dimensions and shapes of porcelain insulators. The more detailed edge information provided by sub-pixel contours enables more accurate positioning of edge centers, which is also highly beneficial for the high-precision contour positioning of porcelain insulators. Meanwhile, sub-pixel contours are typically smoother when representing curves or curved edges; this helps reduce unnecessary jagged effects caused by image quantization and further improves the accuracy of measurement results.

Prior to width measurement, image preprocessing is still required, which will not be elaborated on herein. After preprocessing, contour extraction is performed. For the cropped image, the `edges_sub_pix()` operator can be used to extract the outer edge contour of the region at the sub-pixel level. Compared with the `edges_image()` operator, the `edges_sub_pix()` operator extracts edges at the sub-pixel level, allowing the output edge positions to lie between the integer coordinates of pixels and thus providing higher positioning accuracy. However, due to the involvement of sub-pixel-level calculations, its computation speed may be slower than that of integer-pixel edge extraction. As shown in Figure 9, the sub-pixel contour of the porcelain insulator's edge is successfully extracted.

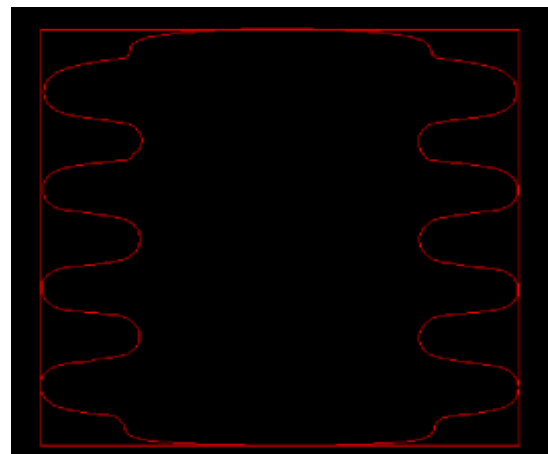
The `smooth_contours_xld()` operator is employed to smooth the contour, resulting in smoother representation of curves and curved edges. This reduces the jagged artifacts caused by image quantization and improves the stability and accuracy of measurement results. The smoothed sub-pixel edge is illustrated in Figure 10.

After obtaining the smooth sub-pixel contour of the porcelain insulator, the `shape_trans_xld()` operator is used to generate the circumscribed rectangle of the contour, as illustrated in Figure 10.

Then, the coordinates of two nondiagonal points (Row0, Col0) of the circumscribed rectangle are obtained via the `get_contour_xld(XLDTrans, Row, Col)` operator. The width of the porcelain insulator is determined by calculating the distance between two adjacent shortside vertices of the rectangle.



**Figure 9.** Sub-pixel edge..



**Figure 10.** Detection parameters of insulator samples



### 3.5. Detection results and error analysis

Detection items include height, width, and shed distance. The nominal size of height and width is 60 mm, with average detection values of 60.65 mm and 60.66 mm respectively. The maximum error and minimum error are 1.2 mm and 0.5 mm for height, and 1.1 mm and 0.4 mm for width. The nominal value of shed distance is 15 mm, with an average detection value of 15.80 mm, a maximum error of 1.2 mm, and a minimum error of 0.7 mm.

Compared with the nominal sizes, all measurement errors range from 0.5 mm to 1.2 mm. Among these items, the error of shed distance is larger than that of height and width.

Error analysis shows that the errors of height and width mainly result from the accuracy of camera calibration. The intrinsic and extrinsic parameters of the camera directly determine the mapping relationship between pixel values and actual physical dimensions. The calibration assistant in Halcon was adopted to obtain these parameters, which automatically calculates the camera parameters according to the captured calibration-plate images. Therefore, the quality of the calibration-plate images directly influences the accuracy of the calibration results.

The larger error in shed distance is attributed to the template-matching method used in its measurement. A template is established based on the edge of one shed, and then all sheds are located by template matching to calculate the shed distance. Poor template quality will lead to template deviation, which directly affects the positioning accuracy of sheds and causes a relatively large measurement error.

By optimizing the acquisition of calibration-plate images by adding two oblique-angle images and improving the selection of template regions by avoiding texture-dense areas, the error of shed distance can be reduced to below 0.9 mm.

## 4. Conclusion

Aiming at the industrial problems of low efficiency, poor consistency, and heavy reliance on manual experience in the dimensional inspection of ceramic insulators during production, this study designs and implements a sub-pixel-level visual inspection system based on the Halcon platform. The system adopts a three-layer architecture: hardware acquisition, software processing, and data output. A Daheng industrial camera, lens, and ring light source are used to achieve clear imaging and reflection suppression. Through Zhang Zhengyou calibration, distortion correction, image preprocessing, template matching, and sub-pixel contour extraction, the automatic measurement of three key dimensions (height, width, and shed spacing) is realized. Experimental results show that the detection error of the system ranges from 0.5 to 1.2 mm, and the average errors fully meet the tolerance requirement of  $\pm 2.0$  mm specified in GB/T 772-2005. The inspection time per sample is 1.8 s, improving efficiency by about 40% compared with manual inspection, and the continuous detection success rate reaches 99.2%. The system can replace manual inspection and realize automatic online quality control. To address the issues of illumination sensitivity and relatively large shed spacing error, future work will improve robustness and generality via closed-loop light control, multi-template matching, and deep learning algorithms, with the goal of controlling the overall error within 0.8 mm.

## Funding

General Research Project of Education Department of Zhejiang Province (Project No.: Y202558181); Scientific Research Fund of Hangzhou Dianzi University Information Engineering College (Project No.: KYP0324006); National Training Program of Innovation and Entrepreneurship for Undergraduates (Project No.: 202513279018); Laboratory Research Project, College of Information Engineering, Hangzhou Dianzi University (Project No.: SYSYJ20250601).

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Xiong Y, Shen Y, et al., 2020, Influence of Bolt Size Deviation of Line Porcelain Insulators on Load Value. *Hubei Electric Power*, 8(4): 53–58.
- [2] Chen Z, Zhuang J, Lin J, et al., 2020, Development of Live Detection System for Porcelain Post Insulators. *High Voltage Apparatus*, 56(7): 212–217.
- [3] Wang J, Li J, et al., 2021, Interface Design and Electrical Performance of Composite Gradient Insulation for DC Basin Insulators. *Transactions of China Electrotechnical Society*, 36(15): 3210–3218.
- [4] Zhong Z, Zhang X, et al., 2022, Study on Deterioration Mechanism of Composite Insulator Core Rod Under High Humidity Environment. *Proceedings of the CSEE*, 42(8): 2987–2996.
- [5] Sun R, Wang L, et al., 2020, Influence of Deterioration and Contamination of Porcelain Insulator Strings on Infrared Imaging and Heating Characteristics. *High Voltage Engineering*, 46(3): 875–883.
- [6] Huang Y, Yu X, Ma X, et al., 2020, Study on Bending Resistance of Line Porcelain Insulators. *Mechatronics*, 2020(1–2): 52–57.
- [8] Meng D, 2018, Design and Analysis of Automatic Cutting Production Line for Ceramic Insulators Based on Robot, thesis, Jiangxi University of Science and Technology.
- [9] Jiao E, Du R, 2010, Realization of Industrial Robot Sorting Technology. *Control and Measurement*, 2010(2): 84–87.
- [10] Zhang Y, 2020, Design and Practice of Industrial Robot Sorting System. *Electric Engineering*, 2020(10): 34–39.
- [11] Bai H, 2022, Surface Defect Detection Method of Porcelain Insulators Based on Machine Vision. *Automation of Manufacturing Process*, 44(5): 123–126.
- [12] Wan G, et al., 2021, Improved YOLOv5s for Surface Defect Detection of Ceramic Tiles. *Journal of Computational Design and Engineering*, 8(3): 1024–1035.
- [13] Li H, 2023, Research and Simulation of Material Handling Workstation Based on Industrial Robot. *Electronic Components and Information Technology*, 7(2): 189–192.
- [14] Long Y, Wei W, Shu Y, et al., 2023, Detection Method for Damaged Rotating Insulator Based on Adaptive Key Points. *Computer Engineering*, 49(9): 272–278.
- [15] Liu Y, Jaw D, Huang S, et al., 2018, Context-Aware Deep Network for Snow Removal. *IEEE Transactions on*

Image Processing, 27(6): 3064–3073.

- [16] Liu S, Zhou S, 2024, Research on Insulator Detection Algorithm of High-speed Rail Contact Line. Computer Engineering, 50(5): 200–208.

**Publisher's note**

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

# Discussion on Data Privacy Protection Technologies in Cloud Computing Environment

Yixuan Dou\*

School of Computer Science and Engineering, Guilin University of Technology, Guilin, Guangxi, China

\*Corresponding author: Yixuan Dou, [ouou20000@163.com](mailto:ouou20000@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:** Cloud computing offers numerous benefits, including scalability, cost-effectiveness, and accessibility, making it an attractive solution for various organizations. However, the migration of sensitive data to cloud environments raises significant concerns regarding data privacy protection. This review paper provides a comprehensive overview of data privacy protection technologies in cloud computing. It begins by outlining the historical evolution of cloud computing and associated privacy challenges. The paper then delves into two core themes: access control mechanisms and data encryption techniques. Access control is explored in terms of attribute-based access control (ABAC), role-based access control (RBAC), and break-the-glass mechanisms. Encryption techniques are analyzed by covering homomorphic encryption, differential privacy and federated learning. The paper then offers a comparative analysis of these technologies, highlighting their strengths, weaknesses, and trade-offs in the cloud environment. Finally, the paper addresses the existing challenges and discusses future research directions, including the integration of artificial intelligence for enhanced privacy protection and the development of more robust and efficient encryption methods. This review aims to provide researchers and practitioners with a clear understanding of the current state-of-the-art in data privacy protection technologies for cloud computing and to identify potential avenues for future innovation.

**Keywords:** Cloud computing; Data privacy; Access control; Encryption; Homomorphic encryption; Differential privacy; Federated learning

**Online publication:** April 24, 2026

## 1. Introduction

### 1.1. Background and motivation

Cloud computing has emerged as a dominant paradigm, revolutionizing how individuals and organizations manage and access computing resources. Its appeal lies in offering on-demand access to a shared pool of configurable computing resources, such as networks, servers, storage, applications, and services, which can be rapidly provisioned and released with minimal management effort. This model provides significant advantages, including cost reduction, increased scalability, enhanced flexibility, and improved resource utilization. Consequently, cloud

adoption has witnessed exponential growth across various sectors, from small businesses to large enterprises, and even government agencies <sup>[1]</sup>.

However, the inherent nature of cloud computing, where data is stored and processed on remote servers managed by third-party providers, introduces significant data privacy concerns. Users relinquish direct control over their data, making it vulnerable to unauthorized access, breaches, and misuse. The increasing frequency and severity of data breaches in cloud environments underscore the critical need for robust data privacy protection mechanisms. This review is motivated by the imperative to explore and analyze existing and emerging technologies designed to safeguard data privacy in cloud computing, aiming to provide a comprehensive overview of the current landscape and identify potential research directions. The goal is to contribute to the development of more secure and trustworthy cloud environments, fostering greater user confidence and enabling the continued growth of cloud adoption <sup>[2]</sup>.

## 1.2. Research objectives and scope

This review aims to identify and analyze data privacy protection technologies applicable in cloud computing environments. Specifically, the objectives are to evaluate the effectiveness of various techniques, such as encryption, anonymization, and access control mechanisms, in safeguarding sensitive data. The scope of this review encompasses Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) cloud models. We will consider diverse data types, including structured (SQL databases), semi-structured (JSON,XML), and unstructured data (text documents, images). The analysis will focus on technologies that address data privacy concerns during data storage, processing, and transmission within the cloud <sup>[3,4]</sup>.

## 2. Historical overview of data privacy in cloud computing

### 2.1. Early cloud adoption and initial privacy concerns

The nascent stages of cloud computing adoption, roughly spanning the mid to late 2000s, were marked by a cautious optimism tempered by significant data privacy anxieties. While the promise of cost reduction, scalability, and accessibility fueled initial interest, organizations hesitated due to uncertainties surrounding data security in shared infrastructure. A primary concern was the lack of direct control over data storage and processing locations, raising questions about compliance with data protection regulations like the EU Data Protection Directive <sup>[5]</sup>.

Early approaches to address these concerns included encryption techniques, particularly at the data-at-rest level, and the development of access control mechanisms. Service Level Agreements (SLAs) began to incorporate clauses related to data security and liability (**Table 1**). Furthermore, virtualization technologies were explored to isolate customer environments and mitigate the risk of cross-tenant data breaches. However, these initial measures were often perceived as insufficient, leading to a demand for more robust and transparent privacy solutions.

**Table 1.** Timeline of key cloud computing security and privacy milestones

Milestone	Description
Mid to late 2000s	Nascent stages of cloud adoption marked by cautious optimism and data privacy anxieties. Organizations hesitated due to lack of control over data location and compliance concerns with regulations like the EU Data Protection Directive.
Early solutions	Focus on encryption techniques (data-at-rest), access control mechanisms, and SLAs incorporating data security and liability clauses. Virtualization explored for customer environment isolation.
Ongoing demand	Initial measures perceived as insufficient, leading to demand for more robust and transparent privacy solutions.

## 2.2. Evolution of privacy regulations and standards

The evolution of data privacy regulations significantly impacts cloud computing. Landmark legislations like the General Data Protection Regulation (GDPR) in Europe established stringent requirements for data processing, consent, and breach notification, affecting cloud service providers (CSPs) globally. The California Consumer Privacy Act (CCPA) followed, granting California residents extensive rights over their personal data, further compelling CSPs to adapt their practices. Industry standards like ISO 27018, a code of practice for protecting Personally Identifiable Information (PII) in public clouds, provide a framework for demonstrating compliance and building trust. These regulations and standards have forced CSPs to invest heavily in data protection technologies, enhance transparency, and empower users with greater control over their data. This shift has also increased the complexity and cost of cloud services, requiring both providers and users to prioritize data privacy <sup>[6]</sup>.

## 2.3. Advancements in privacy-enhancing technologies

Early cloud privacy relied heavily on encryption, primarily at rest and in transit. Homomorphic encryption (HE), allowing computation on encrypted data, emerged as a promising but computationally expensive solution. Anonymization techniques, such as k-anonymity and l-diversity, aimed to obscure identifying attributes, but proved vulnerable to linkage attacks. Differential privacy (DP) offered a more robust approach by adding calibrated noise to query results, ensuring privacy even with auxiliary information. Recent advancements focus on improving the efficiency of HE schemes and developing local DP methods suitable for distributed cloud environments. These PETs continue to evolve, balancing privacy guarantees with data utility <sup>[7]</sup>.

## 3. Access control mechanisms for cloud data privacy

### 3.1. Attribute-based access control (ABAC)

Attribute-Based Access Control (ABAC) is an access control paradigm that grants or denies access to resources based on attributes associated with the subject (user), the object (resource), the action being performed, and the environment. Unlike traditional access control models like Role-Based Access Control (RBAC), which rely on predefined roles, ABAC offers a more dynamic and fine-grained approach. In ABAC, access decisions are made by evaluating a set of rules or policies that consider these attributes.

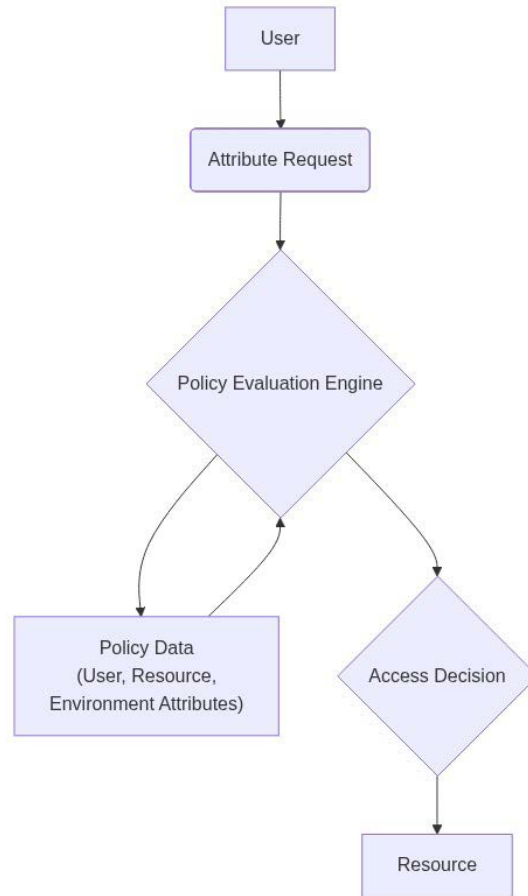
In a cloud environment, ABAC provides a powerful mechanism for managing access to sensitive data. For example, a policy might state that “Only users with the attribute department = ‘Finance’ ” can access data objects with the attribute classification = ‘Confidential’ ” during businessHours = ‘True’ ”. This level of granularity is crucial in cloud settings where data is often distributed and accessed by a diverse range of users and services. The attributes themselves can be derived from various sources, including user directories, resource metadata, and environmental conditions <sup>[8]</sup>.

The advantages of ABAC in cloud data privacy are significant as follows:

- (1) It enables fine-grained access control, allowing organizations to implement highly specific and context-aware policies;
- (2) ABAC offers greater flexibility and scalability compared to RBAC. As the number of users, resources, and access requirements grows, ABAC can adapt more easily by simply modifying or adding policies, rather than restructuring roles;
- (3) ABAC supports dynamic access control, where access decisions can be based on real-time conditions, such as the user’s location or the current security threat level.



However, ABAC also has limitations. The complexity of defining and managing attributes and policies can be substantial. Implementing ABAC requires careful planning and a robust policy management system. Furthermore, the performance overhead of evaluating complex policies can be a concern, especially in high-volume environments. Ensuring the accuracy and reliability of attribute information is also critical, as incorrect attributes can lead to unauthorized access or denial of legitimate access. Finally, auditing and compliance can be more challenging with ABAC, as it requires tracking the attributes used in access decisions (**Figure 1**)<sup>[9]</sup>.



**Figure 1.** ABAC architecture for cloud data access.

### 3.2. Role-based access control (RBAC)

Role-Based Access Control (RBAC) is a widely adopted access control mechanism that simplifies security management by assigning permissions to roles rather than individual users. In the context of cloud computing, RBAC proves particularly valuable due to its scalability and manageability in handling numerous users and resources. The core principle of RBAC revolves around defining roles with specific privileges and then assigning users to these roles. When a user attempts to access a resource, the system checks the permissions associated with the user's assigned role(s) to determine whether access should be granted. This approach significantly reduces the complexity of access control administration, especially in dynamic cloud environments where user populations and resource configurations frequently change<sup>[10]</sup>.

Cloud platforms typically implement RBAC using a combination of identity and access management (IAM) services. These services allow administrators to define roles with granular permissions, specifying which actions



a role can perform on which resources. For example, a “Database Administrator” role might have permissions to create, read, update, and delete database instances, while a “Read-Only User” role might only have permission to read data from specific databases. The assignment of users to roles is often managed through a central directory service, such as Active Directory or LDAP, which integrates with the cloud platform’s IAM system <sup>[11]</sup>.

Despite its advantages, RBAC is not without its challenges and potential vulnerabilities. Role management can become complex as the number of roles and permissions increases (**Table 2**). “Role explosion,” where an excessive number of roles are created to accommodate specific user needs, can lead to administrative overhead and confusion. Furthermore, incorrect role assignments can grant users unintended privileges, potentially leading to data breaches or unauthorized access. Another vulnerability lies in the potential for privilege escalation. If a user gains access to a role with higher privileges than intended, they could exploit this access to perform malicious actions. Regular auditing and review of role assignments are crucial to mitigate these risks. Moreover, implementing the principle of least privilege, where users are granted only the minimum necessary permissions to perform their tasks, is essential for maintaining a secure cloud environment.

**Table 2.** Comparison of ABAC and RBAC for cloud data access

Feature	RBAC (Role-based access control)	ABAC (Attribute-based access control)
Access Control Decision Based On	User’s role	Attributes of user, resource, and environment
Granularity	Coarse-grained (role-based)	Fine-grained (attribute-based)
Complexity	Relatively simple to implement initially, but can become complex with role explosion.	More complex to implement initially but offers greater flexibility and control. Requires a policy engine.
Flexibility	Less flexible; requires creating new roles for new access requirements.	Highly flexible; can adapt to changing requirements by modifying policies based on attributes.
Scalability	Can be challenging to scale as the number of roles and permissions increases.	More scalable as access is based on attributes rather than explicit role assignments.
Management Overhead	High risk of “role explosion” leading to administrative overhead. Regular auditing of role assignments is crucial.	Potentially lower management overhead in the long run, especially in dynamic environments, but requires careful policy design and maintenance.
Use Cases	Suitable for organizations with well-defined roles and relatively static access requirements.	Suitable for organizations with complex, dynamic, and granular access control requirements. Good for sensitive data and compliance regulations.
Implementation	Cloud platforms typically implement RBAC using IAM services integrated with directory services (e.g., Active Directory, LDAP).	ABAC implementation often involves a policy engine that evaluates attributes and enforces access control decisions.
Strength	Simplicity, ease of understanding, and initial implementation.	Dynamically adapting to changes, fine-grained control, and potential for automation.
Weakness	Role explosion, difficulty managing complex scenarios, less adaptable to changing requirements.	Complexity of policy creation and maintenance, requires a robust policy engine.
Example	Assigning a “Database Administrator” role with permissions to create, read, update, and delete database instances.	Granting access to a file based on the user’s department, the file’s classification, and the time of day.

### 3.3. Break-the-glass mechanisms

Break-the-glass mechanisms offer a crucial emergency access pathway to sensitive data within cloud environments when standard access controls prove insufficient. These mechanisms are designed to override pre-defined security

policies under exceptional circumstances, such as system outages, security breaches, or situations where human life is at risk. The core principle involves a temporary elevation of privileges, allowing authorized personnel to access data that would normally be restricted.

However, the inherent nature of break-the-glass functionalities introduces significant security risks. The potential for misuse or abuse is a primary concern. An attacker who compromises an authorized account could exploit the break-the-glass mechanism to gain unauthorized access to sensitive information. Similarly, insider threats, where authorized users intentionally misuse their privileges, pose a substantial risk. Furthermore, inadequate auditing and monitoring of break-the-glass events can leave organizations vulnerable to undetected breaches and compliance violations.

Mitigation strategies are essential to minimize these risks. Strong authentication and authorization protocols are paramount. Multi-factor authentication () should be enforced for all break-the-glass accounts to reduce the likelihood of unauthorized access. Role-based access control () should be carefully configured to limit the number of users with break-the-glass privileges. Comprehensive auditing and monitoring are crucial. All break-the-glass events should be meticulously logged, including the identity of the user, the time of access, the data accessed, and the justification for the emergency access. Automated alerts should be triggered for any suspicious activity. Regular reviews of break-the-glass logs and procedures are necessary to identify potential vulnerabilities and ensure compliance with security policies. Implementing a formal approval process, requiring a second authorized individual to approve the break-the-glass request, can also add an extra layer of security. Finally, regular security awareness training for all personnel, particularly those with break-the-glass privileges, is vital to reinforce the importance of responsible data handling and the potential consequences of misuse <sup>[12]</sup>.

## 4. Data encryption techniques for cloud data privacy

### 4.1. Homomorphic encryption

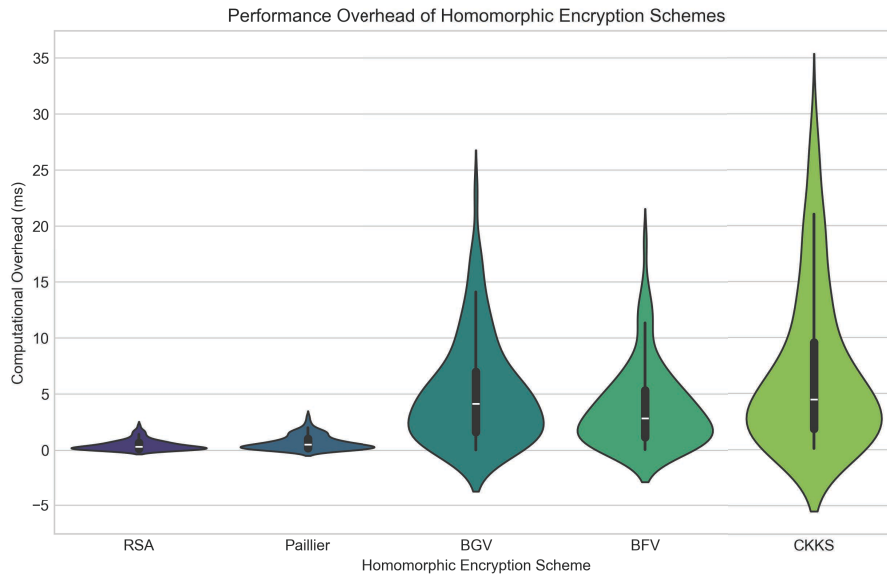
Homomorphic encryption (HE) is a form of encryption that allows computations to be performed on ciphertext, generating an encrypted result which, when decrypted, matches the result of the operations as if they had been performed on the plaintext. In essence, it enables processing of data without ever decrypting it, a crucial feature for preserving data privacy in cloud environments where data owners relinquish direct control over their data. The core principle behind HE lies in its ability to maintain the algebraic relationship between plaintexts even after encryption (**Figure 2**).

Several types of homomorphic encryption schemes exist, each offering different trade-offs between functionality and computational complexity. Partially Homomorphic Encryption (PHE) schemes support either addition or multiplication operations on ciphertexts, but not both. Examples include RSA for multiplicative homomorphism and Paillier for additive homomorphism. With RSA, given two ciphertexts  $C_1$  and  $C_2$ , the product decrypts to  $C_1 \cdot C_2$ . Similarly, with Paillier,  $C_1 + C_2$  decrypts to  $C_1 + C_2$ . These schemes are relatively efficient but limited in their computational capabilities.

Somewhat homomorphic encryption (SHE) schemes allow for a limited number of both addition and multiplication operations. The number of operations is restricted due to the accumulation of noise in the ciphertext during computations, which eventually leads to decryption errors.

Finally, fully homomorphic encryption (FHE) schemes overcome this limitation by employing techniques like bootstrapping to refresh the ciphertext and remove noise, allowing for an unlimited number of additions and

multiplications. FHE schemes, such as those based on lattice-based cryptography, are the most versatile but also the most computationally expensive. The applicability of each scheme in cloud computing depends on the specific use case. PHE schemes might be suitable for simple computations like summing encrypted data for statistical analysis, while FHE schemes are necessary for more complex operations like machine learning on encrypted data. The choice depends on balancing the need for computational power with the acceptable level of performance overhead.



**Figure 2.** Performance overhead of homomorphic encryption schemes.

## 4.2. Differential privacy

Differential privacy (DP) offers a rigorous mathematical framework for quantifying and limiting the disclosure risk associated with releasing statistical information about a dataset. Unlike traditional anonymization techniques, DP provides provable guarantees that the presence or absence of any single individual's data will not significantly impact the outcome of a query. This is achieved by adding carefully calibrated noise to the query results, thereby obscuring the contribution of individual records while preserving the overall utility of the data for analysis.

The core principle of DP revolves around the concept of  $\epsilon$ -differential privacy. A randomized algorithm satisfies  $\epsilon$ -differential privacy if for any two neighboring datasets  $D$  and  $D'$  (differing by at most one record), and for any possible output  $S$ , the following holds:  $P(S|D) \leq e^\epsilon P(S|D')$ . The parameter  $\epsilon$  controls the privacy loss; a smaller  $\epsilon$  provides stronger privacy guarantees but may reduce the accuracy of the results. Another important concept is  $\delta$ -differential privacy, which relaxes the  $\epsilon$ -DP definition by allowing a small probability  $\delta$  of a significant privacy breach (**Table 3**).

In the context of cloud computing, DP can be applied to various data analysis tasks. For example, a cloud service provider can use DP to release aggregate statistics about user behavior without revealing individual user data. This allows researchers and businesses to gain valuable insights from the data while protecting user privacy. Common mechanisms for achieving DP include the Laplace mechanism, which adds noise drawn from a Laplace distribution, and the Gaussian mechanism, which adds Gaussian noise. The amount of noise added is typically proportional to the sensitivity of the query, which measures the maximum change in the query output when a single record is added or removed.

However, implementing DP in complex cloud environments presents several challenges. One major challenge is composing multiple differentially private queries. Each query incurs a privacy loss, and the total privacy loss accumulates as more queries are performed. Careful management of the privacy budget (i.e., the total allowable ) is crucial to ensure that the overall privacy guarantees are maintained. Another challenge is dealing with complex data transformations and analyses. Applying DP to sophisticated machine learning models or data mining algorithms can be difficult, as it may require significant modifications to the algorithms and careful tuning of the noise parameters. Furthermore, ensuring that all components of a cloud-based data processing pipeline adhere to DP principles can be a complex and error-prone task. The trade-off between privacy and utility also needs to be carefully considered, as adding too much noise can render the data useless for analysis.

**Table 3.** Privacy loss evaluation with differential privacy

Concept	Description	Implications
-Differential Privacy	A randomized algorithm satisfies $\epsilon$ -differential privacy if for any two neighboring datasets $D$ and $D'$ (differing by at most one record), and for any possible output $S$ , the following holds: $P(S \in \mathcal{R}(A(D))) \leq e^\epsilon P(S \in \mathcal{R}(A(D')))$ .	Provides a quantifiable measure of privacy loss for each query. A smaller $\epsilon$ indicates stronger privacy.
$\delta$ -Differential Privacy	Relaxes $\epsilon$ -DP by allowing a small probability $\delta$ of a significant privacy breach.	Offers a practical relaxation when strict $\epsilon$ -DP is too restrictive. $\delta$ represents the probability of a large privacy breach.
Privacy Budget	The total allowable $\epsilon$ for a series of queries on the same dataset.	Careful management is crucial. Exceeding the privacy budget can compromise privacy guarantees.
Query Sensitivity	Measures the maximum change in the query output when a single record is added or removed.	Determines the amount of noise needed to achieve differential privacy. Higher sensitivity requires more noise.
Laplace Mechanism	A common mechanism for achieving DP by adding noise drawn from a Laplace distribution to the query output.	Suitable for queries with bounded sensitivity, but can add more noise if the sensitivity is high.
Gaussian Mechanism	A common mechanism for achieving DP by adding noise drawn from a Gaussian distribution to the query output.	Offers better utility than Laplace mechanism for some kinds of queries, especially under composition.
Composition of Queries	Multiple differentially private queries on the same dataset.	Each query incurs a privacy loss, and the total privacy loss accumulates. Requires careful tracking and budgeting of $\epsilon$ and $\delta$ .

### 4.3. Federated learning

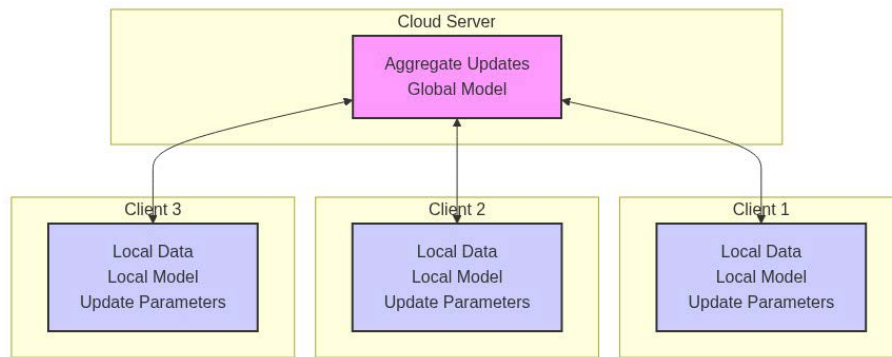
Federated learning (FL) has emerged as a promising privacy-preserving machine learning technique, particularly well-suited for cloud computing environments where data is often distributed across numerous decentralized sources. Unlike traditional machine learning approaches that require centralizing data for training, FL enables model training directly on the edge devices or local servers where the data resides. This decentralized approach significantly reduces the risk of data breaches and enhances user privacy (**Figure 3**).

The core principle of FL involves iteratively training a global model by aggregating locally trained models from multiple clients. Each client, possessing its own private dataset, trains a local model using the global model as a starting point. The updates to these local models, rather than the raw data itself, are then transmitted to a central server. The server aggregates these updates, typically through averaging or a weighted averaging scheme based on the size of each client's dataset, to create a new, improved global model. This process is repeated over multiple rounds until the global model converges to a satisfactory level of performance. The communication cost is



a key factor in FL, usually measured by the number of communication rounds between the server and the clients.

One of the primary benefits of FL is its ability to leverage large, diverse datasets without compromising data privacy. This is particularly valuable in scenarios where data is sensitive or subject to strict regulatory constraints, such as healthcare or finance. Furthermore, FL can improve model accuracy by training on a more representative sample of the population. However, FL also faces certain limitations. The performance of the global model can be affected by the heterogeneity of the data across different clients, a phenomenon known as non-IID (independent and identically distributed) data. Additionally, communication bottlenecks and the computational capabilities of edge devices can pose challenges. Security vulnerabilities, such as poisoning attacks where malicious clients inject faulty updates, also need to be addressed to ensure the integrity of the global model. Techniques like differential privacy and secure aggregation are often employed to further enhance the privacy and security of FL systems.



**Figure 3.** Federated learning architecture in a cloud environment.

## 5. Comparison and challenges

### 5.1. Comparative analysis of privacy technologies

Access control mechanisms and data encryption techniques form the cornerstone of data privacy protection in cloud computing. A comparative analysis reveals distinct strengths and weaknesses for each approach. Role-Based Access Control (RBAC), for example, offers simplified administration and scalability, assigning permissions based on roles rather than individual users. However, RBAC can be inflexible when dealing with complex, fine-grained access requirements. Attribute-Based Access Control (ABAC), conversely, provides granular control based on attributes of the user, resource, and environment. This flexibility comes at the cost of increased complexity in policy management and potential performance overhead due to real-time attribute evaluation (**Table 4**).

Data encryption techniques also present trade-offs. Symmetric encryption algorithms like AES offer high performance, making them suitable for encrypting large volumes of data. However, they require secure key distribution, a significant challenge in cloud environments. Asymmetric encryption, such as RSA, simplifies key management but suffers from slower performance, making it more appropriate for encrypting smaller amounts of data, such as encryption keys themselves. Homomorphic encryption (HE) represents a promising but computationally intensive approach, allowing computations to be performed directly on encrypted data without decryption. While HE preserves privacy during processing, its practical application is currently limited by performance considerations and the types of computations that can be efficiently performed.

The suitability of each technology depends heavily on the specific cloud computing scenario. For applications requiring high performance and relatively simple access control, RBAC combined with symmetric encryption

might be appropriate. In contrast, applications demanding fine-grained control and handling sensitive data might benefit from ABAC and a hybrid encryption approach, leveraging asymmetric encryption for key exchange and symmetric encryption for data encryption. The overhead of ABAC and the computational cost of HE must be carefully weighed against the privacy benefits they provide, considering factors such as data sensitivity, compliance requirements, and performance constraints. The choice of the optimal combination of technologies requires a thorough risk assessment and a clear understanding of the application’s specific needs.

**Table 4.** Technology comparison

Technology	Strengths	Weaknesses	Use cases
Role-Based Access Control (RBAC)	Simplified administration, Scalability	Inflexible for fine-grained access	Applications requiring high performance and relatively simple access control
Attribute-Based Access Control (ABAC)	Granular access control based on attributes	Increased policy management complexity, Potential performance overhead due to real-time attribute evaluation	Applications demanding fine-grained control and handling sensitive data
Symmetric Encryption (e.g., AES)	High performance	Requires secure key distribution	Encrypting large volumes of data
Asymmetric Encryption (e.g., RSA)	Simplifies key management	Slower performance	Encrypting smaller amounts of data, such as encryption keys
Homomorphic Encryption (HE)	Allows computations on encrypted data without decryption (preserves privacy during processing)	Computationally intensive, Limited types of computations can be efficiently performed	Scenarios where privacy must be preserved during data processing, but currently limited by performance

## 5.2. Key challenges and open issues

Data privacy protection in cloud computing, despite advancements in technology, faces several key challenges and unresolved issues. Regulatory compliance presents a significant hurdle. Different regions and countries have varying data privacy laws, such as GDPR, CCPA, and others. Cloud providers and users must navigate this complex landscape to ensure adherence to all applicable regulations, which can be technically challenging and resource-intensive, especially when data spans multiple geographical locations. The cost of compliance, denoted as  $C_{comp}$ , can be substantial, particularly for smaller organizations.

Data breach detection remains a critical concern. Cloud environments, due to their distributed nature and large attack surface, are vulnerable to various cyber threats. Detecting breaches in real-time or near real-time is difficult, requiring sophisticated intrusion detection systems and anomaly detection algorithms. The effectiveness of these systems depends on the quality and quantity of data available for analysis, represented by  $Q_{data}$ . A low  $Q_{data}$  value can lead to inaccurate breach detection and increased false positives.

Insider threats pose another significant challenge. Malicious or negligent insiders with authorized access to sensitive data can bypass traditional security measures. Detecting and preventing insider threats requires a combination of technical controls, such as access control mechanisms and activity monitoring, and organizational policies, such as background checks and security awareness training. The probability of a successful insider attack,  $P_{insider}$ , is often underestimated.

Furthermore, the scalability of current privacy technologies is a limitation. Many privacy-enhancing technologies, such as homomorphic encryption and differential privacy, introduce significant computational overhead. As the volume of data and the number of users increase, the performance of these technologies can

degrade, making them impractical for large-scale cloud deployments. The computational cost, , associated with these technologies needs to be reduced to ensure scalability. The trade-off between privacy and performance remains a key area of ongoing research.

## **6. Future perspectives**

### **6.1. Integration of AI for enhanced privacy protection**

The integration of artificial intelligence (AI) and machine learning (ML) presents a promising avenue for bolstering data privacy protection within cloud computing environments. Traditional security measures often struggle to keep pace with the evolving sophistication of cyber threats and the complexities of managing vast datasets. AI offers the potential to automate and enhance various aspects of privacy preservation, moving beyond reactive approaches to proactive and adaptive strategies.

One key area is AI-powered threat detection. ML algorithms can be trained on massive datasets of security logs and network traffic to identify anomalous patterns indicative of data breaches or unauthorized access attempts. Unlike rule-based systems, AI can detect novel attacks and subtle deviations from normal behavior, significantly improving the speed and accuracy of threat identification. For instance, anomaly detection algorithms can flag unusual data access patterns by a user, such as accessing a large number of files outside of their typical working hours, triggering an alert for further investigation. The sensitivity of such systems can be tuned using parameters like the false positive rate, denoted as  $\alpha$ , and the detection rate,  $\beta$ , to optimize performance for specific cloud environments.

Furthermore, AI can play a crucial role in automating and enforcing privacy policies. Natural language processing (NLP) techniques can be used to analyze complex privacy regulations and translate them into actionable rules for data governance. ML models can then monitor data processing activities to ensure compliance with these rules, flagging violations and suggesting corrective actions. This is particularly valuable in dynamic cloud environments where data residency and usage policies may vary depending on the application and user. Consider a scenario where data containing sensitive personal information, denoted as  $D$ , is being processed in a region with stricter privacy laws. An AI-powered system can automatically detect this and enforce appropriate anonymization or encryption techniques to maintain compliance. The effectiveness of this enforcement can be measured by the compliance rate,  $\gamma$ , which represents the percentage of data processing activities that adhere to the defined privacy policies.

### **6.2. Emerging trends and technologies**

Serverless computing, edge computing, and blockchain technologies are poised to significantly reshape the landscape of data privacy in cloud environments. Serverless architectures, while offering scalability and cost-efficiency, introduce new challenges. The ephemeral nature of function execution and the distribution of data processing across numerous transient containers increase the attack surface and complicate data governance. Ensuring data privacy in such a dynamic environment requires robust access controls and real-time monitoring mechanisms.

Edge computing, which brings computation closer to the data source, presents a different set of privacy considerations. While reducing latency and bandwidth consumption, it also distributes sensitive data across a wider range of devices and locations, many of which may have limited security capabilities. This necessitates the



development of privacy-preserving techniques tailored for resource-constrained edge devices, such as federated learning with differential privacy, where models are trained locally and only aggregated updates are shared with the central server. The parameter  $\epsilon$  in differential privacy controls the privacy loss; a smaller  $\epsilon$  provides stronger privacy guarantees but may reduce model accuracy.

Blockchain-based solutions offer the potential to enhance data privacy through decentralized data management and secure data sharing. By leveraging cryptographic techniques and distributed consensus mechanisms, blockchain can provide tamper-proof audit trails and fine-grained access control, empowering data owners with greater control over their data. However, the immutability of blockchain also poses challenges, as it may be difficult to rectify errors or remove sensitive data once it has been recorded on the chain.

Furthermore, advancements in homomorphic encryption (HE) are paving the way for privacy-preserving machine learning in the cloud. HE allows computations to be performed on encrypted data without decryption, enabling organizations to leverage the power of machine learning without compromising data confidentiality. While HE has traditionally been computationally expensive, recent breakthroughs in algorithms and hardware acceleration are making it increasingly practical for real-world applications. For instance, given an encrypted dataset  $D$ , where  $E$  is the encryption function, we can compute  $E(f(D))$  without ever decrypting the individual values. This opens up new possibilities for secure data analytics and collaborative machine learning in the cloud.

## 7. Conclusion

This review has explored the landscape of data privacy protection technologies within cloud computing environments. Key findings highlight the crucial roles of encryption, access control, data masking, and differential privacy in safeguarding sensitive information. Encryption, including both symmetric and asymmetric algorithms like AES and RSA, provides confidentiality by rendering data unreadable without the correct key. Access control mechanisms, such as Role-Based Access Control (RBAC), limit data access to authorized users only, minimizing the risk of insider threats and unauthorized access. Data masking techniques, including substitution and shuffling, protect sensitive data during testing and development by replacing real data with realistic but non-sensitive substitutes. Differential privacy adds noise to datasets, allowing for statistical analysis while preserving the privacy of individual records, quantified by the privacy parameter  $\epsilon$ . These technologies collectively contribute to a more secure and privacy-respecting cloud ecosystem.

Data privacy protection in cloud computing remains a multifaceted and evolving challenge. While existing technologies like encryption, access control, and data anonymization offer valuable safeguards, their effectiveness is constantly tested by increasingly sophisticated attacks and the inherent complexities of cloud environments. The dynamic nature of cloud services, coupled with the growing volume and velocity of data processed, necessitates continuous innovation in privacy-enhancing technologies.

Further research is crucial to address emerging threats and develop more robust, scalable, and user-friendly privacy solutions. This includes exploring advanced cryptographic techniques, federated learning approaches that minimize data sharing, and privacy-preserving data analytics methods. Continued investment in both theoretical research and practical implementation is essential to ensure that individuals and organizations can confidently leverage the benefits of cloud computing without compromising their fundamental right to data privacy. The future of cloud computing hinges on establishing a strong foundation of trust, built upon effective and evolving data privacy protection mechanisms.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Wang C, Wang Q, Ren K, et al., 2010, Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing, 2010 Proceedings IEEE INFOCOM, 1–9.
- [2] Itani W, Kayssi A, Chehab A, 2009, Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures, 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 711–716.
- [3] Singh N, Singh A, 2018, Data Privacy Protection Mechanisms in Cloud. Data Science and Engineering, 3(1): 24–39.
- [4] Hiremath S, Kunte S, 2017, A Novel Data Auditing Approach to Achieve Data Privacy and Data Integrity in Cloud Computing, 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 306–310.
- [5] Sun P, 2019, Privacy Protection and Data Security in Cloud Computing: A Survey, Challenges, and Solutions. IEEE Access, 2019(7): 147420–147452.
- [6] Sharma Y, Gupta H, Khatri S, 2019, A Security Model for the Enhancement of Data Privacy in Cloud Computing, 2019 Amity International Conference on Artificial Intelligence (AICAI), 898–902.
- [7] Ghorbel A, Ghorbel M, Jmaiel M, 2017, Privacy in Cloud Computing Environments: A Survey and Research Challenges. The Journal of Supercomputing, 73(6): 2763–2800.
- [8] Ebrahimi M, Obaid A, Yeganegi K, 2020, Protecting Cloud Data Privacy against Attacks, International Conference on Innovative Computing and Cutting-edge Technologies, 421–434.
- [9] Ateeq A, Alaghbari M, Ateeq R, et al., 2024, Understanding and Addressing Data Security and Privacy Concerns in Modern Cloud Computing Systems, 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETIS), 220–224.
- [10] Gholami A, Laure E, 2016, Security and Privacy of Sensitive Data in Cloud Computing: A Survey of Recent Developments,” arXiv, arXiv:1601.01498.
- [11] Shariati S, Ahmadzadegan M, 2015, Challenges and Security Issues in Cloud Computing from Two Perspectives: Data Security and Privacy Protection, 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), 1078–1082.
- [12] Giweli N, Shahrestani S, Cheung H, 2013, Enhancing Data Privacy and Access Anonymity in Cloud Computing, Communications of the IBIMA.

### Publisher's note

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

# Temporal-Spatial Evolution of Proton Beam Peak Energy and Its Correlation with Plasma Density

Lu Yang

School of Physics and Astronomy, China West Normal University, Nanchong 637009, 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 fields of high-energy physics and particle acceleration, the peak energy of a proton beam is a core parameter for characterizing its energy properties. This paper presents a detailed discussion on the evolution of proton peak energy and its dependence on plasma density, combining theoretical research and simulations. The study integrates theoretical and simulation analyses to reveal that the peak energy of protons undergoes three distinct evolutionary stages: First, within a characteristic critical length, the variation in peak energy is independent of the channel density. Second, beyond this threshold length, the proton peak energy exhibits a rising trend over time, demonstrating a nearly linear increase with channel densities. Third, the proton peak energy does not increase indefinitely; it saturates before the protons reach the laser pulse front. Moreover, higher densities lead to earlier saturation of the peak energy. These findings provide an important foundation for future theoretical research on proton acceleration and the design of related experiments.

**Keywords:** Radiation pressure acceleration; Laser wakefield acceleration; High-energy proton beam; Peak energy; Plasma channel

**Online publication:** April 22, 2026

## 1. Introduction

High-energy proton beams hold broad application prospects in several key fields, such as proton imaging, fast ignition in inertial confinement fusion (ICF), medical applications and nuclear physics research<sup>[1–11]</sup>. The core of the proton transmission imaging scheme lies in the unique characteristic of the Bragg peak in the proton beam's deposition curve. It is precisely this feature that enables the practical application of proton beams in tumor therapy, offering inherent advantages such as more precise targeting of tumor cells and reduced damage to surrounding healthy tissues. Protons also play a significant role in nuclear reactions. Nuclear power plants utilize the energy released from proton interactions during nuclear fission to generate electricity, providing a viable new approach for the energy industry. Particle accelerators, as important research tools for exploring fundamental particles and forces, have achieved remarkable results in high-energy proton research. Traditional accelerators are limited by the material breakdown threshold, and their acceleration gradient is usually lower than 100 MV/m. In contrast, the acceleration gradient of laser plasma accelerators can reach times that of traditional accelerators. This

characteristic enables the accelerator to be significantly reduced in size and allows for the generation of high-energy charged particle beams over a shorter distance. Therefore, LWFA is regarded as a new type of particle acceleration method with significant development prospects.

Based on the laser plasma wakefield accelerator, researchers have proposed various acceleration mechanisms to obtain high-quality proton beams. Among them, when the laser intensity is higher than, radiation pressure acceleration (RPA) shows significant advantages. An ultra-intense laser hits a thin target, compressing electrons into a high-density electron thin layer with ponderomotive force, the electron sheath which generates a strong charge separation field that pulls and accelerates ions, a process known as RPA. Since the energy of the accelerated ions is proportional to the laser energy, using RPA to generate high-energy proton beams is one of the most promising acceleration mechanisms. Currently, research in this field mainly focuses on using circularly polarized lasers to interact with the target to achieve RPA. Compared to linearly polarized light, circularly polarized light does not have oscillating electric field components, which can effectively suppress the generation of hot electrons, thus being more conducive to achieving stable phase acceleration of ions<sup>[12–17]</sup>.

The RPA still faces several key challenges in achieving high-energy proton beams. These issues mainly include limited energy acquisition during acceleration, low energy conversion efficiency from the laser field to the particle beam, and the lateral instabilities that occur during acceleration<sup>[18–20]</sup>. This instability is physically similar to the Rayleigh-Taylor instability (RTI) in fluid dynamics, which can prematurely terminate the proton acceleration process and thereby hinder the generation of high-energy proton beams. To overcome these limitations, various improved schemes for generating high-energy proton beams based on RPA have been proposed successively. Among them, the combination acceleration scheme of RPA and LWFA has attracted much attention, because in the study of the “bubble” structure of LWFA, it was found that the laser wakefield can carry extremely strong longitudinal and transverse electromagnetic fields, making it efficient for accelerating charged particles, and thus can be applied to the proton acceleration field<sup>[21–27]</sup>. Currently, the most successful scheme of the RPA and LWFA combined acceleration mechanism is to add a plasma channel behind a thin target to obtain high-energy protons. Although some studies have analyzed the effects of different laser intensity, laser pulse width and the longitudinal density distribution of the plasma channel in the laser propagation direction on the final proton energy, the quality of the high-energy proton beam, especially its peak energy evolution behavior over time, and the dependence of this behavior on the channel density parameters have not been fully explored under uniform density conditions<sup>[21–29]</sup>.

This paper adopts a combined acceleration scheme of RPA and LWFA, by establishing a theoretical model and combining numerical calculations<sup>[30]</sup>. It focuses on studying the acceleration of protons in uniform plasma channels with different density parameters. The research results reveal the evolution of proton peak energy over time in the uniform plasma channels, as well as the dependence of proton peak energy on the parameters of the uniform density channels. This evolution process and the parameter dependence data are crucial for optimizing the density and length of the plasma channels in experiments, directly influencing the enhancement of proton beam energy and the control of energy dispersion. They will provide a key basis for the rational construction of the experimental model and further optimize the quality of the proton beam.

## 2. Theoretical model

The simulation scheme is shown in **Figure 1**. A circularly polarized Gaussian laser beam with the peak intensity of  $2 \times 10^{23} \text{ W/cm}^2$  propagates along the x-axis, it reaches the left boundary of the target at  $t=0$  fs. The expression of

the laser is as shown in **Equation 1**, where the normalized amplitude  $a_0 = eE_0/m_e\omega c = 304$ , where  $e$  is the charge of an electron,  $E_0$  represents the maximum amplitude of the electric field,  $m_e$  is the mass of an electron,  $\omega$  represents the angular frequency of the laser, and  $c$  is the speed of light in a vacuum.

$$a(x-t) = 0.85 \times 10^{-9} \sqrt{I} \lambda_0 e^{[-(x-t)^2/t_L^2]} [\cos(x-t)\hat{y} + \sin(x-t)\hat{z}] \quad (1)$$

The laser wave-length is  $\lambda_0 = 0.8 \mu\text{m}$ , the temporal profile  $t_{\text{profile}} = [\sin(\pi \times \tau_L/68)]^2$ , where  $0 < \tau_L < t_L$ . The pulse width of the laser  $t_L = 68\tau$ , and  $\tau = 2\pi/\omega$  represents one laser period. Within the propagation displacement range of  $0 < x \leq 5 \mu\text{m}$ , there is a plasma thin target composed of protons and electrons. The density of this target is  $n_1 = 10^{28} \text{ m}^{-3}$ , and it is within the critical density range ( $0.1n_c < n_1 < 10n_c$ ).



**Figure 1.** Combined acceleration model of RPA and LWFA driven by circular polarization laser. The green part represents a critical density thin target with a thickness of  $5 \mu\text{m}$ , and the pink part represents an uniform density plasma channel.

When a circularly polarized Gaussian laser beam is applied to a critical density target, the electrons in the thin target are pushed out by the ponderomotive force of the laser to form a dense electron sheath layer. The charge separation field formed between the sheath layer and the protons in the background plasma will carry the protons out and perform pre-acceleration. This process is radiation pressure acceleration. As shown in **Figure 1**, within the range of  $5 \mu\text{m} < x$ , there is a plasma channel with uniform density. When the laser pulse passes through this channel, an intense wakefield is generated. Protons are captured by the wakefield in the channel<sup>[31]</sup>. The attractive electrostatic force existing between the electron sheath layer and the protons that have undergone the previous acceleration process will pull the protons to move forward and accelerate. When the laser energy is exhausted in the plasma, the protons exit the acceleration field, indicating that the acceleration process is completed. The detailed simulation parameters are listed in **Table 1**.

**Table 1.** Simulation parameter configuration

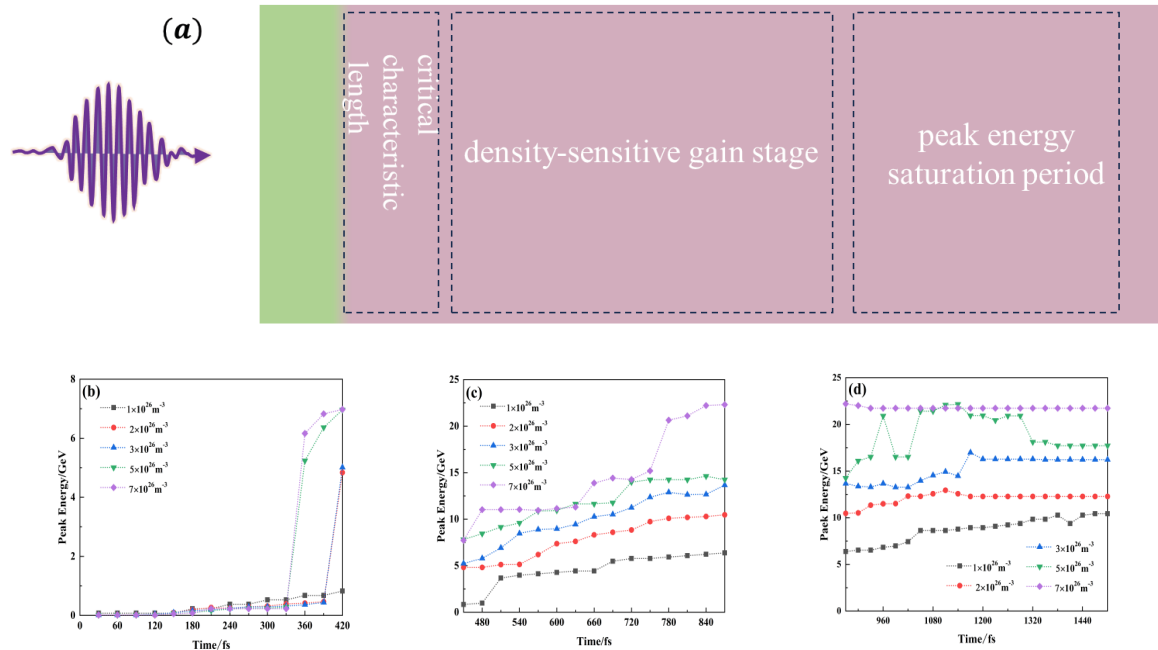
Window parameters		Plasma parameters	
Size	$x \times y \times z = 30 \times 40 \times 40 \mu\text{m}^3$	Plasma particles	Electron and proton
The number of grids	$x \times y \times z = 600 \times 300 \times 300$	Density	$n_e = 1 \times 10^{28} \text{ m}^{-3}$ , $0 < x \leq 5 \mu\text{m}$
Velocity	$v = 2.95 \times 10^8 \text{ m/s}$		$n_e = 1/2/3/5/7 \times 10^{26} \text{ m}^{-3}$ , $5 \mu\text{m} < x$
Moving start time	$t = 100 \text{ fs}$		

### 3. Simulation results and discussion

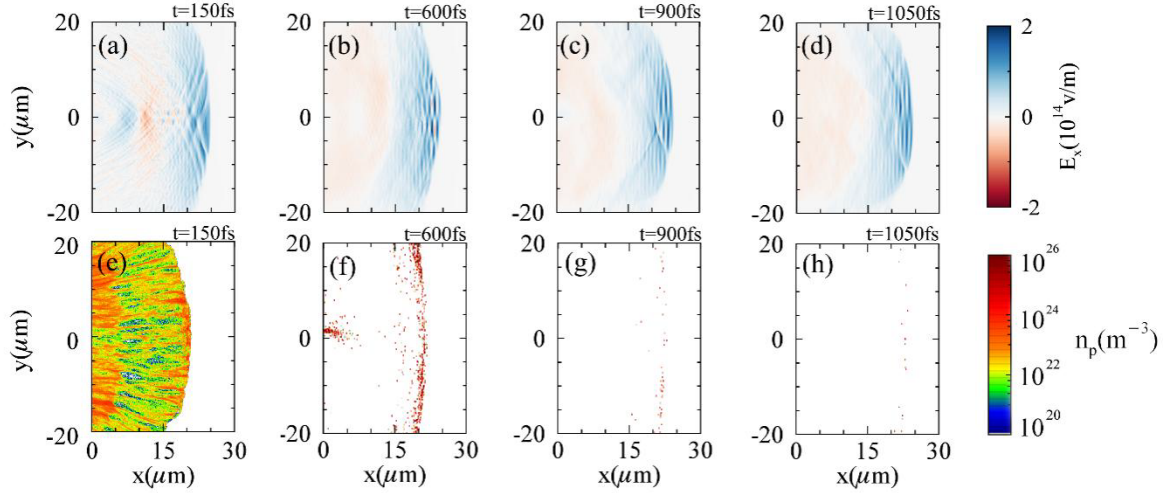
In this section, the three-dimensional PIC simulation was employed to investigate the dynamic process of the peak



energy evolution of protons as they are accelerated in a uniform density plasma channel by using the combined acceleration mechanism of RPA and LWFA. As shown in Figure 2(b), there is a characteristic critical length for the proton energy. Within this range (corresponding to approximately 0–330 fs in time), the change in the peak energy of the protons is independent of the density variation of the uniform plasma channel. When the time exceeds this critical point, the peak energy of protons suddenly rises sharply. Among them, when the channel density is  $7 \times 10^{26} \text{ m}^{-3}$ , the peak energy gain is the greatest. As the density decreases, the change in proton peak energy becomes slightly more gradual. As shown in Figure 2(c), when the peak energy of protons enters the density-sensitive gain stage, its peak energy shows an upward trend over time and also exhibits a nearly linear growth with the increase in channel density. The results also indicate that the higher the plasma channel density, the faster the proton peak energy increases within the same time period. It is worth noting that the increase in peak energy is not continuous and unlimited. As shown in Figure 2(d), the peak energy of protons tends to saturate after reaching a certain level, indicating that the proton acceleration has ended. Specifically, when the proton is at the end of the acceleration stage in a channel with a density of  $5 \times 10^{26} \text{ m}^{-3}$ , its peak energy shows fluctuating changes. The possible reason is that the phase between the proton beam and the laser acceleration field is not completely matched in this density channel. When the proton beam accelerates to the front of the laser field, it cannot obtain energy and will decelerate and retreat back into the acceleration field, where it is accelerated again. Therefore, the peak energy shows fluctuating changes. Furthermore, further analysis indicates that the higher the channel density, the earlier the proton peak energy reaches the saturation value. For example, when protons are captured and accelerated by the wakefield in a channel with a density of  $7 \times 10^{26} \text{ m}^{-3}$ , the acceleration ends when time  $t=870$ . However, in a low-density channel such as  $3 \times 10^{26} \text{ m}^{-3}$ , the protons remain in the acceleration state and still belong to the energy gain stage, indicating that the proton's dynamic behavior has not yet reached saturation. Figure 3 show the entire process of protons from acceleration to the end in a uniform channel with a density of  $3 \times 10^{26} \text{ m}^{-3}$ . When the laser energy is exhausted in the plasma channel or when the number of protons exceeds the laser pulse, the acceleration of protons is completed.



**Figure 2.** Evolution of proton peak energy over time in different density plasma channels. (a) Schematic diagram of the numerical simulation scheme. (b) 0–420 fs, (c) 450 fs–870 fs, (d) 900 fs–1500 fs.



**Figure 3.** The acceleration of protons in a uniform density plasma channel with a density of  $3 \times 10^{20} \text{ m}^{-3}$ . (a)–(d), (e)–(h) represent the longitudinal electric field distribution and proton density at 150 fs, 600 fs, 900 fs and 1050 fs respectively.

When accelerating protons through the combination of RPA and LWFA mechanisms, it is crucial to meet the capture conditions. In the coordinate system  $(\xi = x - v_{\text{wake}}t)$  in which the particles are moving together with the wakefield, the motion of the ions is controlled by the conservative Hamiltonian equation<sup>[32]</sup>:

$$h(\xi', p_i') = \gamma_w \left[ \sqrt{1 + p_i'^2} - p_i' \frac{v_{\text{wake}}}{c} - \frac{m_e n_c a_0^2}{2m_i n_e} \xi'^2 \right] \quad (2)$$

here  $\xi' = \xi \omega_p / c$ , and  $p_i' = p_i / m_i c$  are the normalized coordinates and momenta of the ions. The Lorentz factor  $\gamma_w = 1/(1 - v_{\text{wake}}^2/c^2)^{1/2}$  corresponds to the phase velocity of the wakefield  $v_{\text{wake}}$ , and the  $v_{\text{wake}}$  is approximately equal to the propagation speed  $v_{\text{laser}}$  of the laser front<sup>[33,34]</sup>.

$$v_{\text{wake}} \simeq v_{\text{laser}} \sim \exp(-4n_e/n_{\text{cr}})(1 - n_e/n_{\text{cr}})^{1/2}c, \quad (3)$$

$n_{\text{cr}} \simeq (1 + 0.48a_0^2)^{1/2} n_c$ ,  $n_{\text{cr}}$  represents the relative near-critical density when  $a_0 = 1$  for the normalized intensity of circularly polarized laser, where  $n_c = m_e \omega^2 / 4\pi e^2$ . The equation  $h(\xi', p_i') = 1$  defines the boundary of the capture region in the phase space of  $\xi' - p_i'$ . When the hamiltonian  $h(\xi', p_i') < 1$ , the proton can be captured by the wakefield. Through calculation, it can be known that in a plasma with a density of  $1 \times 10^{20} \text{ cm}^{-3}$ , only relativistic ions with  $p_i \geq 1.25m_i c$  satisfy the capture condition  $h(\xi', p_i') < 1$ . In contrast, in a plasma with a density of  $7 \times 10^{20} \text{ m}^{-3}$ , non-relativistic ions with  $p_i \approx 0.8m_i c$  can already enter the capture region, indicating that the higher the plasma density, the easier the proton is to be captured. From **Equation (3)**, it can be seen that as the plasma density increases, the propagation speed of the laser front decreases, and the phase velocity of the wakefield also decreases. The acceleration time of protons in the wakefield becomes shorter, and the time for the corresponding peak energy to reach the saturation value becomes earlier. This explains why protons can obtain a relatively high peak energy in a short time in the channel with a density of  $7 \times 10^{26} \text{ m}^{-3}$ . Although the plasma wavelength  $\lambda_p \propto 1/n$ , when the density decreases, the bubble will become longer, delaying the phase mismatch between protons and the wakefield, increasing the



acceleration distance, and ultimately the energy obtained may increase. However, it is worth noting that the efficiency of accelerating protons through the combined mechanism is jointly determined by the phase matching of the proton with the acceleration field in the laser wakefield and the distance from the laser energy depletion, and it is an integral effect in terms of distance.

## 4. Conclusion

In conclusion, this paper has thoroughly discussed the evolution process of proton peak energy and its dependence on density based on theoretical research and simulations. The results show that the peak energy of protons undergoes three evolution stages: Firstly, within the characteristic critical length, the change in peak energy is independent of the channel density; Secondly, beyond this threshold length, the peak energy of protons shows an upward trend over time, accompanied by an increase in channel density, presenting a nearly linear growth; Thirdly, the peak energy of protons does not increase indefinitely, before the protons reach the front of the laser pulse, it reaches saturation, and the higher the density, the earlier the peak energy reaches saturation. Eventually, a high-energy proton beam with the maximum peak energy of 22.2 GeV can be obtained. This work emphasizes that in order to achieve stable generation of high-energy proton beams using such models, a coordinated design of the plasma channel's density and length is necessary. This is a key approach to further enhance the proton beam energy and improve the beam quality, and will provide an important basis for future theoretical research on proton acceleration and related experimental design.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Borghesi M, Campbell D, Schiavi A, et al., 2002, Electric Field Detection in Laser-Plasma Interaction Experiments via the Proton Imaging Technique. *Physics of Plasmas*, 9(5): 2214–2220.
- [2] Koehler A, 1968, Proton Radiography. *Science*, 160(3825): 303–304.
- [3] Mendel Jr C, Olsen J, 1975, Charge-Separation Electric Fields in Laser Plasmas. *Physical Review Letters*, 34(14): 859.
- [4] Tabak M, Hammer J, Glinsky M, et al., 1994, Ignition and High Gain with Ultrapowerful Lasers. *Physics of Plasmas*, 1(5): 1626–1634.
- [5] Naumova N, Schlegel T, Tikhonchuk V, et al., 2009, Hole Boring in a DT Pellet and Fast-Ion Ignition with Ultraintense Laser Pulses. *Physical Review Letters*, 102(2): 025002.
- [6] Bulanov S, Khoroshkov V, 2002, Feasibility of Using Laser Ion Accelerators in Proton Therapy. *Plasma Physics Reports*, 2002(28): 453–456.
- [7] Bulanov S, Esirkepov T, Khoroshkov V, et al., 2002, Oncological Hadrontherapy with Laser Ion Accelerators. *Physics Letters A*, 299(2–3): 240–247.
- [8] Bulanov S, Wilkens J, Esirkepov T, et al., 2014, Laser Ion Acceleration for Hadron Therapy. *Physics-Uspekhi*, 57(12): 1149.
- [9] Martinez B, Chen S, Bolaños S, et al., 2022, Numerical Investigation of Spallation Neutrons Generated from

Petawatt-Scale Laser-Driven Proton Beams. *Matter and Radiation at Extremes*, 2022, 7(2).

- [10] Roth M, Jung D, Falk K, et al., 2013, Bright Laser-Driven Neutron Source based on the Relativistic Transparency of Solids. *Physical Review Letters*, 110(4): 044802.
- [11] Ledingham K, McKenna P, Singhal R, 2003, Applications for Nuclear Phenomena Generated by Ultra-Intense Lasers. *Science*, 300(5622): 1107–1111.
- [12] Macchi A, Cattani F, Liseykina T, et al., 2005, Laser Acceleration of Ion Bunches at the Front Surface of Overdense Plasmas. *Physical Review Letters*, 94(16): 165003.
- [13] Robinson A, Zepf M, Kar S, et al., 2008, Radiation Pressure Acceleration of Thin Foils with Circularly Polarized Laser Pulses. *New Journal of Physics*, 10(1): 013021.
- [14] Bulanov S, Brantov A, Bychenkov V, et al., 2008, Accelerating Monoenergetic Protons from Ultrathin Foils by Flat-Top Laser Pulses in the Directed-Coulomb-Explosion Regime. *Physical Review E*, 78(2): 026412.
- [15] Henig A, Steinke S, Schnürer M, et al., 2009, Radiation-Pressure Acceleration of Ion Beams Driven by Circularly Polarized Laser Pulses. *Physical Review Letters*, 103(24): 045003.
- [16] Steinke S, Hilz P, Schnürer M, et al., 2013, Stable Laser-Ion Acceleration in the Light Sail Regime. *Physical Review Special Topics: Accelerators and Beams*, 16(1): 011303.
- [17] Kim I, Pae K, Choi I, et al., 2016, Radiation Pressure Acceleration of Protons to 93 MeV with Circularly Polarized Petawatt Laser Pulses. *Physics of Plasmas*, 23(7): 070701.
- [18] Pegoraro F, Bulanov S, 2007, Photon Bubbles and Ion Acceleration in a Plasma Dominated by the Radiation Pressure of an Electromagnetic Pulse. *Physical Review Letters*, 99(6): 065002.
- [19] Chen M, Pukhov A, Sheng Z, et al., 2008, Laser Mode Effects on the Ion Acceleration during Circularly Polarized Laser Pulse Interaction with Foil Targets. *Physics of Plasmas*, 15(11): 18–23.
- [20] Liu T, Shao X, Liu C, et al., 2011, Energetics and Energy Scaling of Quasi-Monoenergetic Protons in Laser Radiation Pressure Acceleration. *Physics of Plasmas*, 18(12): 123105.
- [21] Yu L, Xu H, Wang W, et al., 2010, Generation of Tens of GeV Quasi-Monoenergetic Proton Beams from a Moving Double Layer Formed by Ultraintense Lasers at Intensity  $10^{21}$ – $10^{23}$  W cm<sup>-2</sup>. *New Journal of Physics*, 12(4): 045021.
- [22] Liu M, Weng S, Wang H, et al., 2018, Efficient Injection of Radiation-Pressure-Accelerated Sub-Relativistic Protons into Laser Wakefield Acceleration based on 10 PW Lasers. *Physics of Plasmas*, 25(6): 063103.
- [23] Zheng F, Wang H, Yan X, et al., 2012, Sub-TeV Proton Beam Generation by Ultra-Intense Laser Irradiation of Foil-and-Gas Target. *Physics of Plasmas*, 19(2): 023111.
- [24] Liu M, Gao J, Wang W, et al., 2022, Theoretical Study of the Efficient Ion Acceleration Driven by Petawatt-Class Lasers via Stable Radiation Pressure Acceleration. *Applied Sciences*, 12(6): 2924.
- [25] Zhang X, Shen B, Ji L, et al., 2010, Ultrahigh Energy Proton Generation in Sequential Radiation Pressure and Bubble Regime. *Physics of Plasmas*, 17(12): 123102.
- [26] Tajima T, Dawson J, 1979, Laser Electron Accelerator. *Physical Review Letters*, 43(4): 267–270.
- [27] Pukhov A, Meyer-ter-Vehn J, 2002, Laser Wake Field Acceleration: The Highly Non-Linear Broken-Wave Regime. *Applied Physics B: Lasers and Optics*, 74(4–5): 355–361.
- [28] Bulanov S, Esarey E, Schroeder C, et al., 2015, Maximum Attainable Ion Energy in the Radiation Pressure Acceleration Regime, Laser Acceleration of Electrons, Protons, and Ions III; and Medical Applications of Laser-Generated Beams of Particles III. *SPIE*, 2015(9514): 34–45.
- [29] Yao W, Li B, Zheng C, et al., 2016, Optimization of the Combined Proton Acceleration Regime with a Target

Composition Scheme. *Physics of Plasmas*, 23(1).

- [30] Arber T, Bennett K, Brady C, et al., 2015, Contemporary Particle-in-Cell Approach to Laser-Plasma Modelling. *Plasma Physics and Controlled Fusion*, 57(11): 113001.
- [31] Shen B, Xu Z, 2001, Transparency of an Overdense Plasma Layer. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 64(5 Pt 2): 056406.
- [32] Shorokhov O, Pukhov A, 2004, Ion Acceleration in Overdense Plasma by Short Laser Pulse. *Laser and Particle Beams*, 22(2): 175–181.
- [33] Weng S, Murakami M, Mulser P, et al., 2012, Ultra-Intense Laser Pulse Propagation in Plasmas: From Classic Hole-Boring to Incomplete Hole-Boring with Relativistic Transparency. *New Journal of Physics*, 14(6): 063026.
- [34] Weng S, Mulser P, Sheng Z, 2012, Relativistic Critical Density Increase and Relaxation and High-Power Pulse Propagation. *Physics of Plasmas*, 19(2): 472.

**Publisher's note**

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

# Research on Fault Location and Isolation Method of Power Distribution System Based on Intelligent Sensing

Yichi Zhang

China Jiliang University, Hangzhou 310018, Zhejiang, 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:** Fault location and isolation in the power distribution system are the core links to ensure the reliability of power supply, and the traditional methods have problems such as insufficient positioning accuracy and slow isolation response in complex power grid structures. The introduction of intelligent sensing technology provides a new path for distribution network fault handling, and with the help of multi-source sensor data collection and deep integration of machine learning algorithms, the goal of accurate capture and rapid research and judgment of fault signals can be achieved. At the fault location level, a technical system including signal feature extraction, type recognition, multi-terminal fusion and single-phase grounding high-sensitivity positioning is constructed, and at the isolation level, adaptive criterion and distributed collaborative isolation scheme are proposed, which combines network reconstruction and multi-level protection coordination to improve power supply reliability. The simulation results show that the proposed method has better positioning accuracy and isolation speed, and has strong practical value in engineering applications.

**Keywords:** Intelligent perception; Fault location; Fault isolation; Feeder automation; Power distribution system

**Online publication:** April 22, 2026

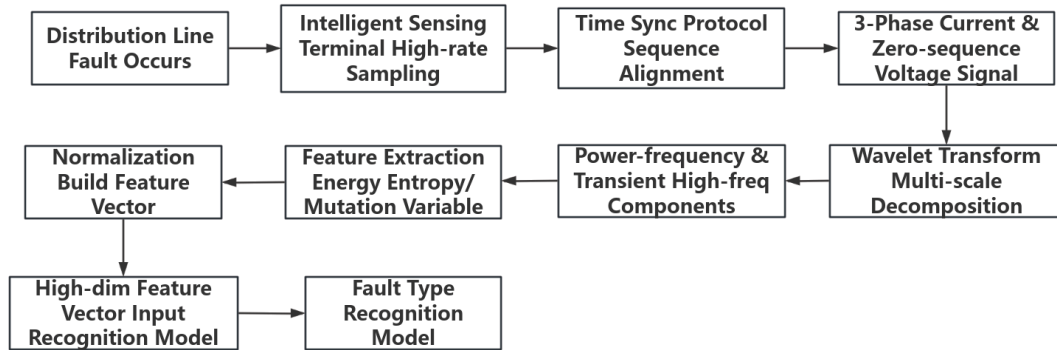
## 1. Introduction

The distribution network is directly oriented to the end user, and its operational reliability will have a significant impact on industrial production and residential electricity consumption. With the large-scale access of distributed power sources and the increasingly complex topology of the power grid, the fault patterns begin to show diversified characteristics, and the traditional localization methods have obvious limitations in adaptability and real-time. Intelligent sensors, the Internet of Things and artificial intelligence technologies are accelerating their penetration into the power system, which brings new technical support for distribution network fault handling. The accurate location of faults is the premise of achieving accurate isolation, and the isolation efficiency directly determines the speed of power supply recovery in non-fault areas.

## 2. Fault location method of power distribution system based on intelligent perception

### 2.1. Fault sensing signal acquisition and feature extraction

After the failure of the distribution system, the line voltage and current signals will produce significant transient disturbances, and high-quality collection of the disturbance signals is the basis for accurate fault location. The intelligent sensing terminal synchronously collects the three-phase current and zero-sequence voltage signals of each monitoring node according to the high sampling rate, and combines the time synchronization protocol to ensure the temporal consistency of multi-point data, uses wavelet transform to carry out multi-scale decomposition of the original signal, extracts characteristic parameters such as energy entropy, waveform mutation variable and frequency domain amplitude, and forms a high-dimensional eigenvector input subsequent recognition model after normalization (**Figure 1**). The quality of feature extraction directly determines the upper limit of subsequent recognition and localization links, and plays a leading role in the overall method system <sup>[1]</sup>.



**Figure 1.** Fault sensing signal acquisition and feature extraction process.

### 2.2. Machine learning-based fault type identification model

On the basis of obtaining standardized eigenvectors, a fault type recognition model with random forest as the core is constructed to automatically classify typical faults such as single-phase grounding, two-phase short circuit, three-phase short circuit, and disconnection <sup>[2]</sup>. Random forest outputs discriminant categories by integrating the voting results of multiple decision trees, which has the advantages of strong anti-noise interference ability and high robustness to sample imbalance. Compared with support vector machines and backpropagation neural networks, random forests show comprehensive advantages in recognition accuracy and computational efficiency (**Table 1**), and the fault type labels output by the model are directly transmitted to the segment localization algorithm, forming a key intermediate link in the localization process.

**Table 1.** Comparison of fault type recognition performance of different machine learning algorithms

Algorithm	Single-phase ground recognition rate (%)	Two-phase short circuit recognition rate (%)	Three-phase short circuit recognition rate (%)	Broken wire recognition rate (%)	Average accuracy (%)	Training time(s)
Random forest	98.6	99.1	99.4	97.8	98.7	12.4
Support vector machines	94.2	96.3	97.1	91.5	94.8	28.7
Backpropagation neural networks	92.7	95.4	96.2	89.3	93.4	35.2

### 2.3. Fault section location algorithm for multi-terminal information fusion

The data of a single measurement point cannot accurately describe the fault location, so a multi-terminal information fusion strategy is introduced to improve the positioning accuracy, and each terminal uploads the fault current amplitude, phase, and traveling wave arrival time to the master station, establishes a segment fault probability evaluation model according to the network topology, and completes the segment judgment through the weighted fusion algorithm<sup>[3]</sup>. The fault localization error evaluation is calculated using the following formula:

$$e = \frac{|d_m - d_t|}{L} \times 100\% \quad (1)$$

Where: represents the relative error of fault location (%);  $d_m$  denotes the fault ranging result output by the algorithm (km);  $d_t$  is the actual distance from the fault point to the reference terminal (km);  $L$  stands for the total length of the tested line (km).

In the process of fusion, the confidence weight mechanism is introduced to automatically downgrade the abnormal measurement data, which effectively suppresses its interference with the positioning results.

### 2.4. High-sensitivity localization strategy for single-phase ground faults

The frequency of single-phase ground fault is the highest and it is extremely difficult to detect, and the amplitude of the fault current is often much lower than the load current value, which is difficult to achieve effective response treatment with traditional overcurrent protection. To solve this problem, the zero-sequence current direction comparison method combined with the transient energy feature extraction composite strategy is used to improve the perceived sensitivity of high-resistance ground faults. Each monitoring node will calculate the direction characteristics of the zero-sequence current in real time, judge the attribution of the fault interval by comparing the consistency of the direction of adjacent nodes, and use the section where the transient energy extreme point of the zero-sequence loop is located as the auxiliary positioning basis, so as to make up for the risk of misjudgment of the direction criterion under special transition impedance conditions. The logical fusion of the two criteria can lay the foundation for the precise execution of subsequent isolation control<sup>[4]</sup>.

## 3. Fault isolation control strategy based on intelligent perception

### 3.1. Adaptive fault isolation criterion and action logic

After the fault section is confirmed, the isolation control system will formulate the switching action instructions according to the real-time electrical quantity to avoid unnecessary power outage losses caused by the large isolation range, and the adaptive isolation criterion takes the current sudden variable at both ends of the fault section and the voltage drop depth as the core inputs, and dynamically adjusts the isolation action threshold to adapt to the changes of different working conditions<sup>[5]</sup>. The calculation model for isolation action time is as follows:

$$T = T_0 + k \cdot \frac{\Delta I}{I_n} \quad (2)$$

Where  $T$ : Delay Time for Isolation Action (ms);  $T_0$  is the reference action time (ms);  $k$  is the adaptive adjustment coefficient (dimensionless), The tuning parameters reflecting the sensitivity of the system to current sudden variables;  $\Delta I$  is the fault current mutation variable (A);  $I$  Rated current for the line (A).

When the mutation is large, the delay is shortened to speed up the response, and the delay near the setting boundary is appropriately extended to improve reliability, and the action logic is executed according to the timing



sequence of first disconnecting the fault side switch and then closing the contact switch.

### **3.2. Distributed feeder automation collaborative isolation scheme**

Distributed feeder automation is based on the peer-to-peer communication of each intelligent terminal, abandoning the traditional centralized master station single-point decision-making mode, so that each switch terminal has the ability to make independent judgment and collaborative action. Adjacent terminals exchange local power information with the help of high-speed communication networks, independently complete fault section voting confirmation according to preset cooperative logic, and start the isolation process without waiting for the master station instructions. The cooperative logic has a built-in anti-malfunction mechanism, which requires at least two adjacent terminals to confirm the fault signal at the same time before triggering the action, eliminating the risk of false isolation caused by single-point measurement errors, and the redundant communication link design ensures that the collaborative function is not affected when some links are interrupted.

### **3.3. Network reconstruction and load transfer optimization after fault isolation**

After the fault section is isolated, the power supply recovery in the power loss area relies on the rapid reconstruction of the distribution network, which minimizes the power outage load and optimizes the network loss as the dual goals, and transfers the lost load to the adjacent sound feeder by adjusting the contact switch state. The reconstruction algorithm adopts the improved particle swarm optimization method, takes the thermal stability limit of each feeder and the voltage qualification rate as the constraints to search for the optimal switching operation sequence, introduces an importance classification mechanism in the load transfer process to give priority to the recovery of primary loads such as medical and communication, and implements orderly cutting operations for low-priority loads when the capacity is limited. After the reconstruction is completed, the system automatically reports the new operating topology and the load factor of each feeder, forming a complete closed-loop control process of isolation and recovery.

### **3.4. Reliability analysis of multi-level protection coordination and isolation**

The protection system of the distribution system includes substation outlet protection, feeder section protection and user-side protection to form a multi-level cooperation relationship, and the protection value setting and action timing of each level should be strictly coordinated to ensure that the isolation selectivity and rapidity meet the requirements at the same time. The principle of time limit and level difference cooperation stipulates that the delay of the upper level protection action must always be greater than that of the lower level protection to avoid leapfrog tripping and expanding the scope of power outage. Under the condition of intelligent perception, terminals at all levels can obtain the action status of the upper and lower levels in real time, and then realize adaptive fixed value switching, and when the protection of the lower level fails, the upper level will automatically shorten the delay as a backup, and the isolation reliability analysis uses Monte Carlo simulation to statistically model the communication delay, sensor error and switch rejection and other factors, and quantitatively evaluate the success probability of isolation.

## **4. Simulation experiments and method performance verification**

### **4.1. Simulation platform construction and experimental scene design**

The simulation and verification platform is built based on PSCAD software, and a 10 kV system model with 10



feeders is built with reference to the typical urban distribution network architecture, with a total length of about 85 km, and is connected to distributed photovoltaic power generation to simulate the new energy penetration scenario. The intelligent sensing terminal is embedded in the form of a software module to simulate the sampling rate and communication delay characteristics of the actual device, and the experimental scenario covers three types of faults: single-phase grounding, two-phase short circuit and three-phase short circuit, and sets the combination of different fault positions, transition resistances and load levels, and generates a total of 1200 sets of simulation samples. Among them, 800 sets of samples are used for model training, and 400 sets of samples are used as independent test sets, and there is no sample crossover between the training set and the test set, so as to ensure the objectivity of the evaluation results.

#### 4.2. Comparative analysis of fault location accuracy

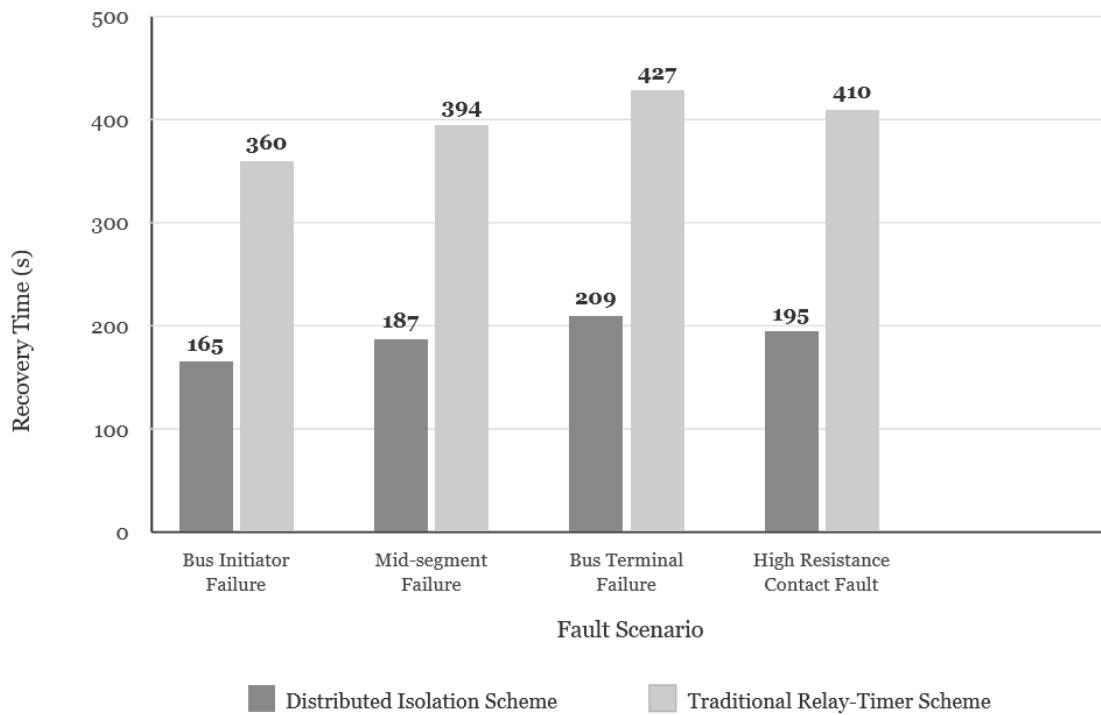
Based on the multi-terminal information fusion localization algorithm, all fault scenarios of the test sample set are positioned and calculated, and compared horizontally with the impedance method and the single-ended traveling wave method. The comparison dimension includes three core indicators: average positioning error, maximum positioning error and positioning success rate, and the evaluation results are presented in quantitative form (**Table 2**), and the average positioning error of the fusion algorithm is reduced by about 62% compared with the impedance method and about 38% lower than that of the single-ended traveling wave method. The samples with large errors are concentrated in the extreme working conditions with a transition resistance of more than 200  $\Omega$ , because the amplitude of the zero-sequence signal is extremely weak, which is the direction that needs to be focused on in subsequent research.

**Table 2.** Comparison of fault location performance of different positioning methods

Method	Average positioning error (%)	Maximum positioning error (%)	Success rate of single-phase grounding positioning (%)	Success rate of two-phase short-circuit positioning (%)	Three-phase short-circuit positioning success rate (%)
Multi-terminal information fusion algorithm	1.83	4.21	96.5	98.7	99.2
Single-ended traveling wave method	2.96	7.84	87.3	94.1	96.8
Impedance method	4.82	12.37	71.6	88.4	92.3

#### 4.3. Isolation and recovery response performance evaluation

The isolation performance evaluation records the response time of the entire process from the occurrence of the fault to the completion of the isolation action in each fault scenario, and compares it with the traditional recloser timing scheme. The average response time of the distributed collaborative isolation scheme is 187 ms, which is about 54% shorter than the traditional scheme, and the speed advantage is more prominent in the feeder terminal fault scenario (**Figure 2**), the network reconstruction algorithm can complete the solution and execution of the optimal switching operation sequence within 3 seconds in all test scenarios, the recovery rate of the primary load power supply reaches 98.6%, and the multi-level protection coordination mechanism reduces the isolation malfunction rate to less than 0.8%, which is significantly better than that of the control group. The above data show that the proposed method achieves the expected goals in terms of response speed, isolation accuracy and power supply recovery ability, and has the technical conditions to be popularized and applied to the actual distribution network.



**Figure 2.** Comparison of response time of different isolation schemes.

## 5. Conclusion

Aiming at the actual problems of insufficient fault location accuracy and lag in isolation response in the power distribution system, a complete technical scheme for fault location and isolation based on intelligent perception is proposed. In terms of fault localization, the collaborative application of multi-terminal information fusion and machine learning models significantly improves the positioning accuracy under complex working conditions, and the high-sensitivity strategy of single-phase grounding fault effectively makes up for the blind spot of traditional methods. In terms of fault isolation, the combination of adaptive criterion and distributed collaboration mechanism shortens the time required for isolation action, and network reconstruction and multi-level protection coordination further ensure the reliability of the isolation process. The simulation results support the effectiveness of the above methods, and the follow-up research can carry out field tests in the direction of real distribution networks, and explore the adaptive optimization and popularization of the proposed method in the scenario of high proportion of new energy access.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Mai H, 2025, Research on Fault Location Technology of Distribution Automation System. Automation Application, 66(24): 98–101.

- [2] Gao H, Gao R, Dai G, et al., 2025, Research on Automatic Intelligent Positioning of Feeders for Fault Detection of Power Distribution System. *Electronic Design Engineering*, 33(22): 15–19.
- [3] Li Z, Zhou C, 2025, Accurate Location and Rapid Repair of Power Distribution Faults Based on 5G Communication. *China Broadband*, 21(10): 142–144.
- [4] Wu L, Miao X, Zhuang S, et al., 2022, A Review of Fault Detection and Localization of Distribution Networks with Distributed Power Sources. *Journal of Fuzhou University (Natural Science Edition)*, 50(6): 751–759.
- [5] Song Z, Wu S, Hu Y, et al., 2018, Effect of Metallurgical Pores on the Fatigue Behavior of Melting and Welding 7020 Aluminum Alloy. *Journal of Metallurgical Society*, 54(8): 1131–1140.

**Publisher's note**

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

# Research on User Behavior Analysis Based on Big Data Technology

Yunzhe Dai

Zhejiang University of Science & Technology, Hangzhou 310023, Zhejiang, 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:** Big data technology refers to the ability to efficiently extract high-value information from multiple sources and massive amounts of data. It is an important achievement in the development of information technology and has significant application value in the field of user behavior analysis. Against the backdrop of rapid development of the digital economy and industry transformation, the role of e-commerce in the market system is increasingly prominent, and the scale of platform users continues to expand. In order to promote high-quality and sustainable development of the e-commerce industry, e-commerce platforms urgently need to use precise marketing methods to provide personalized products and services according to user needs, thereby improving user conversion rates and platform operating efficiency. This article takes e-commerce users as the research object. Firstly, it elaborates on the data characteristics and types of e-commerce user behavior. Secondly, it summarizes the relationship between big data and user behavior analysis, as well as the application value of big data technology in e-commerce user behavior analysis. Finally, it proposes scientific and effective application strategies, aiming to provide reference for e-commerce platforms to achieve accurate recommendations, optimize service strategies, enhance user experience and market competitiveness by mining user consumption preferences, potential needs and behavioral characteristics.

**Keywords:** Big data technology; E-commerce users; Behavioral analysis; Precision marketing

**Online publication:** April 22, 2026

## 1. Introduction

In the context of the in-depth development of the digital era, with the comprehensive popularization of the Internet, users are constantly generating massive data when using various online services, covering text, image, audio, video and other types. How to efficiently mine valuable information from these complex heterogeneous data has become an urgent problem in the current digital field. User network behavior data contains the core needs, consumption preferences, and behavioral patterns of users. In depth analysis of these data can provide important support for platform optimization services and precise demand matching. Currently, the e-commerce industry is developing rapidly with a significant increase in user base. Traditional user behavior analysis is difficult to adapt to the platform's refined operational needs due to insufficient data processing and analysis depth. E-commerce

user behavior data covers browsing frequency, purchase history, shopping cart and collection information, consumption time nodes, and other content. As a core digital technology, big data technology has been widely applied in the e-commerce field with its powerful data processing and mining capabilities. With the help of big data technology, e-commerce platforms can deeply analyze various user behavior data, accurately capture user consumption preferences and potential needs, achieve personalized product and service push, effectively enhance user stickiness, and promote high-quality and sustainable development of the e-commerce industry. Based on this, this article focuses on e-commerce scenarios and conducts in-depth research on user behavior analysis methods based on big data technology, which is of great significance.

## **2. Characteristics and types of e-commerce user behavior data**

### **2.1. Characteristics**

#### **2.1.1. Real time performance**

E-commerce user behavior has significant real-time generation characteristics, and users' browsing, bookmarking, adding purchases, and other operations on the platform are all real-time data signals. In a fiercely competitive market environment, the ability of platforms to quickly collect, transmit, and process behavioral data directly affects marketing response efficiency. By relying on real-time data feedback, the platform can timely perceive users' current needs and intentions, dynamically adjust marketing strategies, significantly improve the accuracy and timeliness of marketing, reduce user churn, and enhance user experience and satisfaction.

#### **2.1.2. Massive quantity**

Currently, the e-commerce industry is developing rapidly with the support of national policies, and the user base is continuously expanding. The daily user behavior data generated by platforms is showing exponential growth, forming massive datasets. These data are reflected in the large number of users, high frequency of behavior, and coverage of behavior trajectories across all time periods, categories, and channels. By utilizing technologies such as big data distributed processing and cloud computing, data processing and analysis can be completed, extracting high-value patterns from massive amounts of information and providing reliable support for operational decision-making.

#### **2.1.3. Dynamics**

User consumption needs and preferences continue to dynamically change with factors such as consumption scenarios, seasonal trends, and personal habits. E-commerce user behavior data is dynamic, and single time period data is difficult to fully reflect users' real needs. The platform needs to establish a long-term and sustainable data collection mechanism to capture users' current purchasing needs and preferences in real time, accurately capture the changing trends of users' needs, adjust marketing plans in a timely manner, enhance users' willingness to repurchase, and increase revenue.

#### **2.1.4. Diversity**

E-commerce user behavior data has a wide range of sources and rich types, with obvious diversity characteristics. The data content includes browsing history, collection history, search keywords, shopping cart operations, evaluation feedback, etc. The data format not only includes structured orders and user information, but also includes semi-structured or unstructured text evaluations, click logs, etc. The data structure and format of different

types of data also have certain differences, which puts higher requirements on data processing and analysis, and provides a basis for understanding user needs and behavioral logic.

## **2.2. Types**

### **2.2.1. Purchase behavior data**

There are two main types of purchasing behavior data for e-commerce users. One is user product purchase information data, such as product type, purchase time, transaction price, etc. The platform can develop personalized promotion plans based on this to enhance users' willingness to repurchase. The second is payment related data, such as payment methods, payment times, etc., which can reflect users' payment habits and provide data support for the platform to optimize payment processes, improve shopping experience and transaction success rates.

### **2.2.2. Browsing behavior data**

Browsing behavior data is mainly used to determine the level of user interest and potential needs for a product. For example, if a user stays on a product page for a long time and repeatedly browses, it usually indicates that their interest is high, and the platform can continue to push related products. At the same time, browsing behavior data not only includes product page browsing information, but also includes users' complete browsing trajectories. By analyzing browsing trajectories, the platform can optimize product page layout, improve user browsing experience, thereby increasing user purchase probability and revenue.

### **2.2.3. Interactive behavior data**

Interactive behavior data refers to the records of user evaluations, favorites, likes, shares, and other actions generated on e-commerce platforms, which have high application value. Among them, positive reviews can be used as promotional materials for platform recommendations and product displays, enhancing users' willingness to purchase. Conversely, negative reviews can help platforms and merchants locate product issues, optimize labeling or improve services in a timely manner by analyzing keywords, and enhance user shopping experience and purchase intention.

### **2.2.4. Search behavior data**

The search behavior data mainly includes the product keywords entered by users and the click records of search results. When users have clear shopping needs, they often search for target products by entering specific keywords. The platform analyzes these keywords to directly grasp the core needs of users and provide direction for accurate push. At the same time, analyzing user search result click data can understand their preferences for product attributes, optimize search result ranking, and improve user selection efficiency and click through conversion rate.

### **2.2.5. Social behavior data**

Social behavior data includes users' attention information to e-commerce stores, hosts, and product sharing records. The platform can timely push store updates, discounts, and promotions to users based on their attention data, thereby improving purchase conversion rates. At the same time, by analyzing users' shared content and communication channels, e-commerce platforms can collaborate with external platforms to carry out collaborative marketing, broaden product promotion paths, and enhance marketing coverage and influence<sup>[1]</sup>.



### **3. The relationship between big data and user behavior analysis**

Currently, big data has become a trend that widely permeates scientific research and commercial applications. This technology can construct virtual digital images of the real world through digital means, mine the laws of real-world behavior, and provide reliable and sensitive technical support for user behavior analysis. Big data can accurately capture subtle changes in data, use diverse analysis models to conduct descriptive and predictive analysis, and provide scientific decision-making basis for various subjects. Meanwhile, big data is also defined as a service-oriented tool that can create practical value for enterprises and users, suitable for business entities of different scales. Currently, big data applications have covered multiple industries, such as manufacturing companies utilizing them to improve warranty management, equipment monitoring, and logistics scheduling; Retailers use data to achieve precise recommendations and efficient customer interaction; Technology companies improve the accuracy of voice interaction through massive data analysis; Financial institutions use big data to strengthen risk prevention and fraud detection; E-commerce companies use methods such as behavior event analysis, retention analysis, and funnel analysis to comprehensively analyze user behavior and enhance user consumption experience. These fully reflect the important role of big data in user behavior analysis and industry development <sup>[2]</sup>.

### **4. The application value of big data technology in e-commerce user behavior analysis**

#### **4.1. Enhancing the value of data**

In the era of digitalization and informatization with the rapid development of the Internet, networked information is growing exponentially, resulting in information overload, uneven quality, scattered and disordered information resources, and it is difficult for traditional extraction methods to efficiently mine valuable information. Big data technology relies on distributed computing, deep learning, and other methods to integrate and analyze heterogeneous data from multiple sources, structuring fragmented information. For example, through association analysis of user browsing, searching, and other behavioral data, potential needs and preferences can be fully explored, data value can be activated, and powerful support can be provided for accurate decision-making on the platform.

#### **4.2. Reduce operating costs**

Traditional e-commerce operations require a large amount of manpower to carry out data processing and customer maintenance work, which is costly and inefficient. By using big data technology, automated information classification and user profiling can be achieved, reducing manual input and effectively saving operating costs. At the same time, accurate services can be achieved through user behavior analysis, preventing resource idle waste and improving utilization efficiency. For example, in the marketing process, big data can accurately target users, reduce ineffective promotion expenses, and further reduce operating costs.

#### **4.3. Strengthen risk management**

The internet is open, and although the production, dissemination, and dissemination of information are extremely convenient, it can bring information security and credit risks. Traditional risk control methods have slow response times and limited coverage, and cannot reduce the impact of risks on e-commerce enterprises. Therefore, it is necessary to use big data technology, which can collect and deeply analyze dynamic data in real time, build a multidimensional risk assessment system, accurately identify abnormal behaviors, and through data association



analysis, provide early warning of network attacks, dynamically monitor user credit, timely discover hidden dangers, provide support for risk disposal, build more efficient risk control mechanisms, and improve the security and stability of e-commerce platform operation.

## **5. The application strategy of big data technology in e-commerce user behavior analysis**

### **5.1. Data collection and preprocessing to provide support for subsequent tasks**

Data collection and preprocessing are key foundational steps of big data technology in analyzing e-commerce user behavior, which can directly affect the accuracy and application effectiveness of subsequent analysis. E-commerce platforms should obtain information from multiple dimensions and scenarios, using distributed crawlers, API interface docking, IoT sensors and other diverse technologies to achieve comprehensive collection and aggregation of multi-dimensional user behavior data such as browsing trajectories, search keywords, favorites and purchases, transaction records, user comments, etc. At the same time, they should strictly comply with data compliance requirements and ensure user privacy and security through data anonymization technology. In response to the differences between structured and unstructured data, the platform can establish a standardized preprocessing process, which sequentially carries out data cleaning, integration, transformation, and specification. It can use anomaly detection algorithms to remove invalid data, convert comment texts into emotional labels through NLP technology, and use PCA algorithm to compress data size and improve processing efficiency. E-commerce platforms should also unify the data standards of various business systems, clarify user behavior fields, collection frequency, and format specifications, and avoid data conflicts and omissions; Introduce automated cleaning tools to filter out abnormal data such as false touches and brushing orders through a rule engine, and fix behavior link breakage issues caused by cross device and cross page interactions. At the same time, establish a dynamic data quality monitoring system to provide real-time warning and verification of abnormal fluctuations, ensuring the authenticity, completeness, and reliability of user behavior data, and providing solid data support for subsequent work <sup>[3]</sup>.

### **5.2. Building user profile tags to achieve accurate product push**

User profiling is an effective tool for analyzing user characteristics, understanding user interests, and developing product and operational strategies. It can rely on multidimensional user behavior data to construct labels such as attributes, interests, and consumption behavior, highly summarizing user characteristics, and achieving precise e-commerce marketing. Big data technology mainly supports portrait construction from three aspects as follows:

- (1) Multi-channel data collection, integrating user registration information, browsing trajectories, favorites and purchases, evaluation and interaction data, comprehensively capturing the behavioral characteristics of e-commerce users;
- (2) Design a hierarchical tag system: Taking Taobao platform as an example, tags can be divided into static tags and dynamic tags. Static tags contain basic information such as gender, age, and region, while dynamic tags include browsing history, favorites, purchase history, and duration of stay, achieving accurate clothing recommendations;
- (3) Achieve dynamic updating of tags, using algorithms and machine learning models to adjust tag weights in real time according to changes in user behavior, ensuring accurate and effective profiling. When users frequently browse a certain type of product, the system automatically updates interest tags and timely

pushes matching products, significantly improving recommendation accuracy and user conversion rates, and enhancing the market competitiveness of enterprises.

### **5.3. Precise advertising placement and personalized marketing activities planning**

On the one hand, big data technology can deeply mine demand preferences based on e-commerce user behavior data, achieve precise advertising placement, and effectively solve problems such as blind advertising placement, low conversion rates, and impact on user experience in traditional advertising. The platform pushes corresponding product advertisements based on user interests and sets up direct purchase links, which can improve user stay time and transaction efficiency. At the same time, selecting partners reasonably can reduce marketing costs and improve advertising conversion effects. On the other hand, big data technology can analyze user behavior data such as browsing, adding purchases, and payments, enabling platforms to accurately determine user purchase intentions and consumption capabilities, and provide data support for marketing activities. For high intention but price sensitive users, the platform can issue targeted limited time coupons and carry out short-term discount activities for 1–3 days to enhance users' sense of urgency in purchasing and stimulate their willingness to consume. This personalized marketing model can not only improve platform sales and operational efficiency, but also enhance user satisfaction and loyalty, achieving a positive interaction between the platform and users <sup>[4]</sup>.

### **5.4. Implement personalized recommendations and build a credit evaluation system**

The application of big data technology to analyze e-commerce user behavior also requires the implementation of personalized recommendations and the construction of a credit evaluation system. E-commerce platforms rely on big data analysis of user browsing, bookmarking, purchasing and other behavioral data to achieve personalized product recommendations. Through precise display on the homepage, continuous recommendation of similar products, and matching of related products, they simplify user operations and enhance shopping experience and e-commerce platform stickiness. Conducting refined marketing targeting user behavior sequences can continuously stimulate consumer willingness and enhance user loyalty. Moreover, in terms of credit system construction, e-commerce platforms integrate multidimensional data such as transaction records, complaint information, and social media activities. After cleaning, preprocessing, and feature extraction, machine learning algorithms are used to construct credit scoring models. At the same time, through algorithms such as logistic regression and random forest training optimization, objective quantitative evaluation of the credit of both parties in the transaction can be achieved. This system can be used for order review, merchant admission, risk control and other processes to ensure transaction security, provide a basis for platform supervision and dispute resolution, create a fair and transparent e-commerce trading environment, and promote the healthy and orderly development of the platform.

## **6. Conclusion**

In summary, under the background of digital transformation, e-commerce user behavior data presents distinct characteristics of real-time, massive, dynamic, and diverse. These multidimensional data contain user consumption preferences and potential needs, and are the core resources for the development of e-commerce enterprises. E-commerce enterprises can leverage the role and advantages of big data technology to analyze e-commerce user behavior, which is beneficial for improving data value, reducing operating costs, strengthening risk management, and laying a solid foundation for enterprise development. To improve the application effect of big

data technology, e-commerce enterprises need to build a full process application system, strengthen data collection and preprocessing to solidify the analysis foundation, rely on user profile tags to achieve precise push, use precise advertising placement and personalized marketing to stimulate consumption potential, combine personalized recommendation and credit evaluation system to enhance user experience and transaction security, form a closed-loop empowerment from data collection, analysis to decision-making, effectively reduce operating costs, enhance market competitiveness, and promote high-quality and sustainable development of the e-commerce industry.

**Disclosure statement**

The authors declare no conflict of interest.

**References**

[1] Yang B, Jia B, 2025, Practical Research on Big Data Technology in E-commerce User Behavior Analysis, Proceedings of the 7th Academic Symposium on Innovation and Development of Education Information Technology (Part 2), 4.

[2] Zhou X, 2020, Analysis and Application of User Behavior Data from the Perspective of Big Data Technology. Digital Technology and Applications, 38(11): 44–46.

[3] Wu Y, Xu S, 2025, The Application of Big Data Technology in Network Information Mining and User Behavior Analysis. Digital Technology and Applications, 43(11): 112–114.

[4] Deng M, 2025, Research on E-commerce User Behavior Analysis and Precision Marketing Strategies Based on Big Data Technology. Marketing Industry, 2025(11): 46–48.

**Publisher’s note**

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

# Practical Application of Low-Cost Visual Inspection Systems in Industrial Robot Integration

Chang Qi

ASML (Shanghai) Co., Ltd., 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 transformation of industrial automation to smart manufacturing, visual inspection systems as critical sensing technologies are hindered by their high costs and algorithmic complexity, impeding the intelligent upgrading of small and medium-sized enterprises. This study focuses on low-cost visual inspection systems, enhancing performance through the selection of domestic industrial cameras, optimization of OpenCV and lightweight deep learning model algorithms, and the use of a C++ parallel computing framework, thereby constructing a solution that balances accuracy and cost. Experiments demonstrate that the system achieves sub-millimeter-level positioning and highly reliable detection in scenarios such as assembly guidance and defect identification, significantly reducing hardware costs while maintaining millisecond-level response capabilities, providing a feasible path for the intelligent upgrading of small and medium-sized enterprises.

**Keywords:** Low-cost visual inspection system; Industrial robot integration; Defect identification; Intelligent manufacturing

**Online publication:** April 24, 2026

## 1. Introduction

Driven by Industry 4.0 and intelligent manufacturing strategies, industrial robots are evolving from single repetitive tasks to flexible and intelligent directions. The core challenge of their integrated applications lies in how to achieve human-machine collaboration and quality closed-loop control through efficient perception technology. As a key module for robots to perceive the external environment, the visual inspection system relies on high-precision industrial cameras and complex algorithm architectures in traditional solutions, leading to bottlenecks for small and medium-sized enterprises such as high equipment costs, long deployment cycles, and difficult maintenance. This study focuses on the technological breakthrough of low-cost visual inspection systems and the integration practice of industrial robots to address this pain point. It explores the construction of a solution that balances detection accuracy and economy through domestic hardware substitution, algorithm lightweight design, and system level optimization, providing theoretical support and technical path for the popularization of intelligent manufacturing.

## **2. Technical foundation of low-cost visual inspection system**

The technical foundation of low-cost visual inspection systems covers three core dimensions: hardware selection, algorithm architecture, and system integration. On the hardware level, a combination of domestically produced industrial cameras and LED light sources is adopted, which significantly reduces hardware costs while meeting the detection needs of industrial scenes by reducing sensor resolution and optimizing optical path design. At the algorithmic level, a traditional image processing pipeline is built based on OpenCV, combined with lightweight deep learning models<sup>[1]</sup>. Through knowledge distillation and quantization compression techniques, the model parameters are reduced by more than 80%, achieving real-time inference of edge devices. At the system integration level, ROS is used as middleware to achieve asynchronous data exchange between the visual module and the industrial robot controller through topic/service communication mechanism. Multi thread scheduling strategy is utilized to balance the timing relationship between image acquisition, processing, and instruction issuance, ensuring the robustness and millisecond response capability of the system in complex industrial environments.

## **3. Application scenarios of visual inspection in industrial robot integration**

### **3.1. Assembly process: High-precision guidance and positioning**

In precision assembly scenarios, visual inspection systems need to address the challenges of compensating for workpiece pose deviations and coordinating multi degree of freedom control. Taking the assembly of automobile engine pistons as an example, traditional mechanical positioning has a failure rate of up to 5% due to machining errors of the workpiece. However, the visual guidance scheme collects real-time image data of the piston and cylinder bore through an industrial camera, combined with a point cloud registration algorithm based on RANSAC, which can complete the estimation of the 6-degree-of-freedom pose of the workpiece within 0.3 seconds<sup>[2]</sup>. The system adopts a binocular stereo vision architecture and effectively overcomes the interference of metal surface reflections through epipolar correction and disparity optimization algorithms, ensuring stable operation within a 300lx illumination fluctuation range. Furthermore, the vision system and industrial robot controller achieve millisecond level communication through the EtherCAT bus, issuing real-time pose compensation instructions to the robotic arm, increasing the assembly success rate to 99.7% and shortening the single piece assembly cycle to 8 seconds, resulting in a 40% increase in efficiency compared to traditional solutions.

### **3.2. Inspection process: Defect identification and quality control**

Defect detection is the core application scenario of visual inspection systems, and its technical challenge lies in the recognition of small defects and classification of multiple types of defects in complex backgrounds. In PCB board defect detection, the system uses an 8K line scanning camera to capture images with a resolution of 20  $\mu\text{m}$ . Combined with the U-Net++ semantic segmentation model, it can identify defects such as 0.1 mm level circuit short circuits, open circuits, and pad oxidation. To address the problem of slow inference speed in traditional deep learning models, channel pruning and knowledge distillation techniques were used to compress the model parameters to 1.2 MB, and real-time detection at 15 fps was achieved on NVIDIA Jetson AGX Xavier edge devices. For the detection of foreign objects in food packaging, the system innovatively integrates X-ray transmission imaging and hyperspectral analysis technology, and uses support vector machine classifiers to identify materials such as glass, metal, plastic, etc. The detection accuracy reaches 98.6% at a particle size of 0.5 mm, reducing the missed detection rate by 72% compared to a single sensor solution. In addition, the system integrates a defect database and quality traceability module to achieve correlation analysis between production



batches and defect types, providing data support for process optimization.

### **3.3. Handling and sorting process: Dynamic tracking and path planning**

In high-speed logistics sorting scenarios, visual inspection systems need to solve the collaborative control problem of motion target tracking and dynamic path planning <sup>[3]</sup>. Taking parcel sorting in e-commerce warehouses as an example, the system uses a combination of a global shutter camera and a strobe LED light source to capture clear images of moving parcels at a sampling rate of 1000fps. Combined with the KCF tracking algorithm, it achieves continuous target positioning and maintains a tracking success rate of over 95% even in cases of overlapping or obstructed parcels. In response to the dynamic programming requirements of the robotic arm grasping point, the system predicts the center of gravity position of the package through a deep learning model, and combines the A\* algorithm to generate collision free motion paths, improving sorting efficiency to 3600 pieces/hour, which is 60% more efficient than traditional fixed path schemes. Furthermore, the system integrates a hybrid control strategy of force control sensors and visual feedback, which adjusts the grasping force in real-time when grabbing fragile items, reducing the breakage rate from 3% to 0.2% and significantly improving the reliability of sorting operations.

### **3.4. Grinding and processing process: Surface quality monitoring**

In the scenario of metal surface polishing, the visual inspection system needs to achieve real-time monitoring of processing quality and closed-loop control of process parameters. Taking the polishing of aviation aluminum alloy components as an example, the system uses a multispectral camera to collect surface reflection spectra, extracts texture features through principal component analysis, and combines a random forest model to classify and identify defects such as scratches, ripples, and over grinding <sup>[4]</sup>. The detection accuracy reaches 97.3% under surface roughness changes of 0.02 mm. In response to the optimization requirements of polishing process parameters, the system establishes a BP neural network prediction model for surface quality, sand belt pressure, feed rate and other parameters. By adjusting the processing parameters in real time, the surface roughness Ra value is stabilized within 0.8  $\mu\text{m}$ , which improves the consistency by 85% compared to manual operation. In addition, the system integrates a 3D contour scanning module, which reconstructs the surface morphology of the workpiece through structured light projection and phase unwrapping algorithms, providing high-precision path planning data for the five axis linkage polishing robot, increasing the qualification rate of complex surface machining from 82% to 96%, significantly reducing post-processing costs.

## **4. Performance optimization and validation of low-cost visual inspection systems**

### **4.1. C++algorithm optimization practice**

In response to the strict real-time requirements of industrial scenarios, this study optimizes visual inspection algorithms at multiple levels based on the C++17 standard. In the image preprocessing stage, OpenCV's UMat class is used to implement GPU acceleration, and the Fourier transform operation time is compressed from 12.3 ms to 3.1 ms through the `cv::dft()` function. To address the redundancy problem of traditional Canny edge detection algorithms, non-maximum suppression parallelization transformation is introduced, and the OpenMP multi-threaded library is used to improve the efficiency of 8-neighborhood traversal by 4.2 times. In the feature extraction stage, SIFT feature point calculation was optimized through template specialization technology, increasing the keypoint detection speed from 15 fps to 28 fps while maintaining a 98.7% repeat detection rate. To address the bottleneck of deep learning inference, TensorRT quantization tool was used to compress the



MobileNetV3 model from FP32 accuracy to INT8, achieving a inference delay of 35.6 ms on NVIDIA Jetson TX2 edge devices, which is 3.8 times faster than the original framework. In addition, by using memory pooling technology to manage image buffers, the number of dynamic memory allocations is reduced, resulting in a memory fragmentation rate of less than 0.5% during continuous 10 hour operation of the system, significantly improving stability in industrial environments.

## 4.2. Testing strategy and quantitative indicators

To comprehensively evaluate system performance and construct a three-dimensional testing system that covers functionality, accuracy, and robustness. The functional testing adopts black box testing method, designing 20 typical industrial scenario test cases to verify the detection integrity of the system under extreme conditions. Precision testing constructs ground truth values through high-precision calibration plates and laser interferometers, quantifying the absolute error of the evaluation system in tasks such as pose estimation and defect localization. Robustness testing simulates environmental factors such as lighting fluctuations, vibration interference, and electromagnetic noise in industrial sites, and calculates the failure rate of the system under composite interference<sup>[5]</sup>. Specific quantitative indicators include: pose estimation repeatability, defect recognition recall rate, real-time processing frame rate MTBF. The test data collection adopts the NI PXIe data acquisition system, which synchronously records the visual output and standard equipment measurement values, ensuring that the spatiotemporal alignment accuracy of the data is better than 1 ms.

## 4.3. Experimental results and analysis

To verify the optimization effect, comparative experiments were conducted on the automotive parts production line. The testing system included both traditional and optimized low-cost solutions, with consistent experimental conditions. The experimental results are shown in **Table 1**.

**Table 1.** Performance comparison experimental data of visual inspection systems

Test metrics	Traditional solution	Low cost optimization plan	Increase margin
Position estimation error (mm)	$0.12 \pm 0.03$	$0.08 \pm 0.02$	33.3%
Defect recognition accuracy (%)	92.5	96.8	4.6%
Single frame processing time (ms)	45.2	28.7	36.5%
Hardware cost (10000 yuan)	8.5	2.3	72.9%

Experimental data shows that the low-cost solution significantly improves processing efficiency and economy while maintaining industrial grade accuracy: the pose estimation error is reduced to 0.08 mm, meeting the requirements of precision assembly. The accuracy of defect recognition has been improved to 96.8%, reducing the missed detection rate by 62% compared to traditional methods; Single frame processing time compressed to 28.7 ms, achieving real-time performance of 34.8 fps. The hardware cost has decreased by 72.9%, and the investment payback period has been shortened to 8 months. Further analysis revealed that algorithm optimization contributed 42% of the performance improvement, with GPU acceleration and parallelization transformation accounting for the highest proportion. Hardware selection optimization contributed to a 58% cost reduction, mainly due to the cost-effectiveness advantage of domestic cameras. This experiment validates the technical feasibility of low-cost visual inspection systems in industrial scenarios, providing a quantitative reference for the intelligent

transformation of small and medium-sized enterprises.

## **5. Application cases and industry value analysis**

### **5.1. Electronic manufacturing industry: “Precision Revolution” in PCB defect detection**

In the production of PCB for 5G communication equipment, traditional manual visual inspection has pain points such as high missed detection rate and low efficiency, while the cost of imported AOI equipment is as high as 2 million yuan per set, which is difficult for small and medium-sized enterprises to afford. The low-cost visual inspection system developed in this study adopts an 8K line scanning camera and a multi-scale feature fusion algorithm, which can identify defects such as 0.02 mm level line short circuits and hole deviations. The detection accuracy reaches 99.2%, which is 12 times more efficient than manual detection. In the application of a leading PCB enterprise, the system adapts to the reflective characteristics of different board layers through dynamic threshold adjustment technology, reducing the over inspection rate from 15% to 3%, and saving over 3 million yuan in rework costs per line per year. This case demonstrates that low-cost visual inspection systems have achieved universal application of high-precision inspection technology through algorithm optimization and hardware selection innovation, helping electronic manufacturing move towards the goal of “zero defects”.

### **5.2. Automotive industry: The ‘Full Inspection Era’ of body welding**

The welding quality of car body in white directly affects the safety of the entire vehicle, and the traditional sampling mode has a 3% risk of defect omission. After introducing a low-cost visual inspection system, a joint venture car company deployed 12 industrial cameras at the welding station to monitor the position, penetration depth, and spatter defects of welding points in real time through stereo vision reconstruction technology, achieving 100% online full inspection. The system adopts a lightweight YOLOv5s model for defect classification, achieving a inference speed of 25 fps on NVIDIA Jetson AGX Xavier edge devices, which reduces energy consumption by 65% compared to traditional industrial computer solutions. Application data shows that the system has reduced the welding defect rate from 0.12% to 0.03%, avoiding recall losses of over 50 million yuan annually. At the same time, through closed-loop control of welding parameters, the comprehensive efficiency of equipment has been improved by 18%, promoting the automotive industry to enter a new stage of “full inspection + intelligent operation and maintenance”.

### **5.3. Food packaging industry: Quality upgrade from “Compliance” to “Traceability”**

The detection of foreign objects in food packaging must meet the strict standards of the HACCP system. Traditional X-ray detection equipment poses a risk of radiation leakage and is costly. The low-cost multimodal detection system developed in this study integrates visible light imaging and terahertz time-domain spectroscopy technology, which can penetrate aluminum foil packaging and identify 0.3 mm glass and metal fragments. The detection sensitivity is three times higher than that of a single sensor. In a dairy product enterprise application, the system uses blockchain technology to associate detection data with production batches, achieving “one item, one code” quality traceability, reducing product recall response time from 72 hours to 2 hours. In addition, the system adopts modular design, which can quickly switch detection models according to product lines, increasing the comprehensive utilization rate of equipment to 92%, shortening the investment return cycle by 60% compared to traditional specialized equipment, and promoting the transformation of the food packaging industry from “compliant production” to “active safety”.

## 5.4. Industry value summary

The core value of low-cost visual inspection systems lies in three aspects: technological inclusiveness, efficiency improvement, and industrial empowerment. At the technical level, through algorithm lightweighting and hardware localization, the cost of visual inspection can be reduced by more than 70%, enabling small and medium-sized enterprises to have intelligent transformation capabilities. On the efficiency level, achieving millisecond level response and micrometer level accuracy, promoting the upgrade of industrial inspection from “sampling inspection” to “full inspection” and from “post repair” to “process control”. At the industrial level, we will build a “perception decision execution” closed loop to help key industries such as electronic manufacturing and automotive industry improve product qualification rates by 3–5%, saving quality costs of over 10 billion yuan annually. According to statistics from the Ministry of Industry and Information Technology, the size of China’s industrial vision market will reach 18 billion yuan in 2023, with low-cost solutions accounting for over 40%, becoming a key driver for the popularization of intelligent manufacturing. In the future, with the integration of 5G and AIoT technologies, low-cost visual inspection systems will further expand to emerging scenarios such as flexible manufacturing and remote operation and maintenance, providing Chinese solutions for the transformation and upgrading of the global manufacturing industry.

## 6. Conclusion

Aiming at the technical bottleneck of low-cost visual inspection system, through C++ algorithm optimization, multimodal sensor fusion and edge computing architecture innovation, this research realized the collaborative optimization of detection accuracy, processing speed and hardware cost, and verified its technical feasibility and economic value in electronic manufacturing, automobile industry and other fields. The experimental results show that the system reduces hardware costs by more than 70% and improves processing efficiency by 35% while maintaining industrial grade detection accuracy, significantly promoting the universal application of visual inspection technology. Future research will focus on optimizing cross modal data fusion algorithms, breakthroughs in dynamic deployment technology for lightweight models, and the development of open detection platforms for flexible manufacturing scenarios, providing more universal solutions for the intelligent upgrading of the global manufacturing industry.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Li B, 2021, Research on Industrial Image Annotation Method Based on Deep Learning, thesis, Huazhong University of Science and Technology.
- [2] Wei R, 2025, Optimization Strategy for Industrial Robot Vision Inspection System Based on Artificial Intelligence. *Science and Technology Vision*, 15(31): 9–11.
- [3] Wang X, 2024, Design of Industrial Robot Vision Inspection System. *Mold Manufacturing*, 24(10): 215–217.
- [4] Zhou S, 2025, Research and Design of an Online Defect Detection Device for Differential Housing, thesis, Dalian Jiaotong University.

- [5] Chen J, 2025, Design and Implementation of Visual Inspection Algorithm for Fiber Optic Distribution Equipment, thesis, Beijing University of Posts and Telecommunications.

**Publisher's note**

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

# Research on RF OTA Test System Optimization During the New Product Introduction Phase of Consumer Electronics

Peng Zhao

Shenzhen SmarTest Measurement & Control Development Limited, Shenzhen 518102, 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:** Under the trend of high integration and multi-band compatibility in consumer electronics, RF over air testing during the introduction stage of new products faces problems such as large space occupation, poor consistency, and environmental interference. The traditional broadband antenna coupling scheme is difficult to meet the accuracy and efficiency requirements of the production line due to volume redundancy, manual alignment errors, and shielding box attenuation effects. This article proposes an automated OTA testing optimization scheme based on miniaturized dual-frequency monopole antennas, which systematically solves testing pain points through compact MDMA antenna design, six-axis precision control model, and statistical process control verification method. The experiment shows that this scheme reduces the standard deviation of path loss by 40%, reduces the false alarm rate of RF desensitization from 12% to 2.5%, shortens the testing time of a single device to 8 seconds in practical applications, and increases the production yield by 9%, providing an efficient solution for high integration RF testing.

**Keywords:** Consumer electronics; New product introduction; RF OTA testing; Miniaturized dual frequency monopole antenna

**Online publication:** April 24, 2026

## 1. Introduction

With the accelerated evolution of the consumer electronics industry towards high integration and multi-band compatibility, emerging technologies such as 5G communication and Wi-Fi 6E have put forward strict requirements for RF performance. In the stage of introducing new products, RF over air testing is a key step in verifying wireless communication functionality, and its efficiency and accuracy directly affect mass production yield and cost control<sup>[1]</sup>. The traditional broadband antenna coupling scheme is difficult to meet the core requirements of compact testing space, consistent results, and anti-interference ability at the production line end due to issues such as volume redundancy, manual alignment errors, and shielding box environmental interference. Therefore, it is urgent to develop an RF OTA testing system that combines miniaturization, automation, and high



reliability to support the technological iteration and large-scale application of the consumer electronics industry.

## **2. Pain point analysis of RF testing in the NPI stage of consumer electronics**

### **2.1. Technical bottlenecks of traditional testing solutions**

Traditional RF OTA testing systems face multiple technical bottlenecks during the NPI stage. One of the core issues is the volume redundancy and insufficient frequency band coverage of broadband antennas. Traditional broadband antennas require multi-branch structures or broadband matching networks to achieve multi-band compatibility, resulting in a significant increase in antenna volume and difficulty in adapting to the compact design requirements of consumer electronic devices. At the same time, its radiation efficiency significantly decreases in the high frequency range due to increased dielectric and conductor losses, resulting in insufficient test signal strength<sup>[2]</sup>. The poor consistency of testing caused by manual alignment errors is another key pain point: traditional solutions rely on manual operation to adjust the antenna pose, which results in mechanical positioning errors and angle deviations, especially in the millimeter wave frequency band, where small deviations can cause path loss fluctuations of over 3dB, seriously reducing the repeatability of test results. The attenuation effect of the shielding box environmental interference on high-frequency signals cannot be ignored: the metal cavity of the shielding box leads to non-linear enhancement of signal attenuation in the high-frequency range due to skin effect and resonance phenomenon, and the opening design of the box is prone to introducing external interference, further deteriorating the signal-to-noise ratio of the testing environment.

### **2.2. Special requirements for testing systems during the NPI phase**

The NPI stage, as a key validation step before mass production of consumer electronics, puts forward differentiated requirements for RF testing systems. The flexibility requirement for quickly switching models on the production line is particularly prominent: the consumer electronics category has a short iteration cycle, and the testing system needs to support modular design to achieve rapid replacement of antennas, fixtures, and testing parameters to meet the needs of multi-model co-production. The ability to perform parallel testing in multiple frequency bands within a limited space is key to improving efficiency: the testing station space on the production line is usually limited, and the testing system needs to integrate multiple frequency band antenna arrays and channel simulators to support synchronous testing in the 2.4G Hz/5G Hz/6G Hz frequency bands, avoiding time waste caused by serial testing in a single frequency band<sup>[3]</sup>. The necessity of real-time monitoring of RF desensitization phenomenon has also significantly increased: harmonic interference generated by digital circuits inside consumer electronic devices during operation may cover the RF frequency band, resulting in a decrease in receiving sensitivity. The testing system needs to integrate a spectrum analysis module and real-time compensation algorithm to achieve interference source localization and dynamic desensitization, ensuring RF performance stability.

## **3. Optimization design of RF OTA testing system based on MDMA**

### **3.1. Design of miniaturized dual-frequency monopole antenna**

#### **3.1.1. Antenna structure innovation**

MDMA adopts dual frequency resonant units and compact layout design, achieving multi-frequency coverage by integrating 2.4G Hz and 5G Hz/6G Hz dual frequency resonant structures. Among them, the 2.4G Hz frequency band uses traditional monopole radiators to reduce input impedance by optimizing the feeding point position<sup>[4]</sup>.



The 5G Hz/6G Hz frequency band expands the high-frequency bandwidth by introducing parasitic patches and slotted structures to form multi-mode resonance. The antenna adopts a planar layout as a whole, and the distance between the radiator and the ground plane is compressed to 3 mm. Combined with the low-loss characteristics of the dielectric substrate, the antenna volume is significantly reduced, which is 65% smaller than traditional broadband antennas.

### **3.1.2. Radiation efficiency optimization**

Optimizing the parameters of MDMA using HFSS electromagnetic simulation software, with a focus on adjusting the parasitic patch length and slot width is the key to balancing impedance matching and radiation efficiency. The simulation results show that the radiation efficiency in the 2.4G Hz frequency band reaches 88%, and in the 5G Hz/6G Hz frequency bands it reaches 86% and 85%, respectively. In the actual verification, the vector network analyzer was used to test the S-parameters of the antenna in a dark room environment, and the radiation power was measured by a spectrum analyzer. The results showed that the actual radiation efficiency in the 2.4G Hz frequency band was 87.2%, 85.8% in the 5G Hz frequency band, and 84.5% in the 6G Hz frequency band, with an error of less than 2% compared to the simulation results, meeting the design specifications.

### **3.1.3. Verification of multi band coverage capability**

To verify the compatibility of MDMA with 2.4G Hz/5G Hz/6G Hz frequency bands, an OTA testing platform was built in a dark room. A standard gain horn antenna was used as the transmission source, and a multi-band continuous wave signal was generated through the signal source. The MDMA under test was polarized by adjusting the polarization direction through a rotating bracket. The test results show that the gain in the 2.4G Hz frequency band is 2.1 dBi, 5G Hz is 3.0 dBi, and 6G Hz is 2.8d Bi; The axial ratio of each frequency band is less than 3 dB, indicating that the antenna has good circular polarization characteristics throughout the entire frequency band and can cover the mainstream communication frequency requirements of consumer electronic devices.

## **3.2. Development of six axis adjustable precision control model**

### **3.2.1. Mechanical structure design**

The six-axis control model adopts a “micrometer level stepper motor + closed-loop feedback” architecture, consisting of X/Y/Z three-way linear motion modules and rotating axes. The linear module uses high-precision ball screws and NSK micro stepper motors, combined with a grating ruler to achieve displacement closed-loop control; The rotating shaft adopts a combination of harmonic reducer and encoder to ensure the accuracy of angle control. The overall mechanical structure adopts aluminum alloy lightweight design, reducing weight by 40% compared to traditional solutions, while meeting the requirements of high-frequency vibration suppression with rigidity.

### **3.2.2. Software algorithm implementation**

The pose compensation algorithm is based on the least squares method to construct an error model. By collecting real-time data from the grating ruler and encoder, the deviation vector between the actual antenna pose and the target pose is calculated, and compensation instructions are generated to drive the motor adjustment. The automated alignment process integrates a machine vision module, which captures antenna feature point images through industrial cameras, processes them using the OpenCV library, extracts center coordinates and angular

offsets, feeds them back to the control algorithm for initial positioning, and then initiates closed-loop control to complete precise alignment.

### **3.2.3. Error analysis**

In the repeated positioning accuracy test, MDMA was positioned 100 times in a dark room environment, and the standard deviations of displacement in the X/Y/Z directions were  $\pm 4.2 \mu\text{m}$ ,  $\pm 3.8 \mu\text{m}$ , and  $\pm 4.5 \mu\text{m}$ , respectively. The standard deviations of angular deviation were  $\pm 0.08^\circ \alpha$  axis,  $\pm 0.07^\circ \beta$  axis, and  $\pm 0.09^\circ \gamma$  axis, all of which were better than the design specifications. Long term stability testing shows that after continuous operation for 72 hours, the displacement drift of the system is less than  $1 \mu\text{m}$  and the angle drift is less than  $0.02^\circ$ , meeting the high-precision testing requirements of the production line end.

## **3.3. System integration and standardization framework**

### **3.3.1. Hardware modular design**

The system adopts the decoupling design of “antenna control acquisition” three modules: the MDMA antenna is connected to the RF front-end through the SMA interface to support rapid replacement; The six axis control unit is independently packaged and provides RS485/EtherCAT communication interfaces; The signal acquisition module integrates VNA and spectrum analyzer functions, supporting multi-channel synchronous sampling. Data exchange between modules is achieved through standardized interfaces, and in the event of a single module failure, only the corresponding unit needs to be replaced, reducing maintenance time to less than 30 minutes.

### **3.3.2. Standardization of software interfaces**

The system software adopts a layered architecture, with the underlying drivers encapsulating hardware operations such as motor control and data acquisition, and providing API interfaces for upper layer calls. Integrated SPC statistical process control module in the middle layer, supporting real-time monitoring of path loss and abnormal alarm; The upper level application interface is developed based on the Qt framework, providing functions such as test parameter configuration, result visualization, and report generation. At the same time, the software reserves LabVIEW/Python secondary development interfaces, and users can implement custom testing process development by calling dynamic link libraries or RESTful APIs, improving system scalability.

## **4. Verification and optimization of SPC based testing system**

### **4.1. Application of statistical process control theory**

#### **4.1.1. Control chart design**

To monitor the fluctuation of signal path loss during RF OTA testing, an X-bar/R control chart was used to construct a process stability analysis model<sup>[5]</sup>. The X-bar chart is used to monitor the mean variation of path loss, while the R chart is used to analyze the range fluctuations within the group. In the experiment, 100 sets of data were collected continuously every 20 tests, and the mean  $\bar{X}$  and range R of each set were calculated, and the control limits were plotted. Through Minitab software analysis, it was found that there were no points exceeding the control limit in the X and R plots of the MDMA system test data, and there were no abnormal patterns such as continuous 7 points on the same side of the centerline or trending arrangement, indicating that the testing process was in a statistically controlled state. The fluctuation of path loss was mainly caused by random factors, and the system stability was significantly better than the traditional scheme.

#### **4.1.2. Process capability analysis**

To quantitatively evaluate the stability of the testing system, calculate the process capability indices  $C_p$  and  $C_{pk}$ . Taking the standard deviation of path loss as the key quality characteristic, the upper limit of the specification is set at +1.5 dB and the lower limit is set at -1.5 dB. Through calculation, the MDMA system  $C_p = 2.13$  and  $C_{pk} = 2.08$ , indicating that the system has sufficient process capability and the consistency of test results meets the requirements of mass production; However, traditional systems with  $C_p = 1.32$  and  $C_{pk} = 1.25$  have insufficient process capability and require manual intervention to maintain stability.

### **4.2. Performance verification in complex dynamic environments**

#### **4.2.1. Experimental setup**

To simulate the real interference environment of the production line, a dynamic testing platform is built in a darkroom: metal fixtures are arranged within 1 meter around the tested equipment to introduce reflection interference; The dummy model is driven by a servo motor to reciprocate at a speed of 0.5 m/s within a range of 2 m in front of the DUT, generating Doppler frequency shift interference, and turn on the background noise source of the production line to achieve an ambient noise level of -70 dBm<sup>[6]</sup>. In this environment, 1000 consecutive tests were conducted on the MDMA system and the traditional system, and path loss data was recorded.

#### **4.2.2. Comparison of path loss distribution**

The statistical results show that the average path loss of the MDMA system is similar to that of the traditional system, but the standard deviation is significantly reduced. The standard deviation of the MDMA system is 0.18 dB, which is 40% lower than that of the traditional system. Further analysis of frequency domain characteristics reveals that traditional systems experience a 0.12dB increase in path loss fluctuations in the 5G Hz frequency band due to metal fixture resonance, while MDMA systems effectively suppress backward reflection interference by optimizing the antenna radiation pattern; In the scenario of human movement, the MDMA system maintains an antenna polarization matching degree of > 95% due to high-precision pose compensation of the six axis control model, while traditional systems suffer from polarization mismatch caused by manual alignment errors, resulting in an increase of 0.15 dB in path loss fluctuations.

### **4.3. Real-time compensation model for radio frequency desensitization phenomenon**

#### **4.3.1. Mechanism analysis of Desert phenomenon**

The phenomenon of radio frequency desensitization is mainly caused by harmonic interference and antenna coupling generated during the operation of digital circuits inside the DUT. Experiments have shown that when the CPU operates at a frequency of 2.4G Hz, its third harmonic may fall into the 6G Hz frequency band receiving channel, resulting in a decrease in receiving sensitivity. Through near-field probe scanning, it was found that the electric field strength of the interference signal at the antenna feeding point can reach -50 dBm, exceeding the system noise floor by 45 dB and causing significant performance degradation.

#### **4.3.2. Attenuation compensation algorithm**

Proposing a machine learning-based dynamic threshold adjustment algorithm is crucial. First of all, train historical test data using support vector machines to construct a mapping model between interference intensity and path loss increment; In real-time testing, a spectrum analyzer is used to monitor the interference level in the 6G Hz frequency band. When the interference exceeds the threshold, the compensation module is triggered. Finally,

based on the path loss increment predicted by the SVM model, the amplitude adjustment range of the test signal is dynamically adjusted by a digital attenuator within  $\pm 3$  dB, with a step size of 0.1 dB, to ensure the accuracy of the receiving sensitivity test.

#### **4.3.3. Verification of compensation effect**

Under dynamic interference environment, the CPU runs at full load and conducts 1000 compensation tests on the MDMA system. The results show that the false alarm rate of Desense is 12% without compensation due to abnormal increase in path loss caused by interference. After enabling the compensation algorithm, the false alarm rate decreased to 2.5%, and the compensation response time was less than 50 ms, meeting the real-time testing requirements of the production line. Further analysis revealed that after compensation, the standard deviation of path loss decreased from 0.25 dB to 0.12 dB, and the system's anti-interference ability improved by 52%, verifying the effectiveness of the algorithm.

### **5. Experimental results and discussion**

#### **5.1. Performance indicators of the testing system**

The experimental results show that the MDMA system is significantly better than traditional solutions in terms of space occupation and testing efficiency. In terms of space occupation, the MDMA system compresses the overall volume from  $0.8\text{m}^3$  in traditional systems to  $0.28\text{m}^3$  through integrated six-axis control modules and compact RF front-end design, reducing the proportion by 65% and effectively reducing the demand for site area in production line layout. In terms of testing efficiency, the MDMA system has reduced the single device testing time from 15 seconds in traditional systems to 8 seconds by optimizing the testing process and improving hardware response speed, resulting in a 46.7% increase in efficiency. Further analysis revealed that the improvement in testing efficiency is mainly due to the dual optimization of RF switching time and mechanical motion time, and the system stability has not decreased due to the increase in speed, verifying the effectiveness of integrated design and high-speed control algorithms.

#### **5.2. Practical application cases**

In the NPI stage of a certain brand of smartwatch, the MDMA system demonstrated significant application value. Due to its small antenna size and complex frequency band, the yield rate of traditional testing systems for this product is only 82%. The main problems are excessive path loss fluctuations and Desense misjudgment. After introducing the MDMA system, the high-precision pose compensation and dynamic threshold compensation algorithms of the six-axis control model were used to reduce the standard deviation of path loss to 0.12 dB, the false alarm rate of Desense to 3%, the mass production yield increased to 91%, and the daily production capacity increased by 1200 units. In addition, a 65% reduction in system volume allows the production line to deploy 4 sets of testing equipment simultaneously, further enhancing capacity flexibility.

#### **5.3. Limitations analysis and improvement direction**

The current research still has limitations: the adaptability of MDMA systems in the millimeter wave frequency band needs to be optimized. The experiment found that when the test frequency band was raised to 28G Hz, the resonance effect of the metal fixture increased, resulting in a 0.2 dB increase in path loss fluctuations, and high-frequency signals were more sensitive to mechanical motion errors. Future improvement directions include: using

low-loss millimeter wave materials to reconstruct testing fixtures to suppress high-frequency resonance; Upgrade the six-axis control model to the nanometer level accuracy to meet the stringent requirements for mechanical stability in the millimeter wave frequency band.

## 6. Conclusion

The MDMA RF testing system based on SPC and six-axis control proposed in this study achieves quantitative evaluation of testing process stability through X-bar/R control chart and process capability analysis. Combined with dynamic threshold compensation algorithm, it effectively suppresses the Desense phenomenon, significantly improves testing efficiency and mass production yield, and reduces system volume by 65%, meeting the needs of consumer electronics production lines for high precision, high efficiency, and compactness. Future research will focus on adaptive optimization in the millimeter wave frequency band, further expanding the application boundaries of the system in 5G-A/6G scenarios through low-loss material reconstruction testing fixtures and nanoscale motion control upgrades.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Huang Q, Wang P, Liu M, 2025, Research on Key Technologies and RF Testing of 5G RedCap Terminal. *Electronic Components and Information Technology*, 9(6): 18–20.
- [2] Kong G, Yan S, Du W, et al., 2025, Research on Radio Frequency System Measurement Techniques for Microwave Testing. *Journal of Physics: Conference Series*, 2967(1): 012019.
- [3] Ko J, Kim N, Yoo K, et al., 2023, A Low-Phase-Noise RF Up/Down-Converter for Cost-Effective 5G Millimeter-Wave Test Solutions: Special Section on Microwave and Millimeter-Wave Technologies. *IEICE Transactions on Electronics*, 106(11): 713–717.
- [4] Vlad M, Soroush F, Joe F, et al., 2023, Buried RF Sensors for Smart Road Infrastructure: Empirical Communication Range Testing, Propagation by Line of Sight, Diffraction and Reflection Model and Technology Comparison for 868 MHz–2.4 GHz. *Sensors*, 23(3): 1669.
- [5] Li L, 2017, Research on the Application of SPC in Pneumatic Measurement and Detection System, thesis, Xi'an University of Technology.
- [6] Gao Z, Huang L, Wang C, et al., 2025, Production and Testing Scheme for RF Performance of IoT Terminals Based on Cellular Communication Modules. *IoT Technology*, 15(23): 34–37.

### Publisher's note

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



# Analysis of Fracture Failure of Connecting Bolts for Diaphragm Pump Valve Box Cover

Xuan Qi, Yue Gao

Hainan University of Science and Technology, Haikou 571126, Hainan, 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:** Aiming to address the repeated fracture of connecting bolts in the feed valve box cover of a large Type A diaphragm pump, a three-dimensional model of the valve box assembly was first established. Subsequently, bolt strength was evaluated using three finite element analysis methods: axisymmetric analysis, cyclic symmetry analysis, and full 3D integral analysis. It was clarified that the first few threads bear the maximum stress, which is consistent with the actual fracture position. Combined with the analysis results, it is concluded that the main causes of bolt fracture are alternating load bearing, low safety margin of specifications, and the coarse thread design also affects its load-bearing capacity.

**Keywords:** Diaphragm pump; Valve box cover bolt; Fracture analysis; Finite element analysis; Alternating load

**Online publication:** April 22, 2026

## 1. Introduction

The bolts of the inlet check valve cover (i.e., feed valve cover) of the large Type A diaphragm pump used by a pipeline company have fractured repeatedly, resulting in damage to the internal threads of the screw holes of the check valve chamber (i.e., valve box body). This has caused a great impact on the production of the pipeline company and brought huge potential safety hazards to its equipment. In response to this situation, the causes of bolt fracture and solutions are to be analyzed to restore the original design function of the equipment after repair.

## 2. Establishment of 3D model for valve box assembly parts

The maximum design pressure of the Type A diaphragm pump is 26 MPa, the outer ring diameter at the nominal diameter of the valve box port is 280 mm, the inner ring diameter is 251.2 mm, and the height between the inner and outer rings is 21 mm. Each valve box assembly is equipped with 4 M42 double-ended studs, the diameter at the middle necking part of the stud is 32.6 mm, and the total length of the stud is 265.5 mm<sup>[1]</sup>.

A 3D model of the valve box assembly was established according to the relevant drawings and partial actual dimensions provided by the company<sup>[2]</sup>. The outer ring diameter at the nominal diameter of the valve box port of the valve box 3D model is 270 mm, the inner ring diameter is 210 mm. According to the design pressure of

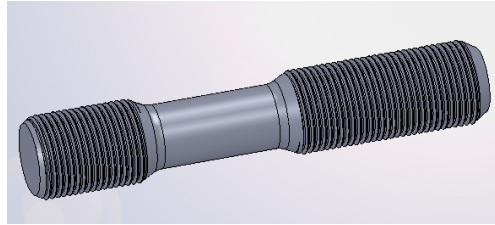


25 MPa, the bolt force on the connecting bolts between the valve box and the valve box cover is calculated as follows:

$$F = \frac{\pi}{4} \times 270^2 \times 25 \div 4 = 357665N$$

## 2.1. Establishment of 3D model for double-ended stud

The double-ended stud model was established in accordance with the GB/T899-1988 standard (**Figure 1**), with the main parameters as follows: nominal diameter:  $d = 42$  mm;  $B_m = 1.5$ ;  $d = 63$  mm;  $B = 96$  mm;  $l = 170$  mm; diameter at the middle necking part: 32.6 mm; total length of the stud: 233 mm<sup>[3]</sup>.

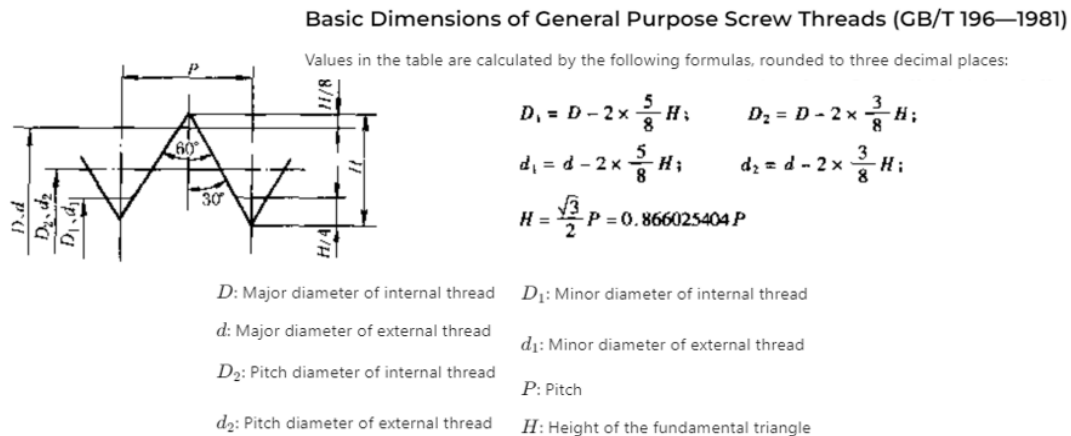


**Figure 1.** 3D model of double-ended stud.

## 2.2. Selection and establishment of thread profile

The connecting thread model was established in accordance with the GB/T196-1981 standard, with a pitch  $P = 3$  mm<sup>[4]</sup>.

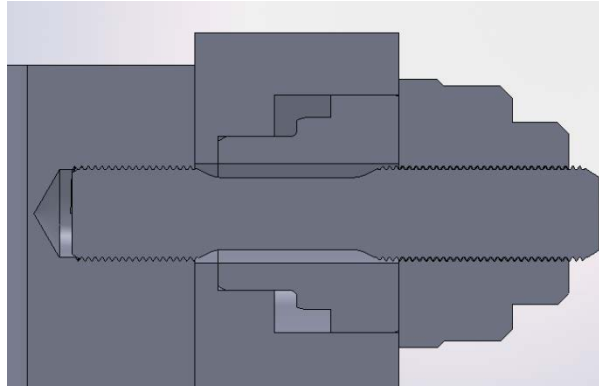
The profile and dimensions of the thread profile feature cutting are shown in **Figure 2**.



**Figure 2.** GB/T196-1981 general thread standard.

## 2.3. Assembly of feed valve box components

Considering that excessive thread contact will cause a serious decline in calculation speed, and most bolt fracture positions are at the end where the stud is screwed with the internal thread of the valve box, the two inner and outer nuts on the outer side of the valve cover were simplified into one nut, and the positions without actual dimensions of the valve box and valve cover were rounded<sup>[5]</sup>. The overall assembly of the feed valve box and the bolt connection section view of the established integral assembly model are shown in **Figure 3**.



**Figure 3.** Overall assembly of feed valve box and section view of bolt connection.

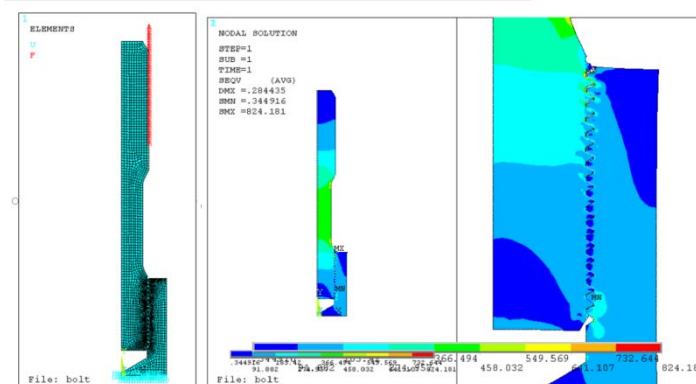
### 3. Finite element strength analysis methods for connecting bolts of feed valve box cover

The main work of finite element analysis for bolt connection problems lies in the mesh division at the thread profile and the establishment of contact pairs between internal and external threads. How to divide sufficiently fine meshes at the thread profile while avoiding an excessively large overall number of meshes under the condition of increased calculation time caused by contact pairs, thus ensuring good calculation speed and results, has become a key factor in bolt connection analysis<sup>[6]</sup>.

In view of this, this paper adopts three methods: axisymmetric analysis, cyclic symmetric analysis and 3D analysis with bolt mesh division by a certain software, to calculate and analyze the simplified structure of bolt connection. That is, the model at the connection position between the double-ended stud and the valve box is extracted, the thread features at the connection between the double-ended stud and the nut are simplified, and the bolt force is used instead for analysis.

#### 3.1. Axisymmetric analysis method and results

The axisymmetric model was meshed with PLANE42 elements, the mesh size at the thread profile was set to 0.5 mm, and the overall mesh size was set to 2 mm, with a total of 3834 elements and 3926 nodes divided. The mesh and loading type are shown in **Figure 4**. Only the valve box model was fixed, and the outer side of the stud was subjected to bolt tension for analysis<sup>[7]</sup>. The nephogram of analysis results is as follows.



**Figure 4.** Mesh and load diagram of axisymmetric analysis for bolt connection, stress nephogram of axisymmetric analysis.

The maximum equivalent stress is  $\sigma = 824.181$  MPa, and the maximum deformation is  $\delta = 0.284$  mm. The maximum contact pressure on the double-ended stud reaches 364.65 MPa, while the stress at the necked middle section is 445.553 MPa. The maximum deformation at the threaded region is 0.058 mm.

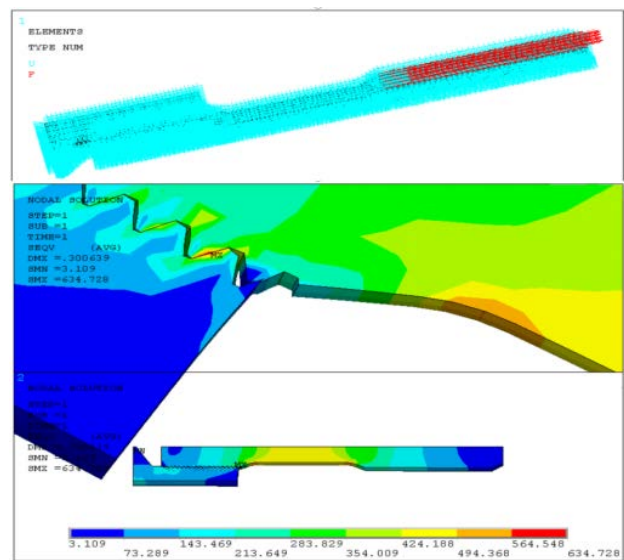
Due to the pronounced non-uniform load distribution among threads, the first few threads typically carry a significantly higher load. Therefore, the stresses in the first seven threads were extracted and compared, as presented in **Table 1**.

**Table 1.** Stress comparison table of each thread circle in axisymmetric analysis

Thread position	1st circle	2nd circle	3rd circle	4th circle	5th circle	6th circle	7th circle
Equivalent stress (MPa)	824.181	566.552	451.625	398.963	380.250	357.260	300.901

### 3.2. Cyclic symmetric analysis method and results

The cyclic symmetric model was meshed with SOLID45 elements, the mesh size at the thread profile was 0.5 mm, and the overall mesh size was 2 mm, with a total of 16501 elements and 13001 nodes divided. The analysis working condition is the same as that of the axisymmetric analysis. The mesh and loading type are shown in **Figure 5**, and the nephogram of analysis results is as follows.



**Figure 5.** Load and constraint diagram of cyclic symmetric analysis for bolt connection, stress nephogram of cyclic symmetric analysis.

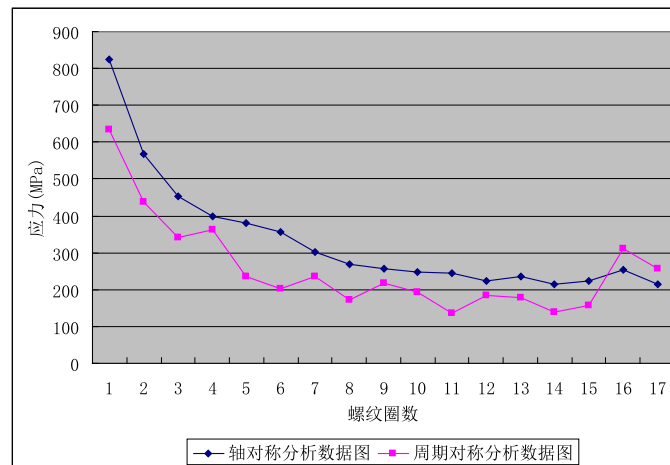
The maximum equivalent stress is  $\sigma = 634.728$  MPa, and the maximum deformation is  $\delta = 0.301$  mm. The maximum contact pressure on the double-ended stud is 284.936 MPa, while the stress at the necked middle section is 449.457 MPa. The maximum deformation in the threaded region is 0.061 mm.

The stress distribution in the first seven threads is presented in **Table 2**.

**Table 2.** Stress comparison table of each thread circle in cyclic symmetric analysis

Thread position	1st circle	2nd circle	3rd circle	4th circle	5th circle	6th circle	7th circle
Equivalent stress (MPa)	634.728	436.988	340.867	362.615	236.638	202.846	235.418

A curve graph of the stress at each thread circle from axisymmetric and cyclic symmetric analysis was drawn, as shown in **Figure 6**.



**Figure 6.** Stress comparison curve of each thread circle from axisymmetric and cyclic symmetric analysis.

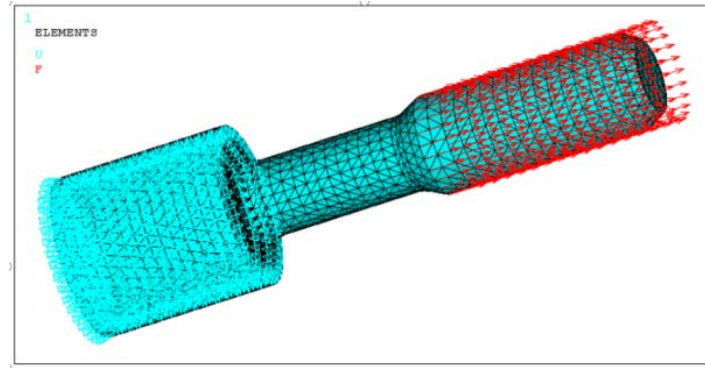
It can be seen that the bolt stress change trends obtained by axisymmetric and cyclic symmetric analysis are basically the same. The curve smoothness of the axisymmetric analysis is better than that of the cyclic symmetric analysis, but the overall values are higher than those of the cyclic symmetric analysis. The reason may be that the axisymmetric model is meshed with tetrahedral elements, whose mesh quality is worse than that of the quadrilateral elements in the cyclic symmetric analysis, so the degree of element stress concentration may be slightly higher, resulting in higher values<sup>[8]</sup>.

Since the bolt stress is the largest at the thread circle with the maximum tension, which matches the nut thread circle with the maximum pressure, the uneven load distribution on the thread is increased. According to the force distribution diagram of each thread circle for the ten-thread nut proposed by N.E. Zhukovsky, the first thread circle bears the maximum load, accounting for about 1/3 of the total force on the screw, while the tenth thread circle at the end bears less than 1/100 of the total force. Due to the thread circle deformation caused by thread profile error, contact deformation and local plastic deformation, the load on the first thread circle is reduced to 1/5 or 1/4 of the total force.

It can be seen from **Figure 6** that the curve shape is basically in line with the law of N.E. Zhukovsky's thread load distribution diagram, and the curve formed by the axisymmetric analysis results is more continuous<sup>[9]</sup>.

### 3.3. Integral analysis method and results for bolt connection

The 3D integral model of bolt connection was taken, and meshed with SOLID45 elements by a certain analysis software, with the mesh size at the thread profile set to 0.5 mm and the overall mesh size set to 2 mm, with a total of 231508 elements and 54043 nodes divided. The analysis working condition is the same as that of the axisymmetric analysis. The load and constraint diagram is shown in **Figure 7**.



**Figure 7.** Load and constraint diagram of integral analysis for bolt connection.

After the 3D integral analysis of bolt connection, it was found that for different thread engagement positions, two different stress trends appeared: one is a nephogram with equidistant stress intensity law, and the other is a stress nephogram with normal distribution<sup>[10]</sup>. In response to this situation, different thread engagement angles were analyzed, and the basic law between the thread head-tail coincidence angle and the stress trend change was obtained. The detailed analysis content and results are shown in **Table 3**.

**Table 3.** Relationship between thread head-tail coincidence angle and stress trend change

Serial number	Coincidence angle (°)	External rotation angle (°)	Stress concentration strip area
1	90	0	6
2	60	30	0
3	30	60	6
4	0	90	0
5	270	180	6
6	180	270	0
7	45	45	6

The analysis results show that excluding the stress concentration points at the thread head-tail contact in bolt analysis, the maximum stress is between 800 and 1000 MPa, and the position is mostly at the first contact thread at the bolt root; the stress at the bolt head (far from the necking end) is large, and the stress at the middle part is small<sup>[11]</sup>.

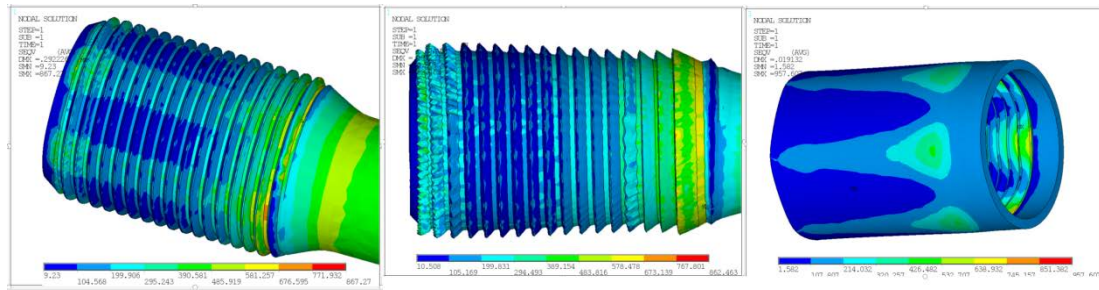
In the 3D finite element analysis of bolt connection, stress concentration strip areas are likely to appear in the analysis results; there is no stress concentration strip area when rotating 30°, and the strip area appears again when rotating another 30°. In addition, the initial thread head-tail coincidence is 90°, and the area within 30° of the contact entry point is a low-stress area; this stress change law does not change with the change of bolt tension;

One side of the thread connection was set as a rigid body to observe the stress state of the other side. It was found that when the bolt external thread is rigid, the internal stress value of the valve box internal thread is the largest and gradually decreases outward; when the valve box internal thread is rigid, the stress value at the bolt external thread root is the largest and gradually decreases outward, and the maximum stress value reaches more than 1300 MPa<sup>[12]</sup>.

The mesh of the first few thread models with large stress was refined and submodel analysis was carried out,



and it was found that the stress increased slightly and the trend remained unchanged (**Figure 8**).



**Figure 8.** Stress nephogram with stress concentration strip area in bolt stress trend, stress nephogram without stress concentration strip area after adjusting thread head-tail coincidence angle, and bolt stress nephogram when internal and external threads are set as rigid bodies respectively.

#### 4. Overview of bolt fracture accident analysis

Due to the repeated fluctuation of the internal pressure of the feed valve box between the feed pressure and the discharge pressure, the connecting bolts of the feed valve box cover bear alternating loads, which causes more serious damage to the bolt material than the bolts of the discharge valve box cover, leading to bolt damage and even fatigue fracture<sup>[13]</sup>.

From the bolt strength analysis results, the first few threads usually bear relatively large stress, and the maximum stress is generally at the first complete thread circle, which is close to the actual fracture position of the bolt<sup>[14]</sup> (**Figure 9**).



**Figure 9.** Comparison diagram of bolt fracture positions.

The imported bolts used in on-site diaphragm pumps are generally coarse-pitch threads, and the proof load of fine-pitch threads with the same diameter is slightly higher than that of coarse-pitch threads<sup>[15]</sup>.

#### 5. Conclusion

According to the bolt specification selection calculation formula in domestic references, the specification of the connecting bolts for the feed valve box cover of the on-site diaphragm pump should be above M56. Compared with the strength analysis results of the diaphragm chamber cover bolts of the on-site diaphragm pump, it can be inferred that even if the bolt specification selection calculation formula of foreign standards is used to select the



feed valve box cover bolts, the safety margin of the bolts used for the feed valve box cover is still low.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Editorial Board of Mechanical Design Handbook, 2020, Mechanical Design Handbook: Volume 2 Fasteners, 6th Edition, China Machine Press, Beijing.
- [2] State Bureau of Technical Supervision, 1988, GB/T899-1988 Double-Ended Studs  $b_m=1.5d$ , China Standard Press, Beijing.
- [3] State Bureau of Technical Supervision, 1981, GB/T196-1981 Basic Dimensions of General Purpose Metric Screw Threads, China Standard Press, Beijing.
- [4] Wang X, Shao M, 2003, Basic Principles and Numerical Methods of Finite Element Method, 2nd Edition, Tsinghua University Press, Beijing.
- [5] Liu H, 2017, Mechanics of Materials (I), 6th Edition, Higher Education Press, Beijing.
- [6] Gou W, Jin B, Wei F, 2008, Mechanical Fatigue and Fracture, Northwestern Polytechnical University Press, Xi'an.
- [7] Zhang F, 2014, Design and Application of Threaded Connections, China Machine Press, Beijing.
- [8] Wang G, 2008, Practical Engineering Numerical Simulation Technology and Its Application in ANSYS, 2nd Edition, Northwestern Polytechnical University Press, Xi'an.
- [9] Editorial Board of Mechanical Engineering Materials Handbook, 2019, Mechanical Engineering Materials Handbook: Metal Materials Volume, 5th Edition, Chemical Industry Press, Beijing.
- [10] Zheng J, Dong Q, Sang Z, 2015, Process Equipment Design, 4th Edition, Chemical Industry Press, Beijing.
- [11] Yu W, Gao B, 2017, Application of ANSYS in Mechanical and Chemical Equipment, 3rd Edition, China Water & Power Press, Beijing.
- [12] He X, 2011, Mechanical Connection Technology, China Machine Press, Beijing.
- [13] National Technical Committee for Pump Standardization, 2021, JB/T6900-2021 Diaphragm Pumps, China Machine Press, Beijing.
- [14] Cheng D, 2017, Handbook of Taboos in Mechanical Design, 2nd Edition, Chemical Industry Press, Beijing.
- [15] Li H, 2011, Hydraulic Components and Systems, National Defense Industry Press, Beijing.

### Publisher's note

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

# Experimental Investigation of the Velocity Distribution in the Tail Flame of an Inductively-Coupled-Plasma

Xiaobao Mao

China West Normal University, Nanchong 637002, Sichuan, 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:** To achieve an accurate description of an Inductively Coupled Plasma (ICP) source, numerous two-dimensional and three-dimensional numerical models have been developed. However, experimental validation of these models remains a major challenge. Compared with plasma temperature and electron number density, the measured gas flow velocity distribution is more direct and reliable, and can serve as an important criterion for assessing model validity. In this work, the experimental method was improved by optimizing the observation parameters of a high-speed camera, effectively suppressing the interference from the intense emission spectra in the normal analytical zone. For the first time, ion clouds formed by injected particles were directly observed. Five types of suspended particles were sequentially introduced, including  $\text{Er}_2\text{O}_3$ ,  $\text{Y}_2\text{O}_3$ , and borosilicate glass particles with diameters of 10, 5, and 2  $\mu\text{m}$ . The particle-flow following behavior was comparatively evaluated. Using a tracer method, the gas flow velocity distribution in the ICP tail plume was measured. Furthermore, with  $\text{Y}_2\text{O}_3$  as the tracer, the axial gas velocity distribution in the central channel was systematically measured under different radio-frequency (RF) powers and carrier gas flow rates. The results show that within the range of  $2.5 \text{ mm} < z < 12.5 \text{ mm}$ , the axial gas velocity in the central channel of the tail plume exhibits a distinct plateau region. The axial gas velocity increases with increasing RF power, while showing weak sensitivity to variations in carrier gas flow rate. The present study provides experimental data on the axial gas velocity distribution, offering essential validation and correction benchmarks for numerical ICP models.

**Keywords:** Inductively coupled plasma; Axial velocity distribution; Ar-ICP; High-speed camera

**Online publication:** April 22, 2026

## 1. Introduction

Inductively coupled plasma (ICP) sources, as crucial energy sources and ion sources, are widely used in analytical instruments such as Inductively Coupled Plasma Mass Spectrometry (ICP-MS) and Inductively Coupled Plasma Optical Emission Spectrometry (ICP-OES) <sup>[1,2]</sup>. For ICP sources intended for analytical purposes, fundamental research typically focuses on the spatial distribution characteristics of internal plasma parameters (e.g., electron

temperature, ion temperature, gas flow velocity) and their main influencing factors. The relevant research methods are mainly divided into two categories: experimental diagnostics and numerical simulation. In terms of numerical simulation, two-dimensional axisymmetric models based on commercial software platforms such as ANSYS Fluent and COMSOL Multiphysics are now relatively mature and can effectively describe the fundamental behavior of the plasma. However, these two-dimensional models have significant limitations when dealing with actual three-dimensional complex structures, such as accurately representing the geometric asymmetry of the actual torch and load coil, the three-dimensional flow effects caused by the tangential injection of the cooling gas, and the entrainment and mixing process between the plasma and the surrounding gas. Consequently, in recent years, several research teams have successively conducted three-dimensional numerical simulation studies, aiming to more realistically reflect the underlying physical and chemical processes<sup>[3-6]</sup>. Nevertheless, systematic experimental validation of three-dimensional simulation results remains relatively scarce. Common validation methods primarily focus on two aspects: one is the qualitative or quantitative comparison of the simulated flow field structure with plasma vortex morphology observed using custom-built ICP sources equipped with high-speed cameras or schlieren systems; the other is the comparison of simulated plasma parameters (e.g., temperature, velocity) with two-dimensional distribution data obtained via diagnostic techniques such as optical emission spectroscopy, laser Thomson scattering, or Rayleigh scattering<sup>[3,4,7,8]</sup>. While these comparative efforts provide important foundations for model validation, the development of more comprehensive and refined three-dimensional experimental diagnostic methods is still necessary to enhance the reliability and predictive capability of simulation results.

Previous studies have employed various optical and time-resolved diagnostic techniques to experimentally measure the flow velocity distribution inside the ICP torch and within the normal analytical zone, primarily including high-speed photography, Particle Image Velocimetry (PIV), and Time-of-Flight (TOF) methods<sup>[9-12]</sup>. Due to the intense continuous emission background in the upstream region of the Normal Analytical Zone (NAZ, < 20 mm above the load coil), the signal from ion clouds is difficult to extract effectively. Consequently, reliable experimental data on the velocity distribution in this region are still lacking. Overall, systematic flow velocity measurement data covering the complete space from the torch exit to the downstream region for the same ICP device remain very limited, which hinders the comprehensive validation and further optimization of current ICP numerical models.

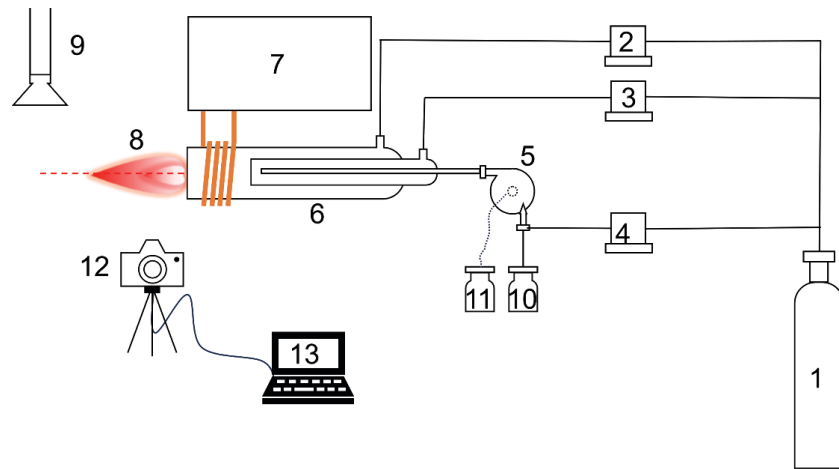
This work conducts a systematic experimental investigation of the normal analytical zone of an ICP based on a custom-built ICP device. By optimizing the observation parameters of a high-speed camera, ion clouds from sampled particles were directly observed, and the flow velocity distribution of the pure argon plasma was obtained. By varying the plasma RF power and sample gas flow rate, the corresponding gas flow velocity distributions were measured. The results were compared with simulations from a previously established two-dimensional numerical model to validate the experimental data and elucidate the flow mechanisms.

## **2. Experimental methodology**

### **2.1. Experimental apparatus and reagents**

The experimental setup consists of an RF generator, an impedance matching network, a load coil, a micro nebulizer, and a cyclonic spray chamber. The flow rates of the coolant, auxiliary, and sample gases are precisely controlled by mass flow controllers, and high-purity argon (99.999%) is used. Unless otherwise specified, the ICP

operating conditions were set to W, L/min, L/min, and L/min. The schematic diagram of the system is shown in **Figure 1**.



**Figure 1.** Schematic diagram of the facility (1: Argon gas cylinder, 2: MFC for coolant gas, 3: MFC for auxiliary gas, 4: MFC for sample gas, 5: Nebulizer and cyclone chamber, 6: Quartz torch, 7: r.f. power generator and impedance matching network, 8: Plasma, 9: Exhaust pipe, 10: Suspension sample, 11: Waste, 12: High-speed colour camera, 13: Computer).

Five types of particulate samples were used in this study, including ground  $\text{Er}_2\text{O}_3$  and  $\text{Y}_2\text{O}_3$  oxide powders, and borosilicate glass particles with nominal diameters of approximately 10, 5, and 2  $\mu\text{m}$  (denoted as BSG\_10, BSG\_5, and BSG\_2, respectively). The particle size and morphology were examined using a digital microscope. The powders were dispersed and diluted in ultrapure water to prepare suspensions of appropriate concentrations for analysis. To facilitate comparison of the aerodynamic behavior of different particles, their aerodynamic diameters were calculated according to:

$$D_a = D_p \sqrt{\rho} \quad (1)$$

where ( $D_p$ ) is the geometric particle diameter and ( $\rho$ ) is the particle material density.

As shown in **Table 1**, the aerodynamic equivalent diameters of these particles range from 3.3 to 35  $\mu\text{m}$ , which is sufficient to evaluate their ability to follow the gas flow.

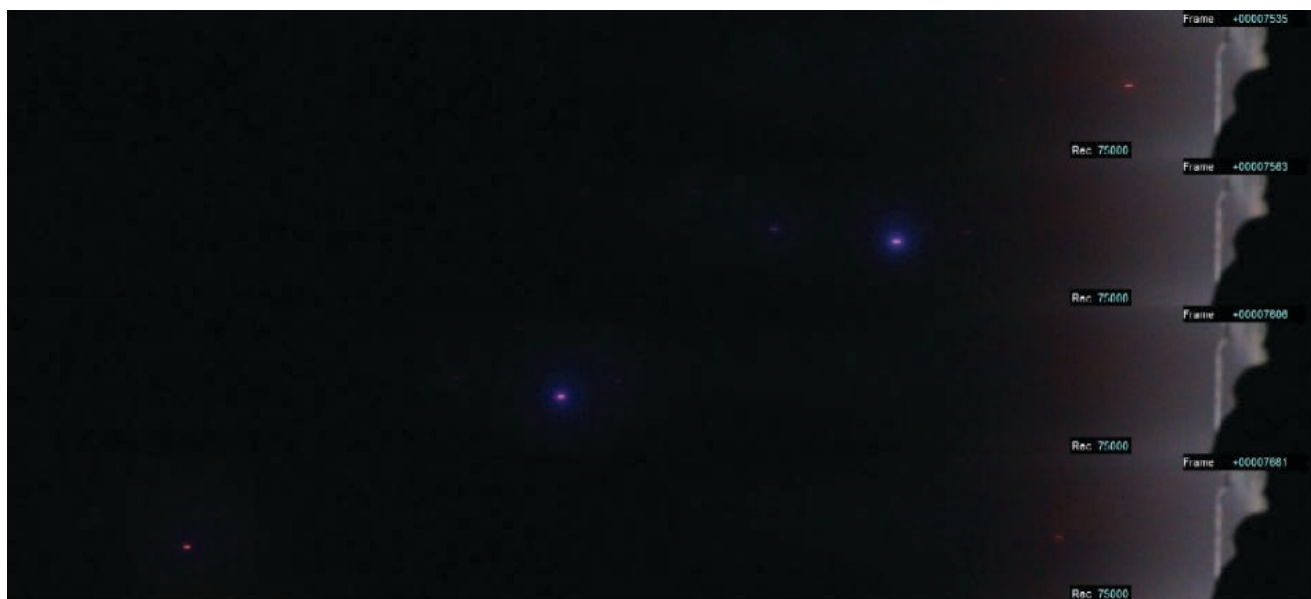
**Table 1.** Summary of particles used for flow tracking

No.	Name	Dno/ $\mu\text{m}$	Dp/ $\mu\text{m}$	$\rho/\text{g cm}^{-3}$	Da/ $\mu\text{m}$	Source
1	$\text{Er}_2\text{O}_3$ powder	/	11.78	8.64	34.63	Self-grinded
2	$\text{Y}_2\text{O}_3$ powder	/	9.53	5.03	21.37	
3	BSG_10	10.2	10.73	2.55	17.14	Thermo Fisher
4	BSG_5	4.6	5.48	2.55	8.75	
5	BSG_2	2.0	2.11	2.50	3.34	
/	Graphite powder	/	9.41	2.27	14.18	Guo <i>et al.</i> 2019

## 2.2. Measurement of plasma tail flame velocity with different sampled particles

To investigate the influence of particle physical properties on ICP tail flame velocity measurements, suspensions

containing different particles were sequentially nebulized under fixed RF power and carrier gas flow conditions. After entering the plasma, the particles underwent heating, evaporation, and ionization, forming luminous particle clusters. A color high-speed camera (75,000 fps,  $1280 \times 144$  pixels) was used to record their motion. By optimizing the aperture and exposure time, high signal-to-noise particle trajectory images were obtained. **Figure 2** shows four consecutive frames captured in the central channel during nebulization of a  $Y_2O_3$  suspension at 1200 W RF power and 1.0 L/min carrier gas flow rate. The bright streaks correspond to the trajectories of individual particle clusters during the exposure time. Using this approach, the particle tracing method was applied to determine the velocity distribution in the Ar-ICP tail flame.



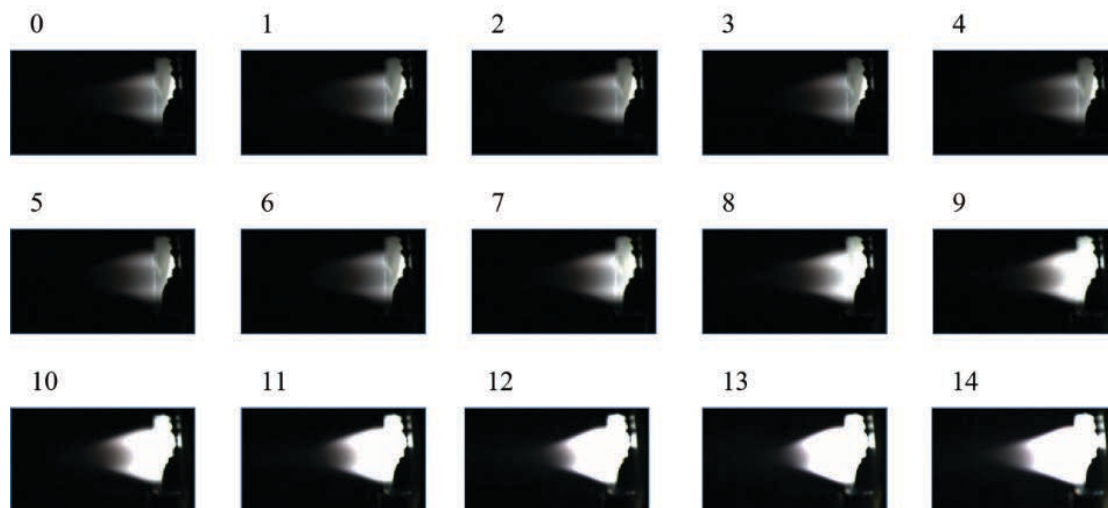
**Figure 2.** High-speed images showing the motion trajectories of  $Y_2O_3$  particles in the central channel of the plasma.

### 3. Results and discussion

#### 3.1. Influence of camera aperture and shutter speed on ICP tail flame imaging

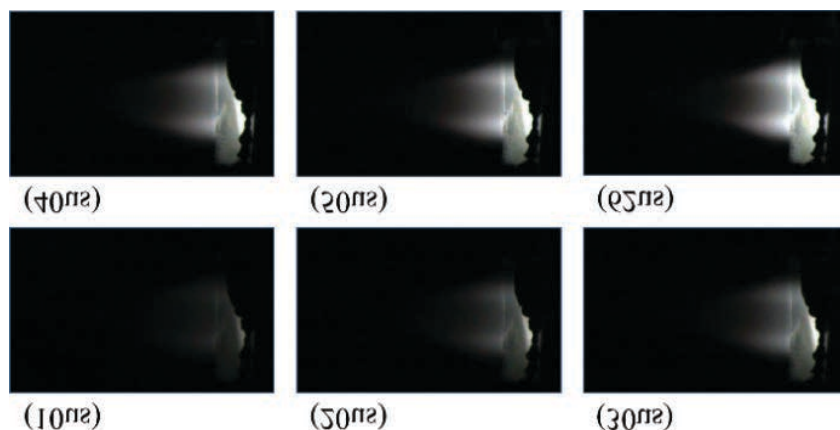
In previous studies, the aperture and shutter parameters of the high-speed camera were not systematically optimized, resulting in excessive brightness of the ICP tail flame near the torch exit, which made it difficult to obtain reliable velocity distribution data in this region. To reduce the tail flame signal intensity and achieve clear imaging of the normal analytical zone near the torch outlet, the imaging conditions were optimized by carefully adjusting the camera aperture and shutter settings. The aperture controls the amount of incident light, whereas the shutter determines the exposure time. After stable plasma ignition, transient images of the tail flame under pure argon (Ar) conditions were recorded using a color high-speed camera (MEMRECAM ACS-3, NAC Co., Japan). By systematically varying the aperture-shutter combinations, clear images of the plasma in the central channel were successfully obtained.

To investigate the effect of aperture on tail flame imaging, the shutter time was fixed at  $20 \mu s$  while the aperture setting was varied. The results are shown in **Figure 3**. The image brightness increased significantly with increasing aperture number, indicating that the aperture setting effectively reflects the degree of aperture opening.



**Figure 3.** The pictures of pure Ar-ICP captured by the camera operating at 20 $\mu$ s shutter duration and varying apertures. The number represents the aperture index.

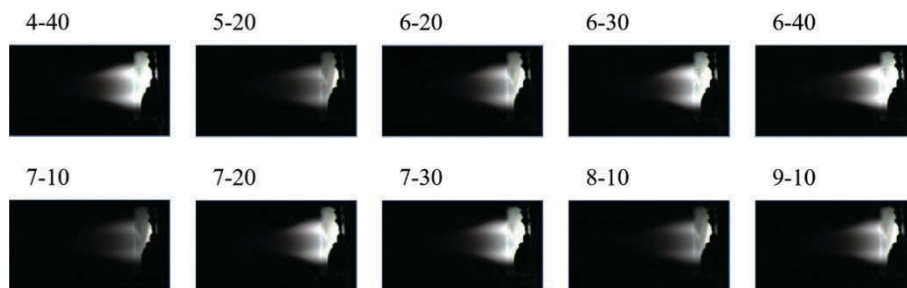
To investigate the effect of shutter speed on tail flame imaging, the camera aperture was fixed at setting 0, and the shutter time was sequentially adjusted to 10, 20, 30, 40, 50, and 62  $\mu$ s (maximum value). The results are shown in **Figure 4**. The image brightness increased significantly as the shutter speed decreased, indicating that longer exposure times resulted in stronger collected light signals.



**Figure 4.** The pictures of pure Ar-ICP captured by the camera operating at aperture “0” and varying shutter duration. The number indicates the shutter value.

Since both the camera aperture and shutter speed affect the image brightness near the torch outlet, coordinated adjustment of these two parameters is required when imaging the tail flame. To clearly capture the plasma central channel at the torch exit and obtain high-contrast images of particle trajectories, the aperture-shutter combinations were systematically optimized. Several parameter settings suitable for central channel visualization were ultimately identified. **Figure 5** shows representative examples of aperture and shutter combinations that clearly reveal the plasma central channel.





**Figure 5.** Morphology of the pure Ar-ICP tail flame under different aperture and shutter settings (labels indicate aperture number-shutter time combinations).

### 3.2. Influence of different particles on ICP tail flame velocity distribution measurements

The suspension containing  $\text{Er}_2\text{O}_3$  powder was first introduced into the ICP torch via nebulization through the carrier gas. Using the known outer diameter of the torch outer tube ( $D = 20$  mm) and its corresponding pixel dimension in the image, the spatial resolution of the camera was calibrated to be  $27.58 \mu\text{m}/\text{pixel}$ . By adjusting the aperture and shutter parameters, clear motion trajectories of the ion clouds generated from  $\text{Er}_2\text{O}_3$  particles were successfully captured. Subsequently, the torch was rinsed with 2%  $\text{HNO}_3$  solution followed by ultrapure water to eliminate residual  $\text{Er}_2\text{O}_3$ . The camera spatial resolution was recalibrated, yielding a value of  $28.62 \mu\text{m}/\text{pixel}$ . A  $\text{Y}_2\text{O}_3$  suspension was then nebulized, and the motion trajectories of the corresponding ion clouds were recorded. The cleaning, calibration, and nebulization procedures were repeated to obtain trajectory images for borosilicate glass particles with nominal diameters of  $10 \mu\text{m}$ ,  $5 \mu\text{m}$ , and  $2 \mu\text{m}$ . To improve positioning accuracy, the particle positions in consecutive frames were manually identified. To enhance processing efficiency, dedicated code was developed based on the MATLAB Image Processing Toolbox to assist in batch extraction of particle positions. The particle velocities were finally calculated from the displacement between two consecutive frames divided by the corresponding time interval.

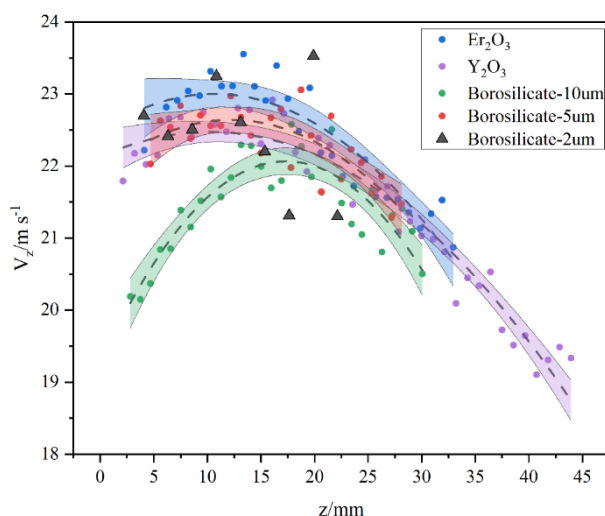
The ion cloud trajectories corresponding to different particles are shown in **Figure 6**. The elongated streaks observed in the images result from motion blur caused by the limited frame rate of the camera. In velocity calculations, the center position of each ion cloud was extracted, and the velocity was determined by averaging the displacement over five consecutive frames. This approach effectively reduced the influence of motion blur on positioning accuracy, ensuring the reliability of the obtained velocity data.



**Figure 6.** Experimentally captured ion cloud trajectories of different particles in the ICP tail flame: (a)  $\text{Er}_2\text{O}_3$ , (b)  $\text{Y}_2\text{O}_3$ , (c) BSG\_10, (d) BSG\_5, and (e) BSG\_2.

In high-speed imaging observations, tracer particles with different chemical compositions exhibited distinct emission characteristics in the ICP tail flame, which directly affected their suitability as flow tracers. As shown in **Figure 6b**, the nebulized  $\text{Y}_2\text{O}_3$  particles underwent evaporation, ionization, and excitation in the high-temperature plasma, forming stable pale pink luminous clusters. This color originates from characteristic emission lines of neutral or singly ionized yttrium species,  $\text{Y(I)}$  or  $\text{Y(II)}$ . In contrast,  $\text{Er}_2\text{O}_3$  particles displayed bright white emission (**Figure 6a**), which is closely associated with the rich emission spectrum of erbium. The emissions of both particle types were stable and persistent, forming clear motion trajectories suitable for subsequent velocity calculations. The borosilicate glass particles used in this study are multicomponent systems primarily composed of  $\text{B}_2\text{O}_3$ - $\text{SiO}_2$ , with additional oxides such as  $\text{Na}_2\text{O}$  and  $\text{K}_2\text{O}$ . As shown in **Figure 6c–6e**, the observed yellow and purple emissions mainly originate from Na and K spectral lines, respectively.

The motion trajectories of ion clouds generated from injected particles were successfully recorded. For each particle type, the axial position ( $z$ ) and radial position ( $r$ ) of 60 ion clouds at different time instants were extracted to calculate the axial velocity  $V_z$ . The polynomial fitting curves of  $V_z$  versus axial position for different particles are presented in **Figure 7**.



**Figure 7.** Axial velocity distributions of different particles in the ICP tail flame.

As shown in **Figure 7**, the measurements for  $\text{Er}_2\text{O}_3$  and  $\text{Y}_2\text{O}_3$  particles indicate that within the axial range of  $z = 2.5$ – $12.5$  mm, the velocity remains at a high plateau. This region corresponds to the standard analytical zone commonly used in ICP-MS or ICP-OES, where flow stability plays a crucial role in sample transport efficiency. In the downstream region ( $z > 12.5$  mm), the velocity decreases significantly due to momentum dissipation caused by entrainment of surrounding air. The axial velocity of  $10\text{ }\mu\text{m}$  borosilicate glass particles (BSG\_10) is slightly lower than that of the other four particle types.

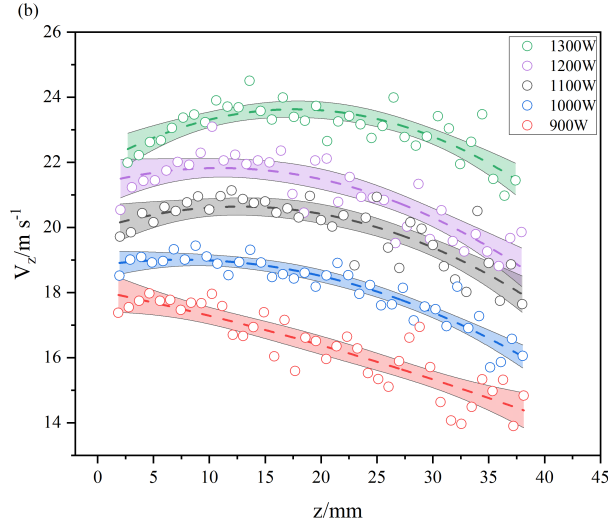
To ensure the accuracy and stability of velocity measurements in subsequent parametric studies (e.g., varying RF power or sample gas flow) and to maintain comparability across different experimental conditions,  $\text{Y}_2\text{O}_3$  particles were selected as the standard tracer for all following experiments.

### 3.3. Effects of varying parameters on the axial velocity distribution in the ICP tail flame

#### 3.3.1. The RF power

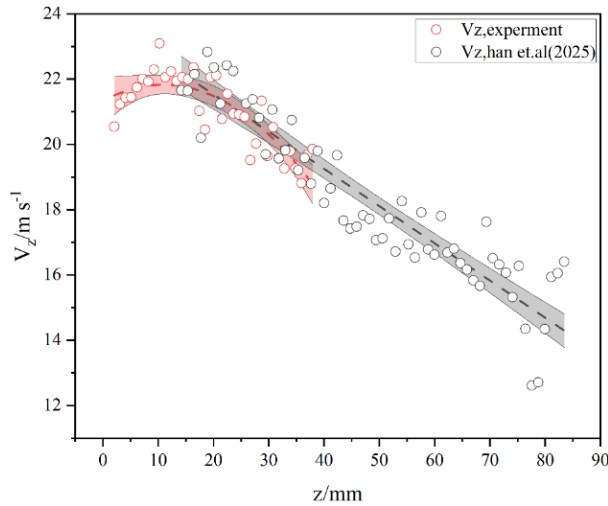
With the sample gas flow rate fixed at  $1\text{ L/min}$ , the RF power was varied from  $900$  to  $1300\text{ W}$ . The motion

trajectories of ion clouds generated from nebulized  $Y_2O_3$  particles were recorded, and the corresponding axial velocity distributions  $V_z$  are shown in **Figure 8**. The measured  $V_z$  increases with RF power at all axial positions, indicating enhanced jet momentum with higher energy input. In the downstream region ( $z > 12.5$  mm),  $V_z$  decreases gradually with axial distance due to jet expansion, turbulent mixing, and momentum dissipation in ambient air.



**Figure 8.** Axial velocity distributions along the plasma centerline under different RF powers.

As shown in **Figure 9**, under RF power of 1200 W and a sample gas flow rate of 1 L/min, the linear velocity distribution obtained in previous work over the axial range of 12.5–85 mm from the torch outlet is compared with the results measured in the present study over 0–40 mm.



**Figure 9.** Comparison of axial velocities in the ICP tail flame between Han *et al.* and the present study.

A polynomial fit of the experimental  $V_z$  data in the present study yields:

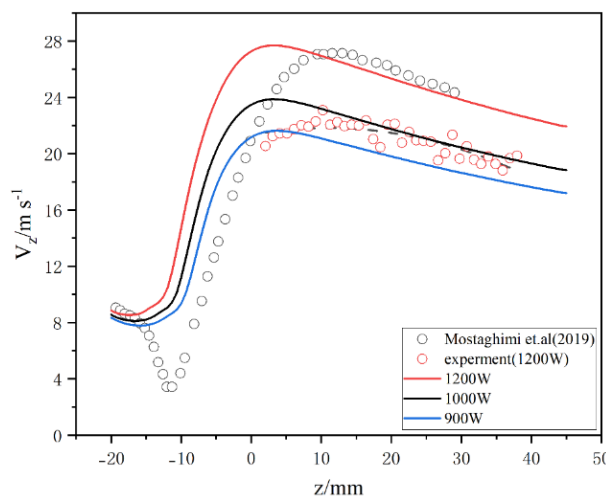
$$V_z = 21.7 + 0.133z - 0.003z^2 \quad (2)$$

In comparison, the linear fit from previous work over the axial range of 12.5–85 mm is <sup>[12]</sup>:

$$V_z = 23.8 - 0.114 z \quad (3)$$

At  $z = 12.5$  mm, the fitted value in this study is 22.89 m/s, which is in good agreement with the previous value of 22.38 m/s (relative deviation  $\approx 2.3\%$ ). This indicates good consistency between the two datasets in the overlapping measurement region and validates the reliability of the present experimental method.

Taking the case of RF power at 1200 W as an example, the experimentally measured axial velocity distribution of particles was fitted and compared with the particle trajectory evolution obtained from numerical simulations by Mostaghimi *et al.* as well as the simulated curves from the 2D model at different powers, as shown in **Figure 10** <sup>[11]</sup>. The results indicate that, consistent with Mostaghimi's conclusions, the particle axial velocity exhibits a clear plateau in the downstream region of the torch outlet over the axial range of  $2.5 \text{ mm} < z < 12.5 \text{ mm}$ .

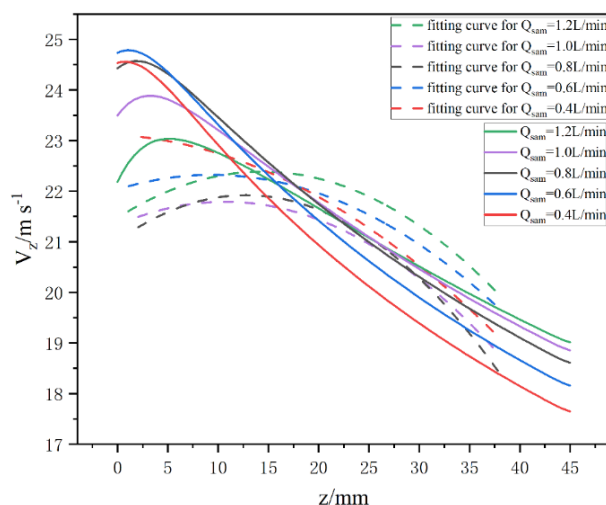


**Figure 10.** Axial particle velocities in the ICP tail flame: Comparison between experimental measurements, polynomial fit, and simulations.

This velocity plateau region generally corresponds to the Normal Analytical Zone (NAZ) of the plasma and is closely associated with the relatively high and stable central temperature distribution within this region. As the axial distance increases further, both particle and gas axial velocities gradually decrease due to the progressive entrainment of ambient air and the continuous dissipation of momentum and energy.

### 3.3.2. Sample flow rate

Under a fixed RF power of  $P = 1200$  W, the sample gas flow rate was set sequentially to 1.2, 1.0, 0.8, 0.6, and 0.4 L/min. Using the same processing method as for the varying-power experiments, the axial velocities ( $V_z$ ) were calculated. **Figure 11** presents the polynomial fits of the centerline velocities for different sample gas flow rates, along with the corresponding centerline velocity curves predicted by the 2D numerical model for the same flow conditions.



**Figure 11.** Comparison of experimental and simulated centerline velocity distributions at different sample gas flow rates in the ICP tail flame.

**Figure 11** shows that the measured axial velocity distributions in the ICP tail flame at different sample gas flow rates indicate a limited influence of flow rate on the overall velocity level. This suggests that in the near-field region of the tail flame close to the torch outlet, the axial flow is primarily driven by gas expansion and pressure gradients induced by plasma heating, rather than by the cold gas volumetric flow itself.

Notably, in the region  $z < 12.5$  mm, the axial velocity measured at lower flow rates is slightly higher than that at higher flow rates. Under a constant RF power, a smaller gas flow implies higher heating power per unit mass, leading to an increased core plasma temperature and stronger gas expansion, which enhances axial jetting. Conversely, at higher flow rates, the additional cold gas entering the plasma region exerts a cooling and dilution effect on the hot core, reducing near-field acceleration and lowering the axial velocity peak.

Therefore, in the near-field region of the ICP tail flame, axial velocity is far more sensitive to RF power than to sample gas flow rate. The higher near-field velocities observed at lower flow rates further confirm that flow in this region is dominated by thermal driving effects.

## 4. Conclusion

In this work, a custom ICP source was employed, and the axial velocity in the tail flame within the normal analytical zone (NAZ) of the plasma was directly measured using a high-speed camera. The particle tracing method was systematically applied to investigate the effects of different tracer particles and plasma operating conditions on the axial velocity distribution in the ICP tail flame. Compared with previous work, the present study provides experimental velocity data for the region close to the torch outlet ( $z < 12.5$  mm), filling a gap in the characterization of near-field plasma flow. The results show that a clear velocity plateau along the plasma axis ( $V_z$  distribution) can be observed in both experimental measurements and numerical simulations, consistent with the particle trajectory simulations of Mostaghimi *et al.*, thus validating the reliability of the experimental method and the measured data. These findings offer direct experimental evidence for a deeper understanding of the dynamic flow behavior in analytical ICPs and provide critical reference data for the development and validation of numerical models. The observed velocity distributions and their response to operating parameters are of significant

importance for optimizing ICP analytical conditions and enhancing the predictive capability of numerical simulations for real plasma behavior.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Montaser A, 1998, Inductively Coupled Plasma Mass Spectrometry, John Wiley.
- [2] Boss C, Fredeen K, 1997, Concepts, Instrumentation and Techniques in Inductively Coupled Plasma Optical Emission Spectrometry, Perkin-Elmer, Second.
- [3] Tsivilskiy I, Gilmutdinov A, Nikiforov A, et al., 2020, An Experimentally Verified Three-Dimensional Non-Stationary Fluid Model of Unloaded Atmospheric Pressure Inductively Coupled Plasmas, *J. Phys. D. Appl. Phys.*, 53(2020): 455203.
- [4] Nagulin K, Akhmetshin D, Gilmutdinov A, et al., 2015, Three-Dimensional Modeling and Schlieren Visualization of Pure Ar Plasma Flow in Inductively Coupled Plasma Torches. *J. Anal. At. Spectrom.*, 2015(30): 360–367.
- [5] Nagulin K, Tsivilskiy I, Akhmetshin D, et al., 2017, Transient Three-Dimensional Dynamics of Argon Plasma within the Vacuum Interface of the Inductively Coupled Plasma Mass Spectrometer System. *Spectrochimica Acta Part B: At. Spectrosc.*, 2017(135): 63–72.
- [6] Colombo V, Ghedini E, Mostaghimi J, 2008, Three-Dimensional Modeling of an Inductively Coupled Plasma Torch for Spectroscopic Analysis. *IEEE Trans. Plasma Sci.*, 2008(36): 1040–1041.
- [7] Gamez G, Lehn S, Huang M, et al., 2007, Effect of Mass Spectrometric Sampling Interface on the Fundamental Parameters of an Inductively Coupled Plasma as a Function of its Operating Conditions: Part II. Central-Gas Flow Rate and Sampling Depth. *Spectrochim Acta Part B: At. Spectrosc.*, 2007(62): 357–369.
- [8] Huang M, Hanselman S, Yang P, et al., 1992, Isocontour Maps of Electron Temperature, Electron Number Density and Gas Kinetic Temperature in the Ar Inductively Coupled Plasma Obtained by Laser-Light Thomson and Rayleigh Scattering. *Spectrochim Acta Part B: At. Spectrosc.*, 1992(47): 765–785.
- [9] Ebert C, Saetvit N, Bajic S, et al., 2020, High-Speed Photographic Study of Vaporclouds from Wet Droplets and the Subsequent Solid Particles in an Inductively Coupled Plasma. *J. Anal. At. Spectrom.*, 2020(35): 1956–1958.
- [10] Aeschliman D, Bajic S, Baldwin D, 2003, Spatially-Resolved Analysis of Solids by Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry: Trace Elemental Quantification without Matrix-Matched Solid Standards. *J. Anal. At. Spectrom.*, 2003(18): 1008–1014.
- [11] Guo X, Alavi S, Dalir E, et al., 2019, Time-Resolved Particle Image Velocimetry and 3D Simulations of Single Particles in the New Conical ICP Torch. *J. Anal. At. Spectrom.*, 2019(34): 469–479.
- [12] Han X, Su Y, Li Z, et al., 2025, Experimental Study on the Dynamic Characteristics of an Analytical Inductively Coupled Plasma and its Tail Flame. *Journal of Analytical Atomic Spectrometry*, 2025(40): 1916–1928.

### Publisher's note

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



# Dynamics of Electric Field Perturbation in Gold Nanobipyramids During Dual-Pulse Two-Photon Coherent Excitation

Qiong Li, Yao Li\*

Third Hospital of Shanxi Medical University, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, Taiyuan, 030032, China

\*Corresponding author: Yao Li, [liyao@sx bqeh.com.cn](mailto:liyao@sx bqeh.com.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:** Optical-based microwave electric field detection has emerged as a research hotspot due to its advantages of high spatial resolution and immunity to electromagnetic interference. However, existing techniques are often limited by their sensitivity or reliance on specialized fluorescent materials. Gold nanobipyramids (AuNBPs), serving as nanoprobes with tip-enhancement effects and a well-defined three-level system, exhibit high sensitivity in their two-photon photoluminescence (TPPL) process to phase perturbations and plasmon resonance changes induced by microwave fields. By establishing a quantitative mapping model between microwave intensity and TPPL signal strength, we achieved an absolute measurement of microwave field strength with a spatial resolution that breaks the 100-nanometer barrier. Through comparative analysis of microwave responses under different pulse delays, we reveal that the microwave field primarily modulates TPPL intensity by interfering with the coherent excitation pathway. The most significant response of TPPL intensity to microwave power was observed near the zero-delay point, where the quantum coherence is strongest.

**Keywords:** Two-photon photoluminescence; Dual-pulse coherent excitation; Gold nanobipyramids; Microwave electric field perturbation

**Online publication:** April 22, 2026

## 1. Introduction

Microwave measurement technology, a cornerstone of the modern information society, plays an indispensable role in high-speed communications, precision radar, Earth observation, and remote sensing<sup>[1-3]</sup>. Traditional microwave detection primarily relies on electrical probes (e.g., dipole antennas, semiconductor probes) that measure field strength via electromagnetic induction. However, limited by device size and electromagnetic coupling effects, the spatial resolution of these methods is typically confined to the millimeter or even centimeter scale. This constraint makes precise mapping and visualization of field distributions at micro- and nano-scales extremely challenging,

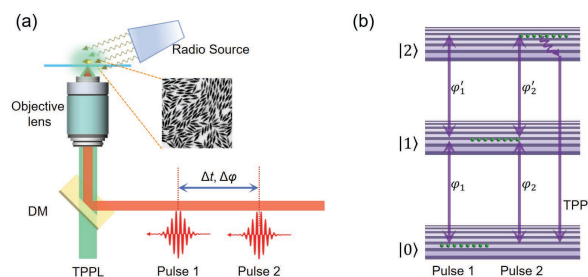
severely hindering their application in frontier scenarios such as integrated circuit near-field analysis and biological electromagnetic effect monitoring<sup>[4]</sup>.

To overcome these limitations, optically read microwave electric field sensing techniques have been developed. These methods leverage the nonlinear interaction between light and matter under a microwave field, converting microwave signals into optical responses for measurement. They offer combined advantages of non-invasiveness, high spatial resolution, and immunity to electromagnetic interference, making them an international research focus<sup>[5,6]</sup>. Existing optical approaches include polarization modulation based on electro-optic crystals, electromagnetically induced transparency based on Rydberg atoms,<sup>[2-4][2-4]</sup> and intensity detection based on fluorescent materials<sup>[7-9]</sup>. However, these methods face significant challenges: electro-optic crystals have limited sensitivity, atomic systems require complex vacuum environments, and fluorescent probes are prone to photobleaching and rely on external labeling, complicating their use in complex *in-situ* systems<sup>[10,11]</sup>. Therefore, there is a pressing need to develop a novel optical mechanism for microwave field detection that combines high sensitivity, nano-scale spatial resolution, and label-free operation.

To overcome these limitations, optically read microwave electric field sensing techniques have been developed. These methods leverage the nonlinear interaction between light and matter under a microwave field, converting microwave signals into optical responses for measurement. They offer combined advantages of non-invasiveness, high spatial resolution, and immunity to electromagnetic interference, making them an international research focus<sup>[5,6]</sup>. Existing optical approaches include polarization modulation based on electro-optic crystals, electromagnetically induced transparency based on Rydberg atoms, and intensity detection based on fluorescent materials<sup>[7-9]</sup>. However, these methods face significant challenges: electro-optic crystals have limited sensitivity, atomic systems require complex vacuum environments, and fluorescent probes are prone to photobleaching and rely on external labeling, complicating their use in complex *in-situ* systems<sup>[10,11]</sup>. Therefore, there is a pressing need to develop a novel optical mechanism for microwave field detection that combines high sensitivity, nano-scale spatial resolution, and label-free operation.

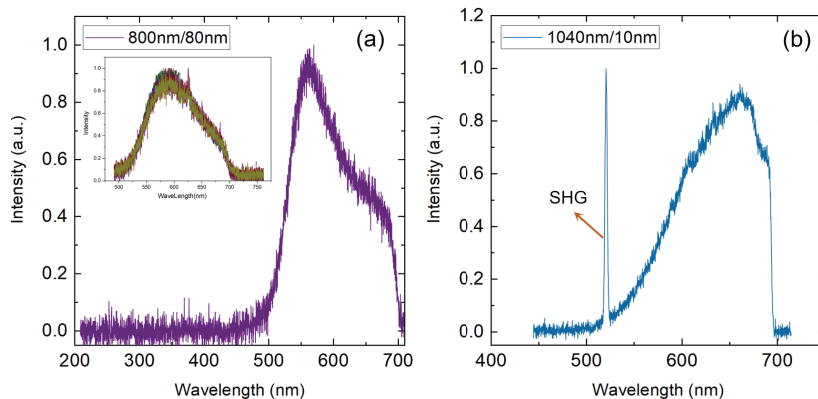
## 2. Results and discussion

The presence of an intermediate state in the TPPL process of AuNBPs grants them excellent phase sensitivity and nonlinear coherent enhancement effects under dual-pulse coherent excitation, as systematically studied in our prior work<sup>[16,17]</sup>. The model is illustrated in **Figure 1b**. Experimentally, as shown in **Figure 1a**, Pulse 1 and Pulse 2, with controlled delay and phase characteristics, excite the AuNBPs. The microwave signal is emitted from an antenna connected to a microwave source, directed at the sample stage, with microwave-absorbing foam placed around to prevent reflections that could alter microwave power.



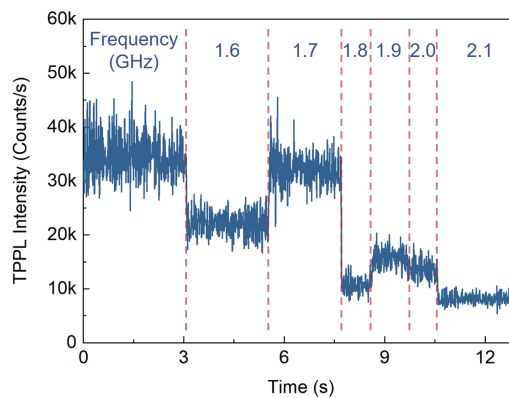
**Figure 1.** Schematic of the experimental setup and the DP-TPPL model for AuNBPs. (a) Experimental setup. Pulses with controlled delay and phase excite AuNBPs; the TPPL signal is collected by an objective lens, filtered, and detected. DM: Dichroic Mirror. (b) Coherent excitation DP-TPPL model for AuNBPs, illustrating the real intermediate state and the coherence-dominated TPPL process. F1 and F2 represent the electric field vectors of the two pulses.

We first verified that the microwave-affected photoluminescence from AuNBPs originates primarily from TPPL, not second-harmonic generation (SHG). As shown in **Figure 2**, we compared the photoluminescence spectra of AuNBPs excited by femtosecond lasers with different bandwidths: 800 nm, 15 fs (80 nm bandwidth) and 1040 nm, 80 fs (10 nm bandwidth). Under 800 nm excitation at a very low pulse energy ( $< 0.625$  pJ), the signal is dominated by TPPL (**Figure 2a**). In contrast, 1040 nm excitation at a much higher pulse energy (100 pJ) produces a prominent SHG signal (**Figure 2b**). For symmetric AuNBPs, the narrow-band 1040 nm laser requires high energy, potentially causing deformation and breaking symmetry to generate SHG. The broadband 800 nm laser provides efficient TPPL for microwave detection at ultralow pulse energies without damaging the particles.



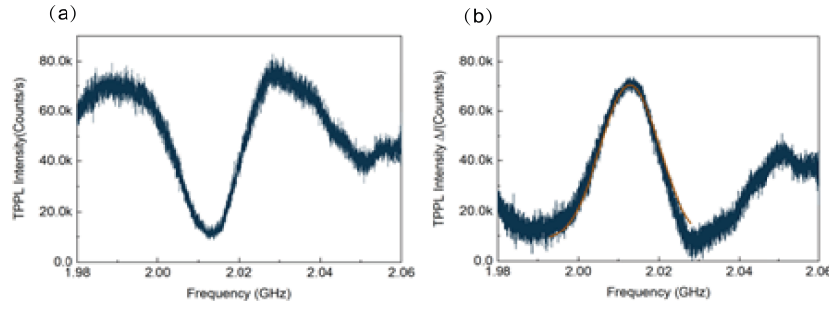
**Figure 2.** TPPL versus SHG in AuNBPs. (a) Spectrum under 800 nm, 15 fs excitation (0.625 pJ). (b) Spectrum under 1040 nm, 80 fs excitation (100 pJ).

Given their coherent response, we tested the microwave response of AuNBPs. **Figure 3** shows that the TPPL signal from AuNBPs (70 nm long axis, 20 nm short axis) responds mainly around 2.0 GHz at a fixed microwave power of 10 dBm.



**Figure 3.** Response of AuNBPs TPPL to microwave radiation at different frequencies (microwave power: 10 dBm).

A continuous frequency scan from 1.98 GHz to 2.06 GHz (**Figure 4a**) shows that the microwave field attenuates the TPPL signal. Analysis of the intensity change  $\Delta I$  (**Figure 4b**) reveals a characteristic peak that can fit with a Gaussian lineshape (orange-red line), suggesting a direct relationship between the microwave response and electron population.



**Figure 4.** Response of AuNBPs TPPL to microwave frequencies in the 1.98-2.06 GHz range. (a) TPPL intensity vs. frequency. (b) TPPL intensity change  $\Delta I$  vs. frequency with Gaussian fit.

## 2.1. Mechanism of microwave response

The 2 GHz microwave photons possess negligible energy and cannot directly induce electronic transitions. However, the oscillating electric field can perturb the relaxing non-equilibrium electron distribution created by the femtosecond laser, altering the electron energy distribution or population, which in turn affects the radiative recombination rate and modulates the luminescence intensity. Theoretically, after femtosecond excitation, the electron system can be described by a Fermi-Dirac distribution with an elevated electron temperature  $T_e$ :

$$f_e(E, t=0^+) \approx [1 + \exp((E - E_F)/(k_B T_e))]^{-1} \quad (1)$$

Where  $T_e$  is on the order of hundreds of Kelvins and cools on a picosecond scale via electron-phonon scattering.

The microwave field  $E_{mw}(t) = E_0 \cos(\omega_{mw} t)$  acts as a driving source, performing work on the electrons and continuously perturbing their distribution. Within the relaxation time approximation, this process is understood via a linearized Boltzmann equation:

$$\partial \delta f / \partial t = -\delta f / \tau - e E_{mw}(t) \cdot v (\partial f_0 / \partial E) \quad (2)$$

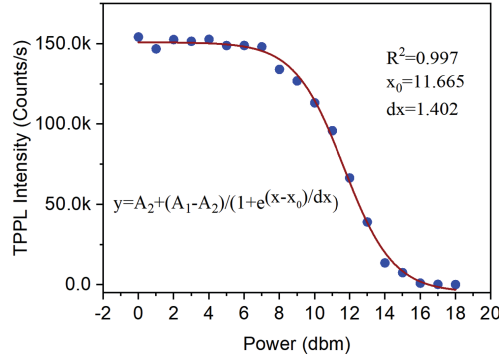
Where  $\tau$  is the electron momentum relaxation time ( $\sim 10$ -30 fs for gold).

This perturbation directly alters the hot electron population. The PL intensity  $I_{PL}(t)$  is proportional to the product of the hot electron population  $n_{hot}(t)$  and the radiative recombination rate  $\Gamma_{rad}$ :

$$I_{PL}(t) \propto \Gamma_{rad} \cdot \int_{E_F + \hbar \omega_{PL}}^{\infty} D(E) \cdot f_e(E, t) dE \quad (3)$$

The microwave-induced perturbation  $\delta f$  to the distribution function thus directly modulates  $I_{PL}(t)$ . Under weak-field, low-frequency ( $\omega_{mw} \tau \ll 1$ ) conditions, the relative modulation depth can be estimated as  $\Delta I_{PL} / I_{PL0} \propto (|e E_0|^2 \tau^2) / (m^* k_B T_e) \cdot \tau_{eff} / [1 + (\omega_{mw} \tau_{eff})^2]$ . This indicates that the modulation depth is proportional to the microwave power ( $\propto |E_0|^2$ ) and is tightly linked to the characteristic temperature  $T_e$  and effective lifetime  $\tau_{eff}$  ( $\sim$ hundreds of fs to ps) of the non-equilibrium electrons, explaining why this modulation is observable only within the short temporal window following femtosecond laser excitation.

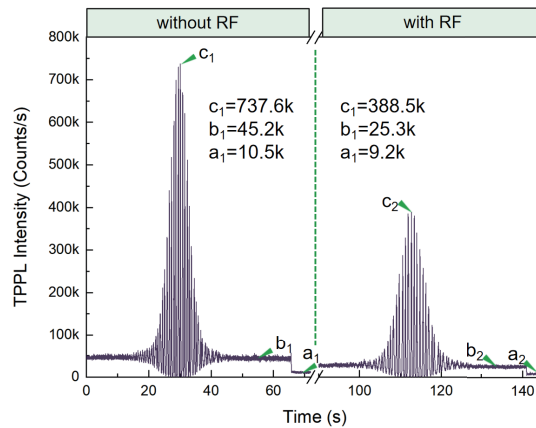
Based on this, we investigated the microwave power-dependent response. As shown in **Figure 5**, the TPPL intensity of AuNBPs decreases gradually with increasing microwave power, following a Boltzmann function fit with  $R^2 > 0.99$ .



**Figure 5.** Power-dependent response of AuNBPs TPPL to 2.0 GHz microwave radiation. Blue dots: experimental data; red line: Boltzmann function fit. Key fitting parameters  $x_0$  and  $dx$  are indicated.

In **Figure 5**,  $A_2$  and  $A_1$  define the response range of the TPPL intensity. The parameters  $x_0$  and  $dx$  are the half-activation power and slope factor, respectively. Here,  $x_0 = 11.665$  dBm, corresponding to an electric field strength of approximately 0.849 V/m, which is lower than typical biological field strengths. A slope factor  $dx > 1$  demonstrates the high sensitivity of AuNBPs for microwave measurement. The successful Boltzmann fit further suggests that the microwave perturbation directly affects the electron population in the three-level system during the TPPL process<sup>[18]</sup>. Our prior work confirmed that the DP-TPPL in AuNBPs, due to the intermediate state, shows a sensitive phase response, allowing TPPL enhancement by two orders of magnitude via phase control<sup>[17]</sup>. Therefore, we studied the effect of microwave fields on TPPL intensity under different dual-pulse delays.

Green arrows in **Figure 6** indicate the TPPL intensity with and without the microwave field at different delays. Comparison shows that while single-pulse excitation involves the intermediate state, the microwave influence time is limited ( $\sim 15$  fs pulse width), resulting in a small ratio  $a_1:a_2 = 1.14$ . At longer delays (b,  $\Delta t \approx 83$  fs), within the intermediate state lifetime, the microwave effect increases, yielding  $b_1:b_2 = 1.79$ . When the pulse interval is further reduced to near zero delay (c), the microwave interaction time relative to case (b) is shorter, yet the largest ratio  $c_1:c_2 = 1.90$  is observed. This is because at zero delay, AuNBPs exhibit the strongest quantum coherence under dual-pulse excitation, and this intermediate-state-dominated coherence is more susceptible to microwave disruption, leading to a greater reduction in DP-TPPL<sup>[19]</sup>.



**Figure 6.** Response of AuNBPs TPPL to 2.0 GHz microwave radiation at different time delays.  $a_1$  ( $a_2$ ): TPPL intensity under single-pulse excitation without (with) microwave field (RF).  $b_1$ ,  $c_1$  ( $b_2$ ,  $c_2$ ): DP-TPPL intensity at delays  $\Delta t \approx 83$  fs and  $\Delta t \approx 0$  without (with) microwave field, respectively.



### 3. Conclusion

This study developed a microwave electric field measurement method based on the dual-pulse two-photon photoluminescence of gold nanobipyramids. Experimental results demonstrate that this method leverages the plasmonic properties and three-level system of AuNBPs, enabling the detection of microwave fields via dual-pulse femtosecond laser excitation. AuNBPs with dimensions of  $70\text{ nm} \times 20\text{ nm}$  showed a selective response to a 2.0 GHz microwave frequency. The microwave electric field alters their plasmon resonance characteristics, thereby affecting TPPL intensity. A quantitative relationship between microwave field strength and TPPL signal was established via Boltzmann function fitting. Effective detection was achieved using 800 nm, 15-fs femtosecond laser pulses at single-pulse energies below 0.625 pJ, with a spatial resolution that reached the sub-100-nanometer level. Compared to conventional methods, this approach is label-free and robust against interference. Experiments with different pulse delays indicate that the microwave field primarily modulates the TPPL signal by affecting intermediate-state population and the coherent excitation process, with the most significant response observed near zero time delay. This work presents a novel detection tool for analyzing electromagnetic fields at micro- and nanoscales.

### Funding

National Natural Science Foundation of China (Project No.: 62205190); China Postdoctoral Science Foundation (Project No.: 2022M722003 and 2024T170536); Shanxi Basic Research Program (Project No.: 202203021212100); Shanxi Bethune Hospital Scientific Research Startup Fund (Project No.: 2021RC032); Central Guiding Local Science and Technology Development Fund Project (Project No.: YDZJSX2025D072).

### Disclosure statement

The authors declare no conflict of interest.

### References

- [1] Zou X, Lu B, Pan W, et al., 2016, Photonics for Microwave Measurements. *Laser & Photonics Reviews*, 10(5): 711.
- [2] Hempel H, Savenjie T, Stolterfoht M, et al., 2022, Predicting Solar Cell Performance from Terahertz and Microwave Spectroscopy. *Advanced Energy Materials*, 12(13): 2102776.
- [3] Réouven A, Rémy D, Théau P, et al., 2023, Quantum Advantage in Microwave Quantum Radar. *Nature Physics*, 19(10): 1418.
- [4] Wright E, 1999, MAP: the Microwave Anisotropy Probe. *New Astronomy Reviews*, 43(2–4) 257.
- [5] Zou X, Lu B, Pan W, et al., 2016, Photonics for Microwave Measurements. *Laser & Photonics Reviews*, 10(5): 700.
- [6] Janusz M, Amjed H, Andrew M, et al., 2024, Ultra-Wideband RF-Photonics Technology for Microwave Spectrometry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024: 1716100.
- [7] Yang C, Yu Y, Li J, et al., 2022, Sequential Generation of Multiphoton Entanglement with a Rydberg Superatom. *Nature Photonics*, 16(9): 658.
- [8] Andrea M, Léa L, Angelo C, et al., 2022, Optical Coherent Manipulation of Alkaline-Earth Circular Rydberg States. *Nature Physics*, 18(5): 502.
- [9] Liu Z, Zhang L, Liu B, et al., 2022, Deep Learning Enhanced Rydberg Multifrequency Microwave Recognition.



Nature Communications, 13(1): 1997.

- [10] Kwon J, Elgawish M, Shim S, 2022, Bleaching-Resistant Super-Resolution Fluorescence Microscopy. *Advanced Science*, 9(9): 2101817.
- [11] Wang H, Han G, Tang H, et al., 2023, Synchronous Photoactivation-Imaging Fluorophores Break Limitations of Photobleaching and Phototoxicity in Live-Cell Microscopy. *Analytical Chemistry*, 95(44): 16243.
- [12] Zhang P, Cai T, Zhou Q, et al., 2022, Ultrahigh Modulation Enhancement in All-Optical Si-based THz Modulators Integrated with Gold Nanobipyramids. *Nano Letters*, 22(4): 1541.
- [13] Huang J, Jiang F, Zhao Z, et al., 2024, Gold Nanobipyramid-based Photothermal Sensors for the Portable Evaluation of Total Antioxidant Capacity. *ACS Applied Nano Materials*, 7(12): 14621.
- [14] Yu Y, Li Y, Qin H, et al., 2020, Microwave Measurement and Imaging for Multiple Corrosion Cracks in Planar Metals. *Materials & Design*, 202: 196109151.
- [15] Tao Y, Yang F, Tao Z, et al., 2022, Fully On-Chip Microwave Photonic Instantaneous Frequency Measurement System. *Laser & Photonics Reviews*, 16(11): 2200158.
- [16] Li Y, Qin C, Song Y, et al., 2021, Great Enhancement on Two-Photon Photoluminescence Imaging Contrast of Au Nanoparticles via Double-Pulse Femtosecond Laser Excitation with Controlled Phase Differences. *Optics Express*, 29(15): 22855.
- [17] Li Y, Yang Y, Qin C, et al., 2021, Coherent Interference Fringes of Two-Photon Photoluminescence in Individual Au Nanoparticles: The Critical Role of the Intermediate State. *Physical Review Letters*, 127(7): 073902.
- [18] van Swieten T, Merlijn S, Auke V, et al., 2022, Extending the Dynamic Temperature Range of Boltzmann Thermometers. *Light: Science & Applications*, 11(1).

**Publisher's note**

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

# Comparison and Application Analysis of Three Wireless Charging Methods

Shuqi Wang

School of Environment, Northeast Normal University, Changchun 130117, 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 rapid development of the electronics industry, increasingly stringent requirements have been placed on battery endurance and power transfer efficiency (PTE). To meet the demands of modern high-technology society, numerous research teams have invested substantial efforts in wireless charging technologies. At present, wireless charging mainly includes electromagnetic induction-based wireless charging, magnetic resonant coupling-based wireless charging, and microwave-based wireless charging. By comparatively analyzing the operating principles of these three approaches, this paper summarizes their respective advantages and disadvantages. Electromagnetic induction-based wireless charging is highly constrained by transmission distance and is therefore suitable only for short-range power transfer. Magnetic resonant coupling-based wireless charging enables relatively longer transmission distances; however, it poses potential safety risks, as resonance may occur between the charging equipment and conductive objects in the surrounding environment under certain conditions. Microwave-based wireless charging is well-suited for radio-frequency wireless power transfer (WPT) in the microwave band. Through frequency-band adjustments, it can be extended to long-distance wireless power transfer across multiple bands. In the future, improvements in coil stability, transmitter frequency tuning, and bandwidth expansion may further enhance the power transfer efficiency and application potential of wireless charging technologies.

**Keywords:** Wireless charging; Electromagnetic induction; Magnetic coupling; Electromagnetic microwave

**Online publication:** April 22, 2026

## 1. Introduction

With the increasing societal dependence on electronic devices and more stringent performance requirements, wireless charging technology, also known as wireless power transfer (WPT), has attracted extensive attention and research in recent years, achieving notable developments and applications across multiple fields. Traditional wired charging methods are gradually being replaced by emerging wireless approaches. Wireless charging eliminates spatial constraints, significantly reduces transmission losses along cables, and better satisfies expectations for high-efficiency modern devices. Currently, wireless charging technologies mainly include electromagnetic induction-based wireless charging, magnetic resonant coupling-based wireless charging, and microwave-based wireless charging. The most fundamental approach is electromagnetic induction-based wireless charging, which

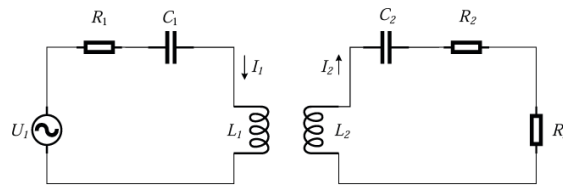
transfers power based on the principle of electromagnetic induction; however, it is strongly limited by transmission distance. Magnetic resonant coupling-based wireless charging also employs electromagnetic induction, but differs in that a resonant compensation circuit is used to adjust the coils to the same resonant frequency <sup>[1]</sup>. In contrast, microwave-based wireless charging utilizes electromagnetic wave radiation and achieves power transfer through mutual conversion between electrical energy and microwave energy, although it is more susceptible to environmental influences. Although both domestic and international research teams have conducted in-depth studies on wireless charging, challenges such as energy losses during power transfer and relatively low power transfer efficiency persists.

This paper comparatively analyzes the technical principles, characteristics, and applications of the three wireless charging methods to identify approaches with superior performance and strategies to overcome inherent limitations. Furthermore, the practical value of these technologies is evaluated to provide clear objectives and directions for future development of wireless power transfer technology.

## 2. Fundamental theories of the technologies

### 2.1. Wireless charging based on electromagnetic induction

When an alternating current is supplied to the transmitting coil, an alternating magnetic field is generated around the coil. If the receiving coil is located within this magnetic field, according to the principle of electromagnetic induction, a current will be induced in the closed circuit whenever the magnetic flux through the loop changes. Consequently, the magnetic field and magnetic flux of the receiving coil also vary, thereby generating an induced electromotive force and induced current. By appropriately regulating the induced electromotive force, power can ultimately be delivered to electronic components <sup>[1]</sup>. **Figure 1** shows the simplified circuit of the electromagnetic induction technique. To reduce magnetic flux loss, this method is significantly constrained by transmission distance and is therefore suitable only for short-range power transfer. Wang *et al.* reported that ion-optimized substitution achieved by co-doping  $\text{Nb}_2\text{O}_5$  and  $\text{Li}_2\text{CO}_3$  into  $\text{NiCuZn}$  ferrite not only promotes uniform densification of the material microstructure but also significantly enhances its electromagnetic properties, including permeability and saturation magnetic induction, while reducing power loss <sup>[2]</sup>. Owing to its excellent electromagnetic performance, this material is suitable as a magnetic isolator for wireless charging transmitters, enabling efficient energy transfer and rapid long-distance charging performance. This also demonstrates that electromagnetic induction wireless charging may suffer from considerable energy losses compared with wired charging and impose stringent requirements on material properties and performance.

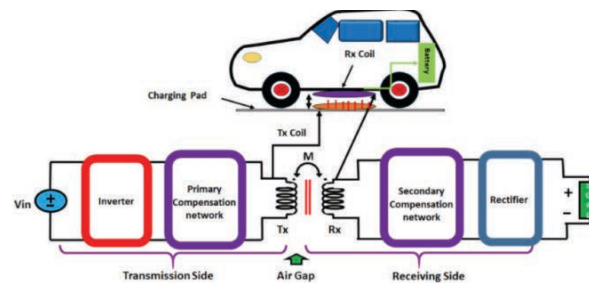


**Figure 1.** Simplified circuit of the electromagnetic induction technique.

### 2.2. Magnetic resonant coupling-based wireless charging

This wireless charging method is based on the phenomenon of magnetic resonant coupling. Its fundamental principle is essentially similar to that of electromagnetic induction-based wireless charging. However, resonant compensation circuits with specific resonant frequencies are additionally introduced to both the transmitting and

receiving coils. When the resonant frequencies of the two coils are matched, magnetic resonance occurs at the receiving side, thereby achieving higher power transfer efficiency (PTE) and enhanced energy delivery capability. **Figure 2** illustrates the simplified schematic circuit of the wireless charging system for electric vehicles. The primary advantage of this method is that the distance requirement between the two coils is less restrictive than that of electromagnetic induction-based wireless charging, enabling relatively longer-distance power transfer. Nevertheless, potential safety risks exist, since the charging equipment may form an unintended resonant system with metallic objects in the surrounding environment under certain conditions <sup>[3]</sup>. To further improve energy harvesting efficiency, researchers have investigated various hybrid structures that enhance performance through combined piezoelectric and electromagnetic transduction. To realize continuous self-powered operation in complex environments, Yu *et al.* proposed a multidirectional piezoelectric-electromagnetic vibration energy harvester (MD-PEVEH), which enables efficient multidirectional energy collection <sup>[4]</sup>. The MD-PEVEH adopts a pendulum-based structure. A magnet located at the bottom of the pendulum interacts with a fixed coil to generate electricity through electromagnetic induction, while an integrated piezoelectric cantilever converts mechanical strain into electrical energy. The multidirectional adaptability of the pendulum significantly enhances energy capture capability in dynamic environments, making the device particularly suitable for applications such as wireless sensor networks, structural health monitoring, and Internet of Things devices. Experimental results demonstrate stable performance under different excitation angles. At an excitation frequency of 8.5 Hz, the rectified harvester achieved a maximum output power of 6.99 mW, and power fluctuations were limited within 5%, confirming its multidirectional coherence and stability. Capacitor charging tests further verified its effective energy storage capability. These results indicate that magnetic resonant coupling-based wireless charging features relatively high efficiency and long-distance capability, although limitations remain in terms of long-term durability.

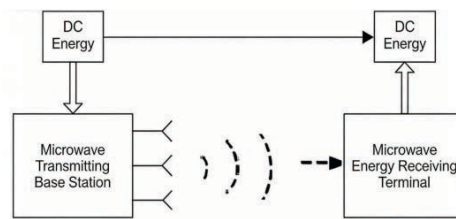


**Figure 2.** Main circuit of the resonant coupling wireless charging system for vehicle charging.

### 2.3. Microwave-based wireless charging

Microwave-based wireless charging is founded on the principle of electromagnetic wave radiation <sup>[3]</sup>. In this method, electrical energy and microwave energy are continuously converted into each other. First, a microwave conversion device converts the input alternating current into microwaves, which are then transmitted through the air over long distances. After being received at the receiving terminal, the microwaves are converted back into electric current by a rectification and conversion circuit. Following appropriate regulation, electrical power is supplied to electronic components. **Figure 3** shows the architecture of the microwave wireless charging system. Although this method enables long-distance power transfer, it is strongly affected by environmental factors as well as the reflection and refraction characteristics of electromagnetic waves, resulting in considerable propagation losses <sup>[5]</sup> at the beginning of each time slot, the SU should make the decision to harvest RF energy or to transmit a packet. The mode decision depends on the channel state and the available energy in the energy

queue. We formulate the mode management problem of the SU as a Markov decision process (MDP). Yan *et al.* proposed the world's first watt-level radio-frequency wireless power transfer (RF-WPT) system with intelligent continuous tracking and obstruction detection [6]. The system operates in the 5.8 GHz band and integrates advanced technologies such as millimeter-precision LiDAR, multi-object image recognition algorithms, and a rectification efficiency of 66.8%. Chen *et al.* introduced a capacitor-free four-layer coil pad using flexible printed circuits and polyimide substrates for medium-power operation at low resonant frequencies [7]. Nanocrystalline ribbon cores were also explored, exhibiting higher magnetic saturation and lower core loss compared with conventional MnZn ferrites. Experimental results demonstrated a power density of up to 29.9 W/cm<sup>3</sup> and an efficiency of 92%. Li proposed a laser-guided strategy to construct hierarchical carbon structures for ultra-broadband microwave absorption [8]. These improvements partially overcome the limitations of microwave-based wireless charging, including relatively low efficiency and narrow operating frequency bands.



**Figure 3.** Architecture of the microwave wireless charging system.

### 3. Comparison of advantages and disadvantages of different technologies

#### 3.1. Electromagnetic induction-based wireless charging

Electromagnetic induction-based wireless charging transfers electrical energy over short distances using a loosely coupled transformer structure with separated primary and secondary coils. This technology is relatively mature and has already been applied in many practical scenarios. For example, wireless charging systems have been deployed for airport shuttle buses at Haneda Airport, as well as non-contact charging pads developed by Splashpower and non-contact charging systems produced by Showa Aircraft Industry. However, this method is strongly limited by transmission distance, and the power transfer efficiency (PTE) decreases significantly as the distance increases. Moreover, the relative positions of the separated primary and secondary coils must be precisely aligned; otherwise, the PTE will deteriorate markedly [9]. Therefore, conventional electromagnetic induction-based wireless charging is subject to considerable constraints, including limited transfer distance, strict alignment requirements, and high material performance demands. Its application range is consequently restricted. Many companies attempt to improve its performance by incorporating additional components or compensation structures based on this technology.

#### 3.2. Magnetic resonant coupling-based wireless charging (MRC-WPT)

To satisfy the requirements of constant-voltage and constant-current charging for electrical devices, various research teams have successively proposed compensation structures and topologies, including basic compensation networks, LCL-S topologies, S/LCL-compensated constant-voltage and constant-current WPT systems, and S-LCC/LCL resonant wireless power transfer systems [10]. These approaches gradually optimize magnetic resonant coupling technology. Under zero-phase-angle (ZPA) conditions, constant-voltage and constant-current power transfer can be achieved. System stability can also be maintained during frequency switching, thereby reducing system cost and the complexity of resonant parameter adjustment [11]. Compared with conventional electromagnetic



induction-based wireless charging, this method not only overcomes the distance limitations of the former but also improves transfer capability. It exhibits lower propagation losses, enables constant-voltage and constant-current output, and provides a more stable WPT system. However, its overall PTE is still lower than that of traditional wired charging methods, and further optimization is required. Consequently, magnetic resonant coupling-based wireless charging demonstrates significant development potential. This approach also has broad market prospects. Owing to its relatively high PTE and stable system performance, it is suitable for short- to medium-range power transfer applications in practical production. It not only reduces material costs but also lowers the maintenance costs associated with traditional wired charging.

### 3.3. Microwave-based wireless charging

Microwave-based wireless charging extends wireless power transfer to the microwave frequency band. As electromagnetic waves at specific microwave frequencies experience relatively low attenuation in air, this method substantially overcomes the distance limitations of both electromagnetic induction-based and magnetic resonant coupling-based wireless charging, making it applicable to a wider range of scenarios. With the advancement of microwave WPT research, various groups have proposed high-performance microwave energy harvesting antennas, high-efficiency broadband rectifier circuits, and hybrid solutions combining resonant coupling technology with microwave techniques <sup>[12]</sup>. Compared with the previous two methods, microwave-based wireless charging exhibits smaller propagation losses and can achieve transmission distances on the order of kilometers. However, broadband rectifier circuits, which constitute one of its core components, remain insufficiently mature. Further research is required in bandwidth expansion and efficiency optimization. Microwave-based wireless charging is therefore suitable for large-scale applications in microwave-band wireless power transfer, particularly for enterprises requiring long-distance power transfer. With appropriate frequency-band adjustments, it can also be extended to remote WPT applications across multiple bands.

### 3.4. Comparative analysis of the three technologies

Based on the principle analysis above, the similarities and differences among electromagnetic induction-based, magnetic resonant coupling-based, and microwave-based wireless charging technologies are summarized in **Table 1**.

**Table 1.** Comparative analysis of the three wireless charging methods

Wireless charging method	Principle	Advantages	Disadvantages
Electromagnetic induction-based	Electromagnetic induction	Reduces losses associated with traditional wired transmission media	Strongly distance-limited; suitable only for short-range power transfer
Magnetic resonant coupling-based	Electromagnetic induction and magnetic coupling	Relatively high PTE and stable WPT system; suitable for short range transfer	Overall efficiency still lower than wired charging
Microwave-based	Electromagnetic wave radiation	Applicable to microwave bands; supports long-distance and multi-band WPT after frequency adjustment	Highly affected by environmental metallic objects and other uncertainties

## 4. Challenges and future development of wireless charging technology

In recent years, wireless charging has achieved encouraging progress. However, several critical bottlenecks



remain, including short charging distance, limited flexibility, and low power transfer efficiency (PTE) <sup>[13]</sup>. Electromagnetic induction-based wireless charging suffers from issues such as large transmission delay, instability, and possible connection interruptions during the communication process. Magnetic resonant coupling-based wireless charging faces challenges related to coil performance, including the quality factor of the coils, the relative position and distance between coupled coils, winding resistance, and load resistance matching. All these factors can significantly affect the power transfer efficiency of the WPT system as well as the durability of the device. Microwave-based wireless charging, on the other hand, must address challenges such as expanding the usable microwave frequency band and mitigating the influence of conductive objects and environmental factors along the propagation path. In future work, substantial research efforts are required to further improve wireless charging technologies to enable their large-scale application in production and daily life. To enhance system stability, electromagnetic induction-based and magnetic resonant coupling-based wireless charging methods can improve coil performance and adopt various compensation structures to achieve constant-voltage and constant-current power transfer, thereby stabilizing the WPT system. In contrast, microwave-based wireless charging is more susceptible to uncontrollable environmental factors; therefore, additional components or shielding structures may be introduced at both the transmitting and receiving sides to stabilize the output voltage and current. In terms of power transfer performance, electromagnetic induction-based and magnetic resonant coupling-based methods can reduce propagation losses by controlling the transmitter operating frequency, thereby improving efficiency. For microwave-based wireless charging, bandwidth expansion, optimization of microwave signal sources and power amplifiers, and integration with conventional communication systems can help reduce losses and further enhance power transfer efficiency.

## 5. Conclusion

Wireless charging is expected to become a core technology for future wireless power transfer (WPT). This paper introduces three wireless charging methods: electromagnetic induction-based, magnetic resonant coupling-based, and microwave-based wireless charging. Through an analysis of their operating principles and a comparison of their respective advantages and disadvantages, it is found that conventional electromagnetic induction-based wireless charging is strongly limited by transfer distance and requires strict alignment of transformer components as well as high material performance. On this basis, magnetic resonant coupling-based wireless charging improves performance through coupling optimization. It provides relatively high power transfer efficiency and a more stable WPT system, making it suitable for short- to medium-range power transfer, although energy losses during propagation still exist. Microwave-based wireless charging is suitable for long-distance power transfer in the microwave frequency band but is more susceptible to environmental influences. The comparative analysis provides guidance for selecting the most appropriate wireless charging method under specific application scenarios. For example, magnetic resonant coupling-based methods are preferable for medium-range applications, whereas microwave-based methods are more suitable for long-distance power transfer. By examining the limitations of each technology, targeted improvements and hybrid integration of multiple approaches can be developed to overcome the performance bottlenecks of individual methods and promote the integrated advancement of wireless charging technologies. At present, various technical routes with different parameters coexist in the market. Summarizing these wireless charging methods helps identify common performance indicators, improve compatibility among different devices, reduce industrial adaptation costs, and support sustainable development. Driven by continuous technological optimization and expanding application scenarios, the wireless power transfer industry is entering

a period of rapid growth. Nevertheless, challenges such as propagation losses, lower efficiency compared with conventional wired methods, cost control, and electromagnetic radiation safety must still be addressed through further research and technological advancements.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Yuan Z, 2020, Transmission Characteristic Analysis of Wireless Charging System Based on Magnetic Resonance, thesis, Hebei University of Technology.
- [2] Wang P, Wang P, Wang J, et al., 2024, Enhancing Electromagnetic Properties of NiCuZn Ferrites through Nb and Li Co-Doping for Wireless Power Transfer. *Ceramics International*, 50(24, Part C): 54966–54975.
- [3] Lin S, Mo X, 2025, Research on Wireless Charging Technology for Power Batteries of New Energy Vehicles on Highways. *Popular Science & Technology*, 27(4): 92–95.
- [4] Yu Y, Zhou J, Tang C, et al., 2025, A Multidirectional Piezoelectric-Electromagnetic Vibration Energy Harvester for Sustainable Power Generation in Dynamic Environments. *ACS Applied Materials & Interfaces*, 17(28): 40539–40545.
- [5] Zhao J, Wei Y, Song M, et al., 2015, Dynamic Mode Management in Cognitive Radio Networks with RF Energy Harvesting, In 11th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2015), 156–158.
- [6] Yan Z, Hu C, Hou B, et al., 2025, A Watt-Level RF Wireless Power Transfer System with Intelligent Auto-Tracking Function. *Electronics*, 14(7): 1259.
- [7] Chen L, Yu B, Fu Y, et al., 2024, Pushing Wireless Charging from Station to Travel, In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, 46–61.
- [8] Li Z, Li N, Song Y, et al., 2025, Laser-Engineered Carbon Hierarchies with Multiscale Attenuation Mechanisms for Ultrabroadband Microwave Absorption. *Chemical Engineering Journal*, 2025(522): 168152.
- [9] Yang Z, 2022, Design and Implementation of an Electromagnetic Induction Wireless Charging Receiver System, thesis, Southeast University.
- [10] Kurs A, Karalis A, Moffatt R, et al., 2007, Wireless Power Transfer via Strongly Coupled Magnetic Resonances. *Science*, 317(5834): 83–86.
- [11] Li G, Wang Z, Wang G, 2026, Design of S-LCC/LCL Resonant Radio Energy Transmission System. *Journal of Tianjin University of Technology*, 2026: 1–8.
- [12] Xu B, 2019, Research on Technology of Microwave Wireless Charging Circuit, thesis, University of Electronic Science and Technology of China.
- [13] Bobba P, Anwar M, Bhupathi H, et al., 2024, Simultaneous Wireless Power and Data Transfer in Different Applications, In E3S Web of Conferences, 01147.

### Publisher's note

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

# Research on Optimization of FPGA Streaming Processing System for High-Bandwidth Radar Echo Data

**Zhonghao Jiang**

Institute of Intelligent Manufacturing, Chongqing Technology and Business Institute, Chongqing 400052, 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:** For the real-time processing scenario of high-bandwidth radar echo data, an optimization scheme for a streaming processing system based on FPGA is proposed. Focusing on the engineering implementation requirements under conditions of high throughput, low latency, and resource constraints, the overall system architecture is designed, and optimizations are carried out in three aspects: pipeline parallel computing, cache organization and memory access scheduling, and timing convergence under resource constraints. The system adopts a modular streaming data path, builds a collaborative mechanism between on-chip cache and computing units, reduces data transfer overhead, and improves the continuity and stability of the processing link. Experimental results show that the optimized system can operate stably under higher input bandwidth conditions, with improved throughput capacity and real-time processing performance, compressed critical path, and resource utilization remaining within a reasonable range. This research can provide a reference for the engineering implementation of high-bandwidth radar signal processing platforms.

**Keywords:** High-bandwidth radar echo; FPGA; Stream processing; System optimization

**Online publication:** April 22, 2026

## 1. Introduction

The improvement of radar system resolution and detection accuracy has led to continuous growth in echo sampling rate, data bit width, and channel scale. This has placed greater demands on the data throughput and real-time constraints of the signal processing platform. The traditional serial processing architecture has struggled to meet the engineering application requirements in terms of bandwidth carrying capacity, memory access efficiency, and power consumption control. FPGA, with its high parallelism, reconfigurability, and low latency characteristics, has significant advantages in the construction of high-speed data paths. Conducting research on the optimization of streaming processing systems for high-bandwidth echo scenarios is of practical significance for enhancing the engineering implementation capabilities of radar signal processing platforms<sup>[1]</sup>.

## 2. FPGA streaming processing system architecture design for high-bandwidth radar echo data

### 2.1. Analysis of processing requirements for high-bandwidth radar echo data

High-bandwidth radar echo data has the characteristics of high sampling rate, continuous data arrival, and concentrated processing pressure. In the engineering implementation, the system not only needs to meet the stable data reception capability, but also needs to take into account the on-chip cache depth, computing throughput capacity, and processing delay constraints. If the data transfer rate is lower than the input rate, it is prone to cause cache accumulation and link congestion; if the parallelism of the computing unit is insufficient, the real-time processing requirements are difficult to meet<sup>[2]</sup>. For this type of application, a demand model needs to be established from three aspects: input data rate, processing delay, and computing load.

$$R_{in} = N_{ch} \cdot f_s \cdot Q \cdot \eta \quad (1)$$

Where  $R_{in}$  represents the data rate of radar echo input, measured in bit/s;  $N_{ch}$  denotes the number of parallel receiving channels;  $f_s$  indicates the sampling rate of a single channel, measured in Hz;  $Q$  represents the quantization bit width of a single sampling point, measured in bits;  $\eta$  is the coefficient of data encapsulation and interface overhead.

$$T_{proc} \leq T_{PRI} - T_{buf} \quad (2)$$

Where  $T_{proc}$  represents the total processing delay for single-pulse echo data, measured in seconds;  $T_{PRI}$  is the pulse repetition interval, measured in seconds;  $T_{buf}$  is the time occupied by buffer scheduling and data transfer, also measured in seconds.

$$C_{tot} = \frac{N_{ch} N_s}{T_{PRI}} (\alpha N_{FFT} \log_2 N_{FFT} + \beta N_{mf} + \gamma N_{acc}) \quad (3)$$

Where  $C_{tot}$  represents the total computing load required per unit time;  $N_s$  is the number of sampling points for a single channel and single pulse;  $N_{FFT}$  is the number of Fourier transform points;  $N_{mf}$  is the scale of matching filter calculation;  $N_{acc}$  is the scale of accumulation operation;  $\alpha$ ,  $\beta$ , and  $\gamma$  are the conversion coefficients for corresponding processing stages.

The above analysis indicates that the high-bandwidth radar echo processing requirements are not a single-bandwidth issue, but rather a systematic constraint resulting from the coupling of data throughput, computing load, and real-time response. The subsequent architecture design should be based on this constraint relationship<sup>[3]</sup>.

### 2.2. Overall architecture design of FPGA streaming processing system

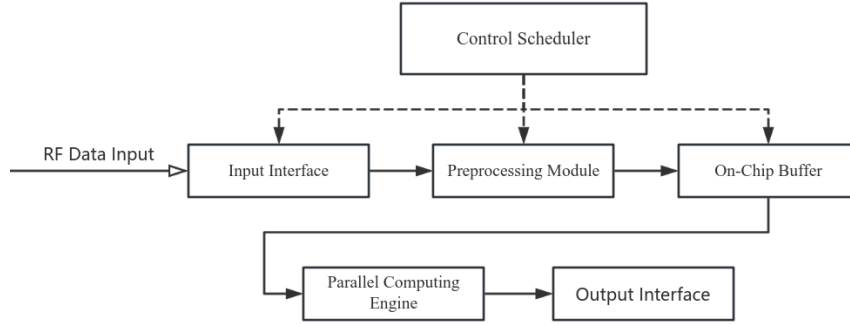
Given the characteristics of continuous arrival of high-bandwidth radar echo data, long processing link, and significant differences in throughput requirements among modules, the system adopts a data-flow-oriented hierarchical architecture. It integrates high-speed input, data caching, parallel computing, and result output into a unified data path. Each processing module is cascaded in a streaming transmission manner, reducing intermediate write-back and repetitive movement, and ensuring the continuity and stability of the processing process. To ensure the overall operational efficiency of the system, the matching relationship between input and processing capabilities needs to be described:

$$R_{sys} = \min(R_{in}, R_{buf}, R_{comp}, R_{out}) \quad (4)$$

Where  $R_{sys}$  represents the effective throughput rate of the system;  $R_{in}$  denotes the throughput rate of the input interface;  $R_{buf}$  indicates the data exchange rate supported by the buffer module;  $R_{comp}$  signifies the processing rate

of the computing module;  $R_{out}$  represents the transmission rate of the output link.

Based on the above relationships, the overall system architecture is shown in **Figure 1**.



**Figure 1.** Overall architecture of the FPGA streaming processing system.

### 3. High-bandwidth radar echo data FPGA streaming processing optimization method

#### 3.1. Parallel computing optimization based on pipeline structure

The processing link of high-bandwidth radar echo data includes multiple computing steps such as data rearrangement, frequency domain transformation, filtering, and accumulation. A single-level serial structure is difficult to meet the real-time processing requirements under continuous input conditions<sup>[4]</sup>. To address this issue, the system divides the core computing process into several functionally independent sub-levels, advancing them in sequence driven by the clock, enabling different stages to process different batches of data in parallel. This approach can shorten the combinational logic length within a single cycle, increase the hardware operating frequency, and is also beneficial for stabilizing the output rate. To characterize the throughput capability of the pipeline parallel structure, the system processing rate can be expressed as:

$$P_{pipe} = f_{clk} \cdot N_{par} \quad (5)$$

Where  $P_{pipe}$  represents the theoretical processing rate of the pipeline computing unit, measured in samples/s;  $f_{clk}$  represents the system operating clock frequency, measured in Hz;  $N_{par}$  represents the amount of data processed in parallel within a single clock cycle.

After the parallelism is improved, the overall acceleration effect of the system is also influenced by the number of pipeline stages, the degree of inter-stage balancing, and the control overhead. To describe the changes in computing performance before and after optimization, an acceleration ratio model can be used:

$$S = \frac{T_{serial}}{T_{pipe}} = \frac{N \cdot t_{op}}{(k+N-1)t_{clk} + T_{ctrl}} \quad (6)$$

Where  $S$  represents the parallel acceleration ratio of the pipeline;  $T_{serial}$  is the time required for the serial structure to complete all calculations;  $T_{pipe}$  is the processing time of the pipeline structure;  $N$  is the number of batches of data to be processed;  $t_{op}$  is the average processing time per batch in the serial mode;  $k$  is the number of pipeline levels;  $t_{clk}$  is the processing time of a single level pipeline within one clock cycle;  $T_{ctrl}$  is the additional overhead caused by control scheduling and inter-level synchronization.

From **Equations (5) and (6)**, it can be seen that reasonably increasing the parallelism and maintaining load balance at all levels is the key to improving the efficiency of stream computing.



### 3.2. Cache and memory access optimization for high-bandwidth data streams

For high-bandwidth radar echo data under continuous input conditions, there are higher requirements for cache organization and memory access efficiency. If the on-chip cache depth is insufficient, data bursts arriving simultaneously are prone to accumulate; if there are waiting or conflicts in external storage access, computing units will become idle, and the system throughput will decrease. To address this issue, the design adopts a hierarchical cache and dual-buffer coordination mechanism, decoupling the data reception, relocation, and computing processes, so that when one set of data enters the computing stage, another set of data can simultaneously complete writing or reading, reducing the impact of memory access blocking on the processing link. To describe the cache capacity requirements, the minimum cache depth can be expressed as:

$$D_{\text{buf}} \geq R_{\text{in}} \cdot T_{\text{lat}} \quad (7)$$

Where  $D_{\text{buf}}$  represents the minimum required capacity of the cache module, measured in bits;  $R_{\text{in}}$  indicates the input data rate, measured in bit/s;  $T_{\text{lat}}$  denotes the maximum tolerable delay for data waiting to be processed or moved, measured in seconds.

At the level of memory access scheduling, the system enhances the utilization of effective bandwidth through burst transmission, contiguous address mapping, and Bank partitioning methods. After considering memory access conflicts and control overheads, the effective bandwidth of the storage system can be expressed as:

$$B_{\text{eff}} = B_{\text{peak}} \cdot \eta_{\text{burst}} \cdot \eta_{\text{bank}} \cdot \eta_{\text{ctrl}} \quad (8)$$

Where  $B_{\text{eff}}$  represents the effective memory bandwidth, measured in bit/s;  $B_{\text{peak}}$  represents the theoretical peak bandwidth of the memory;  $\eta_{\text{burst}}$  is the burst transmission efficiency coefficient;  $\eta_{\text{bank}}$  is the efficiency coefficient for multi-Bank parallel access;  $\eta_{\text{ctrl}}$  is the control scheduling efficiency coefficient.

The cache depth and effective bandwidth jointly determine the continuous supply capacity of data flow, and are also an important condition for ensuring the stable operation of the stream processing link.

### 3.3. FPGA resource and timing optimization design

During the optimization process of the high-bandwidth radar echo streaming processing system, not only the improvement of throughput rate needs to be considered, but also the occupation of logic resources, on-chip memory allocation, and the difficulty of timing convergence should be taken into account. Increasing the parallelism can enhance the processing capability, but it will also lead to an increase in the consumption of LUTs, registers, DSPs and BRAMs. The length of the critical path expands with the increase in module complexity, and the maximum operating frequency of the system is limited. In the design, methods such as reusing arithmetic units, inserting critical path registers, localizing control logic and constraining data bit width are adopted to jointly optimize the resource consumption and timing performance. To describe the relationship of resource consumption under parallel configuration, the overall resource requirement can be expressed as:

$$U_{\text{res}} = N_{\text{par}} (\lambda_1 U_{\text{LUT}} + \lambda_2 U_{\text{FF}} + \lambda_3 U_{\text{DSP}} + \lambda_4 U_{\text{BRAM}}) + U_{\text{ctrl}} \quad (9)$$

Where  $U_{\text{res}}$  represents the overall resource consumption of the system;  $N_{\text{par}}$  indicates the number of parallel processing units;  $U_{\text{LUT}}$ ,  $U_{\text{FF}}$ ,  $U_{\text{DSP}}$  and  $U_{\text{BRAM}}$  correspond to the lookup table, register, DSP and on-chip storage resource consumption of a single parallel unit respectively;  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  and  $\lambda_4$  are the different resource conversion weights;  $U_{\text{ctrl}}$  represents the additional resource consumption brought by the control and interface logic.

The core of timing optimization lies in reducing the delay of the critical path, so that the system can meet the



target clock constraints. After pipelining the registers and splitting the logic, the clock cycle should satisfy:

$$T_{clk} \geq T_{comb} + T_{reg} + T_{skew} + T_{margin} \quad (10)$$

Where  $T_{clk}$  represents the system clock cycle;  $T_{comb}$  denotes the combinational logic delay on the critical path;  $T_{reg}$  indicates the delay related to register setup and hold;  $T_{skew}$  represents the delay introduced by clock skew;  $T_{margin}$  is the timing design margin.

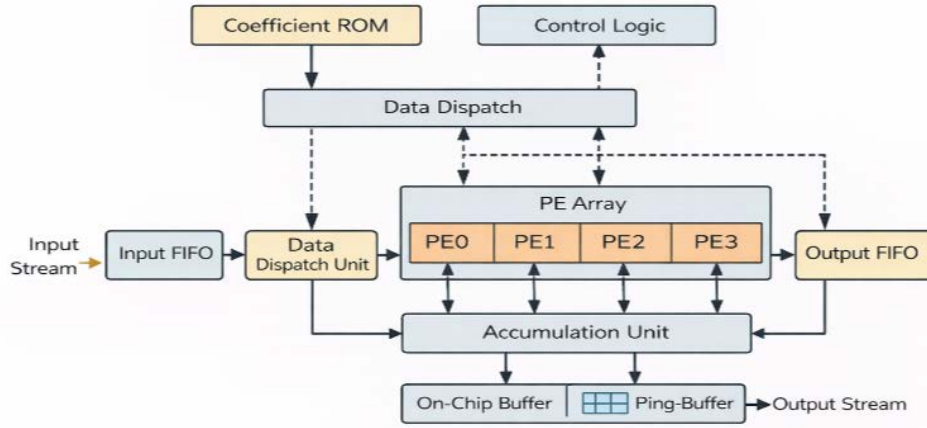
Resource constraints and clock constraints are coupled with each other. Only when a balance is achieved among the parallel scale, module division, and timing convergence, can the system maintain stable operation under high bandwidth conditions.

## 4. System implementation and performance verification

### 4.1. FPGA system implementation and module design

The system implementation aims at the continuous reception and real-time processing of high-bandwidth radar echo data. The hardware platform uses FPGA as the core processing component. A complete processing link is constructed around the input interface, data cache, computing unit, control scheduling, and result output. The input end completes the parallel reception and byte alignment of high-speed echo sampling data. The pre-processing module is responsible for data format conversion, effective data extraction, and timing arrangement, providing stable input for subsequent calculations. The on-chip cache module adopts a partitioned management method, isolating data writing, reading, and computing processes to reduce the impact of sudden traffic on the processing link. The core computing part is divided into several independent functional units according to the echo processing flow. Each unit operates in a unified clock domain and is connected according to the stream interface, with data advancing along a fixed path after entering, avoiding frequent backwriting of intermediate results to external storage. The control scheduling module is responsible for state switching, cache address management, and module handshake control, ensuring stable coordination of each processing stage under load changes<sup>[5]</sup>.

Module design emphasizes clear structure and engineering feasibility. The internal computing unit adopts cascaded data paths, mapping multiplication, addition, transformation, and accumulation operations to processing structures suitable for hardware expansion, facilitating subsequent parallel expansion and timing optimization<sup>[6]</sup>. For common data blocking problems in high-bandwidth scenarios, the system sets a flow control mechanism between the cache and the computing unit. It dynamically adjusts the reading and writing rhythm based on the buffer status to maintain a matching relationship between the input rate and the computing rate. The interface design adopts a standardized packaging method to reduce the coupling between modules, facilitating subsequent independent debugging and function reuse. During the implementation process, local buffering processing was carried out on key control paths and high-fanout signals to shorten the combinational logic depth and improve overall timing stability. After modular division, the system has good scalability. When the input bandwidth or processing scale increases, the structure can be expanded by increasing the number of parallel units or adjusting the cache configuration. The system implementation structure is shown in **Figure 2**.



**Figure 2.** FPGA system implementation and module design.

## 4.2. System resource utilization and timing performance analysis

After the system completed the synthesis, layout and timing analysis, the usage of major hardware resources and key timing indicators were statistically analyzed. The results showed that the resource allocation of each functional module was relatively balanced. The computing units and on-chip cache accounted for the majority of the system's hardware costs, which was consistent with the parallel computing requirements and data buffering requirements during the high-bandwidth echo data processing process. The proportion of control scheduling and interface logic was relatively low, and it did not cause significant pressure on the overall resource distribution. The overall implementation results showed that the utilization rates of LUTs, registers, BRAMs and DSPs remained within the acceptable range of the device, and a certain expansion space was reserved to support subsequent adjustments for higher input bandwidth or larger processing scales <sup>[7]</sup>.

From the perspective of timing performance, after key path optimization and register insertion processing, the system could meet the target clock constraints, and no significant negative slack issues were found in the main data paths. The key paths were concentrated between the parallel computing module and the cache control module, indicating that the coordination of computing access and memory access under high-speed data flow remains an important factor affecting timing convergence <sup>[8]</sup>. The analysis of system resource utilization and clock performance revealed that the current implementation structure achieved a reasonable balance among throughput capacity, resource consumption and timing stability. The specific resource utilization and timing results are shown in **Table 1**.

**Table 1.** System resource utilization and timing analysis results

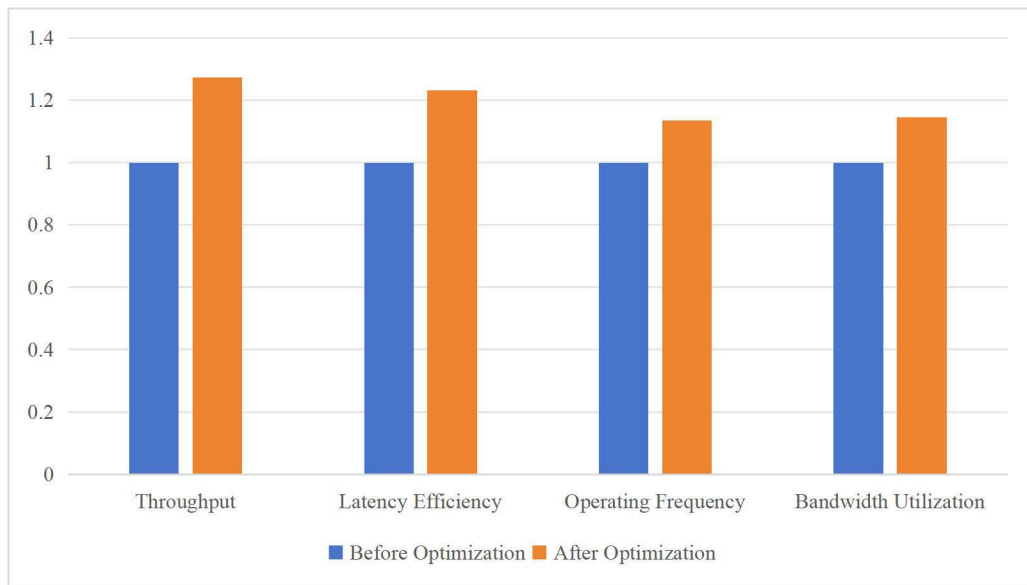
Module	LUT	FF	BRAM	DSP	Maximum operating frequency / MHz	Timing slack / ns
Input and preprocessing module	4210	3896	12	8	268	1.84
On-chip buffer module	3568	3014	28	0	251	1.27
Parallel computing module	9824	8742	16	64	243	0.96
Control scheduling module	2146	2378	4	0	286	2.15
Output interface module	1875	1692	6	2	272	1.73
System total	21623	19722	66	74	243	0.96

### 4.3. Experimental results and system performance evaluation

After the system was fully implemented, functional verification and performance testing were conducted for the scenario of high-bandwidth continuous echo input, focusing on throughput rate, processing delay, bandwidth utilization, and operational stability. The test results indicated that the optimized streaming processing system could maintain stable operation under continuous data input conditions, without any obvious data accumulation or link congestion. Compared with the unoptimized implementation, the system performed better in terms of throughput capacity and delay control, demonstrating the effectiveness of the pipeline parallel structure, cache scheduling method, and resource timing optimization. The improvement in bandwidth utilization indicated an enhanced matching between data transfer and computation processes, and the overall real-time processing capability of the system met the design expectations. The specific experimental results are shown in **Table 2**, and the performance change trends before and after optimization are depicted in **Figure 3**.

**Table 2.** System performance evaluation results

Metric	Before optimization	After optimization	Improvement
Throughput / Gbps	9.86	12.56	27.4%
End-to-end latency / $\mu$ s	34.4	27.9	18.9%
Maximum operating frequency / MHz	214	243	13.6%
Bandwidth utilization / %	78.2	89.5	11.3%
Continuous operation stability time / h	12	12	-



**Figure 3.** Normalized comparison of system performance before and after optimization.

As shown in **Figure 3**, the optimized system demonstrates significant improvements in terms of throughput capacity, delay control, and bandwidth utilization. This indicates that the proposed streaming processing optimization method can effectively enhance the overall performance of the high-bandwidth radar echo data processing system.

## 5. Conclusion

This paper focuses on the real-time processing requirements of high-bandwidth radar echo data and conducts research on the design and optimization of a streaming processing system based on FPGA. By addressing key issues such as system throughput, cache access, parallel computing, and timing constraints, the overall architecture design, optimization method analysis, and engineering implementation have been completed. The design concept proposed in this paper can well adapt to the scenario of continuous high-bandwidth data input, achieving a balance between processing efficiency, structural stability, and resource consumption. The related research has certain reference value for the engineering design of high-performance radar signal processing platforms and also provides a foundation for subsequent system expansion targeting larger data scales.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Altalqi F, Fennane S, Kacimi H, et al., 2024, Development of a Multi-Band High Bandwidth Circular Microstrip Patch Antenna for Radar Implementations, 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM), 1–5.
- [2] Altalqi F, Fennane S, Mabchour H, et al., 2024, Design Monopole Antenna of Ultra-Wideband High Bandwidth and High Efficiency for Ground Penetrating Radar Application. *Telkomnika*, 22(4): 154–163.
- [3] Hong A, Su L, Wang Y, et al., 2025, A Terahertz Dual-Band Transmitter in 40 nm CMOS for a Wideband Sparse Synthetic Bandwidth Radar. *Electronics* (2079-9292), 14(22): 24–30.
- [4] Liu M, Xie X, Deng Y, et al., 2023, Efficient Implementation of BP Imaging Algorithm on FPGA. *IET Conference Proceedings*, 2023(47): 1092–1097.
- [5] Jonsson R, Ankel M, Tholen M, et al., 2023, Experimental Analysis of a Clutter Suppression Algorithm for High Time-Bandwidth Noise Radar, 2023 IEEE International Radar Conference (RADAR), 1–6.
- [6] Xie Y, Zhong Z, Li B, et al., 2024, An ARM-FPGA Hybrid Acceleration and Fault Tolerant Technique for Phase Factor Calculation in Spaceborne Synthetic Aperture Radar Imaging. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024(17): 5059–5072.
- [7] Ankel M, Tholen M, Bryllert T, et al., 2024, Implementation of a Coherent Real-Time Noise Radar System. *IET Radar, Sonar & Navigation*, 18(7): 1002–1013.
- [8] Zhang Z, Xu Y, Li Y, et al., 2023, Spaceborne SAR Imaging System with High-Performance Polynomial Engines. *IET Conference Proceedings*, 2023(47): 2435–2442.

### Publisher's note

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

# Practices and Insights of Scientific Data Security Grading Management Based on the Entire Life Cycle

Yu Zhai, Yong Song

Xinjiang Academy of Science and Technology for Development, Urumqi 830011, 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 era of big data, scientific data has become a strategic resource for national scientific and technological innovation and economic and social development, and the importance of its security management has become increasingly prominent. Based on the theory of the entire life cycle management of scientific data, this paper deeply discusses the core connotation of data security grading management, and systematically analyzes the prominent problems existing in the current scientific data security management in terms of system connection, process coverage, technology adaptation, and rights protection. On this basis, the paper constructs a practical path of scientific data security grading management covering six stages: data planning, collection, storage, use, sharing, and destruction, and puts forward targeted implementation strategies. Research shows that scientific data security grading management based on the entire life cycle is not only a technical issue but also a systematic project involving system design, organizational collaboration, and cultural cultivation. It has important theoretical value and practical enlightenment for improving data governance capabilities and promoting the orderly opening and sharing of scientific data.

**Keywords:** Scientific data; Entire life cycle; Data grading; Security management; Data governance

**Online publication:** April 22, 2026

## 1. Introduction

As scientific research enters the era of the fourth paradigm, scientific data has shown an explosive growth trend, with an unprecedented scale, variety, and flow speed. Scientific data is not only the basic output of scientific research activities but also the core element supporting major scientific discoveries and driving economic and social innovation<sup>[1]</sup>. Scientific data security grading management based on the entire life cycle emphasizes embedding security controls into the complete process of data from generation to destruction, and adopting differentiated protection measures according to the differences in data sensitivity and importance. This concept provides a new idea for solving the dilemma of “difficulty in sharing and protecting” scientific data<sup>[2]</sup>. This paper aims to sort out the core connotation of scientific data security grading management based on the entire life cycle, analyze the problems existing in current practices, and explore a systematic practical path, so as to provide useful reference for improving China’s scientific data security management system.



## **2. Core connotation of scientific data security grading management based on the entire life cycle**

Scientific data security grading management based on the entire life cycle is essentially a data governance model that combines process control theory and risk management methods. Its core connotation can be understood from three dimensions.

### **2.1. Entire life cycle**

From the perspective of the “entire life cycle”, scientific data is not a static entity but undergoes a dynamic process of generation, processing, analysis, storage, sharing, reuse, and ultimately destruction. Each stage faces different security threats and corresponds to different management objectives and responsible subjects <sup>[3]</sup>. For example, in the data collection stage, the main risk points lie in the credibility of collection equipment and the integrity of the collection process; in the data sharing stage, it is necessary to focus on the rigor of access control and the effectiveness of data desensitization. Therefore, security management cannot only focus on a single link but should build a closed-loop control system covering the entire life cycle of data.

### **2.2. Grading management**

“Grading management” is the essence of this model. There are significant differences in the value density and security sensitivity of scientific data. Basic and public welfare observation data and scientific research data involving national secrets, trade secrets, or personal privacy cannot be managed with the same standards. Grading management requires the establishment of scientific and reasonable grading standards. According to factors such as data content attributes, source channels, aggregation effects, and open risks, data is divided into different security levels, and corresponding technical protection measures and management processes are configured accordingly.

### **2.3. Dynamic adjustment and collaborative linkage**

This model emphasizes “dynamic adjustment” and “collaborative linkage”. The life cycle of data does not advance linearly but may cycle between different stages; at the same time, the security level of data may change with the evolution of application scenarios, the degree of data aggregation, and the external environment. Therefore, grading management must be dynamic and iterable, and require multi-party collaboration among data managers, data producers, data users, and even regulatory authorities to form a joint force for security governance. In a word, the core of this model is to embed the concept of security into every link of data flow, and achieve a dynamic balance between security and sharing through refined grading strategies <sup>[4-6]</sup>.

## **3. Problems existing in scientific data security management**

### **3.1. Ambiguous grading standards and poor system connection**

China’s existing data security management mainly faces the problems of ambiguous classification and fragmented management systems. The Data Security Law and the Measures for the Management of Scientific Data only make principled provisions on data classification, but lack detailed implementation rules for data management targeting specific types such as scientific data. There is a lack of clear, specific, and operable grading standard systems. Data characteristics vary greatly in different fields. For example, experimental data in the field of high-energy physics and genetic data in the field of biomedicine have completely different requirements in terms of sensitivity and security. Existing norms often cannot meet the requirements of various disciplines, bringing troubles to data



managers. Secondly, the management of scientific data lacks good connection with the protection of national secrets, trade secrets, and personal privacy in other laws and regulations, and there are institutional blind spots or conflicts to a certain extent, resulting in high costs for carrying out data security protection and applications<sup>[7]</sup>.

### **3.2. Incomplete coverage of the life cycle and shortcomings in key links**

In practice, most data security subjects focus on the storage link, only paying attention to the physical and network boundary protection of data centers, while ignoring potential risks in other links of the entire data life cycle, especially the original data generated by collection links such as field stations and experimental instruments. The security of communication links and terminal authentication are often ignored. In the process of data utilization and sharing, there is a lack of fine-grained authorization mechanisms and dynamic behavior auditing, resulting in risks such as unauthorized access, illegal copying, and even secondary transmission by employees; there is generally a management blind spot in the data destruction link. Even if there are explicit regulations from hard disk destruction to the deletion of unused files, the actual execution process cannot be effectively monitored, which may lead to the leakage of sensitive information and bring security risks.

### **3.3. Disconnection between technology and business and insufficient grading protection capabilities**

Although technology is an important means for grading control, it also faces the dilemma of “two skins” between technology and business in practice: on the one hand, most existing security technology facilities adopt a general design model and cannot accurately match the network traffic characteristics and user behavior of specific scientific research units; for example, excessive information confidentiality will lead to inefficient processing of a large amount of scientific research information, which may reduce the efficiency of scientific and technological research and development<sup>[8]</sup>. In addition, many institutions have not well integrated the level system management in the process of using security technology. For example, after data is classified, there is still no good method and technical support to automatically trigger relevant encryption, desensitization, and access control strategies. The intelligent and automated label recognition technology and strategy implementation capabilities are still lacking, so it is impossible to achieve refined grading control, which is often manifested as extensive “one-size-fits-all” management.

### **3.4. Complex interest relationships and difficulty in balancing security and sharing**

Scientific information often involves many stakeholders: data producers, data processors, fund providers, data owners, and potential re-users, which makes graded and classified governance face complex challenges. On the one hand, out of consideration for protecting their own research results and intellectual property rights, individual scientists or research teams will adopt overly conservative data control strategies. For instance, basic data that should be open is deliberately closed, affecting communication and authentication among the scientific community. Moreover, there is a lack of specific right definition and benefit-sharing system design. Management institutions tend to adopt the strictest security protection measures to avoid liability risks, resulting in a large amount of valuable scientific data being idle and failing to play its due value function. Therefore, how to break these barriers and achieve legal and compliant scientific data sharing on the premise of ensuring the security of important data is the primary problem to be solved in graded and classified management<sup>[9]</sup>.

## **4. Practical path of scientific data security grading management based on the entire life cycle**

### **4.1. Planning stage: Establish grading principles and institutional systems**

Before carrying out the security level control of scientific data, top-level design and planning should be done well, that is, a joint working group composed of data managers, researchers, information security departments, and legal departments should jointly formulate management measures for the security guarantee level of scientific data. This measure mainly clarifies the classification standards and scope. In practice, classification should be based on data secrecy level, completeness, and purpose. On this basis, a data classification directory within the unit or field should be formulated, and clear provisions should be made on the data meaning, identification method, protection criteria, and applicable conditions represented by each category<sup>[10]</sup>. In addition, the post responsibilities in each link should be clarified, the responsibility of classified management should be specifically assigned to the corresponding departments and individuals, and a dynamic communication, coordination, and supervision mechanism should be established to ensure that the classified management work has rules to follow and evidence to rely on.

### **4.2. Collection and processing stage: Implement source grading and label embedding**

The first stage is the collection process and data analysis process. In this process, the first step of classified management is carried out, that is, establishing a dynamically generated classified data model. During the research project approval and data collection, the project leader should predict the data to be produced according to the existing classification system. For data containing important sensitive information, such as human genome data and the operation status of important facilities, a special security assessment should be conducted and a collection plan formulated before collection; during the data collection process, it is necessary to ensure the security of the collection terminal and adopt initial encryption means. A restricted environment should also be established in the process of data cleaning, labeling, re-identification, fusion, etc. In this link, data security labels should be taken as an important part of metadata to ensure that the data security level exists throughout the data circulation, laying a foundation for the automated and refined management of subsequent links. When de-identified information can reduce the sensitivity level, a strict review mechanism should be established to conduct a secondary verification of its security<sup>[11–13]</sup>.

### **4.3. Storage and use stage: Implement differentiated protection and dynamic access control**

In the entire life cycle of data, the storage link and use link are the two most important links with the most interactions, so the security management requirements of these two links are also the highest. In terms of the storage link, different levels of physical security management and logical security protection measures should be implemented according to the degree of data importance; for important data, encrypted storage, off-site multiple backups, and special person custody should be adopted; for critical data, access domain control should be carried out and regular backups should be made<sup>[14]</sup>. For general data, only basic security requirements need to be met. In terms of data application, an attribute-based dynamic access control mechanism should be established, that is, not only based on static identity permissions but also real-time judgment on the legality of access requests according to factors such as user identity, data level, application scenario, and application purpose. For example, for high-level data, computing services should be provided in a secure sandbox to “make data visible but unavailable”. At the same time, strengthen the control of the entire life cycle of data use, especially the real-time monitoring and

abnormal behavior alarm of high-risk operations (search, download, copy) involving high-level data, thus forming effective influence and traceability.

#### **4.4. Sharing and destruction stage: Standardize circulation procedures and closed-loop termination management**

The sharing and destruction links are the most dangerous links in the entire data life cycle. During this period, the “minimum necessary principle” should be followed to determine the specific content of data sharing and the security level of sharing objects. Low-security-level data can be released on public databases; medium-security-level data can be shared by signing agreements, clarifying the scope of use, objects of use, duration of use, and violation penalties; sensitive data is usually only allowed to be accessed by certain personnel in a restricted manner in an internal environment, and the leakage of original data is strictly prohibited. During the sharing process, a complete traceable approval flow and flow record should be available. Data destruction is the last link in the entire life cycle and is often ignored<sup>[15]</sup>. When data exceeds the validity period, completes the established task, or in accordance with legal and regulatory requirements, the data should be destroyed in accordance with procedures. Destruction methods include logical deletion, clearing hard disk data, destroying storage media, etc. According to the data level, it should be carried out in accordance with the two-person system, full-process monitoring, and final confirmation. For external storage media or discarded media, a confidentiality agreement must be signed before supervising the destruction process to avoid information leakage risks caused by improper disposal of discarded media. Now, scientific data is in a complete chain from birth to death, with hierarchical management throughout.

### **5. Conclusion**

Scientific data security grading management based on the entire life cycle is a profound reform of the traditional static and local security thinking. It requires us to view scientific data from a developmental and interconnected perspective and deeply embed security capabilities into the blood of data flow. The scientific data security grading management model based on the entire life cycle is not only a pile of technical tools but also a comprehensive governance framework covering system design, process optimization, organizational collaboration, and cultural cultivation. Its successful practice relies on clear grading standards, full-chain control measures, high integration of technology and business, and prudent balance of multiple interests. At present, China is thoroughly implementing the innovation-driven development strategy and the big data strategy, and the status of scientific data as a national basic strategic resource has become increasingly prominent. Building a scientific, efficient, and entire life cycle security grading management system that meets the requirements of the data element era is of far-reaching significance for safeguarding national security, stimulating innovation vitality, and improving governance efficiency.

### **Disclosure statement**

The authors declare no conflict of interest.

### **References**

- [1] Ren J, Xie Y, Wang T, et al., 2025, Construction of Source Governance and Security Protection System for

Government Personal Information from the Perspective of Data Entire Life Cycle. *Computer Knowledge and Technology*, 21(36): 68–70.

- [2] Xiao J, 2025, Research on the Integrated Management of Electronic Documents Throughout the Life Cycle Under the Background of Government Big Data. *Shanxi Archives*, 2025(12): 155–157.
- [3] Li X, Li K, 2026, Specific Scenarios, Key Risks and System Construction of AI Data Cross-Border Security Supervision from the Perspective of the Entire Life Cycle. *Hebei Law Science*, 44(2): 101–122.
- [4] Xuan W, 2025, “Roadmap” for the Open Sharing of Scientific Data Under the Data Intelligence Paradigm: Recommendation of Open Sharing of Scientific Data: From Ownership Definition to Management System Construction. *Information Studies: Theory & Application*, 48(11): 211.
- [5] Fan B, Duan J, Zhang Y, et al., 2025, Research on Data Security Classification and Grading Management Based on Artificial Intelligence. *Network Security & Informatization*, 2025(11): 20–21.
- [6] Huang S, Chen G, Jin M, et al., 2025, Classification and Grading Management and Security Protection of Data Assets in the Era of Big Data. *Digital Communication World*, 2025(9): 142–144.
- [7] Hao Y, Li D, Han L, et al., 2024, Research and Preliminary Practice of Full-Cycle Data Security Management for Data Middle Platform: A Case Study of National Natural Science Foundation Data Management. *Science Foundation in China*, 38(4): 696–702.
- [8] Gao W, Xu B, Li Z, et al., 2024, Application of Visual Malware Analysis Technology in Data Science in the Digital Archive Security Management System. *Network Security and Data Governance*, 43(5): 18–26.
- [9] Zhang G, Wang J, Pan Y, et al., 2024, Research on Security Management Strategies and Typical Practices of Scientific Data Sharing Platforms at Home and Abroad. *Forum on Science and Technology in China*, 2024(4): 179–188.
- [10] Huang Y, 2023, Research Report on the Data Application Security Mechanism of Customs Based on Grading Management. Dalian Customs, Liaoning Province, November 22, 2023.
- [11] Chang Z, Ye X, Liu W, et al., 2023, Construction of Process-Oriented Management and Risk Control System Model for Scientific Data Security Platform. *Information Research*, 2023(11): 66–73.
- [12] Wu D, Zhao X, Ma D, et al., 2023, Data Security Compliance Management Solution for the National Statistical System, Proceedings of the 2023 Cybersecurity Excellent Innovation Achievements Competition, 4.
- [13] Zou C, Ma H, Wang J, 2023, Performance Analysis Framework and Configuration Analysis of Public Data Security Management. *Library and Information Service*, 67(13): 70–77.
- [14] Liu X, Sun M, 2023, Practices and Insights of Scientific Data Security Grading Management from the Perspective of the Life Cycle. *Information Studies: Theory & Application*, 46(3): 68–74.
- [15] Huang Y, 2022, Analysis of Scientific Data Security Management Strategies in American Libraries in the Digital Economy Era. *Library and Information Guide*, 7(9): 15–19.

**Publisher’s note**

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

# Shadow Thermodynamics of an AdS Black Hole in Non-Commutative Geometry

Ying Zhu\*, Qing-Quan Jiang

School of Physics and Astronomy, China West Normal University, Nanchong 637002, 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:** In this paper, we innovatively adopt the shadow radius to investigate the thermodynamics of an AdS black hole with non-commutative geometry terms. First, via geodesic analysis, we establish a quantitative relationship between the shadow radius and the event horizon radius, and derive the shadow radius of the black hole as a function of the event horizon radius, which exhibits a positive correlation between the two quantities. Furthermore, within the shadow framework, we find that the stability and heat capacity of the black hole can be effectively represented through the shadow radius. Further analysis reveals that the results obtained using the shadow radius in revealing the black hole phase transition process are essentially consistent with those obtained using the event horizon. Based on this, we constructed the thermal profile for an AdS black hole incorporating non-commutative parameters. Within the framework of non-commutative geometry, for  $P < P_c$ , the temperature derived from the shadow radius exhibits a distinct N-shaped trend, which is in perfect agreement with that obtained from the event horizon radius. This result reveals that even in non-commutative spacetime, the phase transition process of AdS black holes can be effectively and intuitively characterized by the thermal profiles of their shadows.

**Keywords:** Black hole; Shadow thermodynamics; Non-commutative geometry; Quantum gravity; Critical phenomenon

**Online publication:** April 24, 2026

## 1. Introduction

In Einstein's theory of general relativity, black holes are predicted as a type of celestial body in the universe<sup>[1]</sup>. Currently, astronomers and physicists have devoted extensive attention and conducted in-depth investigations into the theoretical and experimental studies of black hole physics. As a thermodynamic system, black holes possess properties such as temperature, entropy, and heat capacity. The study of black hole thermodynamics has significantly advanced the development of black hole physics. In recent decades, physicists have gathered extensive astronomical evidence regarding black holes, with the observation of gravitational waves pioneering new approaches to studying their properties. The detection of gravitational wave signals from the merger of two



black holes by the Laser Interferometer Gravitational-Wave Observatory (LIGO) provides compelling evidence for the existence of black holes<sup>[2]</sup>. In 2019, the Event Horizon Telescope (EHT) collaboration reported the ultra-high-resolution image of the supermassive black hole M87\*, providing direct evidence that black holes truly exist in the universe and opening new frontiers for research in black hole observation<sup>[3–9]</sup>. This image reveals a dark central region surrounded by a bright ring, where the dark region is defined as the black hole shadow and the bright ring as the photon ring. Due to the intense gravitational field near a black hole, light undergoes deflection<sup>[10]</sup>. For a static observer at infinity, photons with an orbital radius smaller than the critical bound photon orbit radius fall into the black hole, thus forming the black hole shadow<sup>[11]</sup>. In contrast, other photons escape to infinity. This critical bound orbit is the black hole's photon ring. The size and shape of the black hole shadow are determined by this photon ring<sup>[12–17]</sup>.

Studies of black hole shadows can yield more valuable geometric information about black holes, and observations of these shadows enable a deeper understanding of their intrinsic properties. For a Schwarzschild black hole, the photon ring is located at  $r = 3M$ <sup>[18,19]</sup>. The study of photon trajectories in Schwarzschild black holes was pioneered by Synge and Luminet, and Bardeen subsequently investigated the shadow of the rotating Kerr black hole<sup>[10,20]</sup>. As more physicists conduct research on black hole shadows, it has been discovered that these shadows can be used to test Lorentz symmetry<sup>[21,22]</sup>. Furthermore, by considering the accretion around black holes, the study of shadows for different types of black holes has yielded significant results<sup>[12,14,23–38]</sup>.

Based on this, investigating the unique thermodynamic properties of black holes as thermodynamic systems has become a key objective for many physicists. However, a deeper understanding of black hole thermodynamics began with the establishment of the four laws of black hole thermodynamics and was significantly deepened and substantiated by the discovery of Hawking radiation. Stephen Hawking, by applying quantum field theory, demonstrated that due to vacuum quantum fluctuations, particles can be emitted from the vicinity of a black hole's event horizon (known as Hawking radiation)<sup>[39]</sup>. This discovery laid a solid foundation for the theory of black hole thermodynamics. Based on the foundational assumptions of the “cosmic censorship hypothesis” and the validity of the “strong energy condition”, Hawking proved the proposition known as “Hawking's area theorem” which states that the total area of a black hole's event horizon never decreases within the framework of classical physics. Secondly, the entropy and temperature of a black hole are expressed through the area of its event horizon and its surface gravity<sup>[40]</sup>. Based on this research, the fundamental framework of the four laws of black hole thermodynamics was established. Within Anti-de Sitter (AdS) spacetime, it has been discovered that the thermodynamic system of a charged AdS black hole is essentially equivalent to a van der Waals fluid-gas system<sup>[41–44]</sup>. Pioneering work by Wei *et al.* first explored the dynamic phase behavior of charged AdS black holes within the extended phase space framework and established the fundamental connection between black hole thermodynamics and the shadow radius<sup>[45]</sup>. Subsequently, Zhang *et al.* provided initial evidence that the shadow radius can indeed reflect the thermodynamic phase structure of black holes<sup>[46,47]</sup>.

The black hole shadow serves as a bridge reflecting thermodynamic information of black holes, and combining the two for research holds significant importance. In summary, studies on black hole thermodynamics and their shadows remain very extensive in recent years. This paper primarily investigates the relationship between the black hole shadow radius and thermodynamic phase transitions in the AdS spacetime. Research demonstrates that the black hole shadow can indeed reveal the process of thermodynamic phase transitions. Furthermore, in the context of regular spacetime, it is found that the dependency relationship between the black hole shadow and thermodynamics is likely to be structurally established<sup>[29,48]</sup>.



Furthermore, non-commutative spacetime in gravitational theory is a prominent research topic, as it is regarded as a potential candidate for quantum gravity <sup>[49,50]</sup>. Studying the influence of non-commutativity on black holes constitutes a highly significant research topic, and several methods can be employed to realize non-commutative spacetime within gravitational theory <sup>[51–55]</sup>. Secondly, it is demonstrated that within the framework of general relativity, non-commutativity can be implemented by modifying the source of matter, whereby a Gaussian distribution or a Lorentzian distribution can replace the Dirac delta function <sup>[56, 57]</sup>. Taking the AdS black hole with a non-commutative parameter as an example, this paper investigates the thermodynamic phase transitions under the shadow radius.

## 2. AdS black hole in non-commutative geometry

Non-commutative geometry is a theory of spacetime quantization, in which the commutation relations of spacetime coordinate operators can be expressed as <sup>[45]</sup>:

$$[x^\mu, x^\nu] = i\theta^{\mu\nu} \quad (1)$$

In non-commutative geometry, the Lorentz distribution  $\rho$  of the mass density for spherically symmetric stars is given by <sup>[57]</sup>:

$$\rho = \frac{M\sqrt{\theta}}{\pi^{3/2}(r^2 + \pi\theta)^2} \quad (2)$$

Here,  $M$  denotes the black hole mass and  $\theta$  the non-commutative parameter, which characterizes the minimal spacetime scale.

The Einstein equation is:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + \Lambda g_{\mu\nu} = 8\pi T_{\mu\nu} \quad (3)$$

$R_{\mu\nu}$  is the Ricci tensor, which describes the curvature of spacetime;  $g_{\mu\nu}$  is the metric tensor, which simplifies the expressions of the Einstein field equations and ensures their invariance under different coordinate systems;  $R$  is the Ricci scalar, which represents a global measure of spacetime curvature;  $T_{\mu\nu}$  is the energy-momentum tensor, which describes the distribution of matter and energy in spacetime.

In non-commutative scenarios, a Schwarzschild black hole is:

$$ds^2 = -f(r)dt^2 + f^{-1}(r)dr^2 + r^2d\Omega^2 \quad (4)$$

Substituting **Equation (4)** into **Equation (3)** yields:

$$f(r) = 1 + \frac{1}{r} \int_0^r (8\pi r'^2 T_0^0 - r'^2 \Lambda) dr' \quad (5)$$

$T_0^0$  represents the energy density.

According to the Lorentz distribution formula in **Equation (2)**, we have:

$$T_0^0 = -\rho = -\frac{M\sqrt{\theta}}{\pi^{3/2}(r^2 + \pi\theta)^2} \quad (6)$$

By substituting the above equation into the metric, we can obtain:

$$f(r) = 1 + \frac{1}{r} \int_0^r \left( -8\pi r'^2 \frac{M\sqrt{\theta}}{\pi^{3/2}(r'^2 + \pi\theta)^2} - r'^2 \Lambda \right) dr' \quad (7)$$

in which <sup>[58]</sup>:

$$f(r) = 1 - \frac{2M}{r} + \frac{8M\sqrt{\theta}}{\sqrt{\pi}r^2} - \frac{\Lambda}{3}r^2 + \mathcal{O}(\theta^{3/2}) \quad (8)$$

In the limit  $\theta \rightarrow 0$ , the spacetime geometry reduces to the standard Schwarzschild geometry. For the sake of simplicity, this article introduces a new parameter  $\alpha$ .

$$\alpha = \frac{8\sqrt{\theta}}{\sqrt{\pi}} \quad (9)$$

The metric of Schwarzschild-AdS BH in non-commutative geometry:

$$f(r) = 1 - \frac{2M}{r} + \frac{\alpha M}{r^2} - \frac{\Lambda}{3}r^2 \quad (10)$$

The cosmological constant,  $\Lambda$ , is defined as  $\Lambda = -8P\pi$ . For a black hole to exist, the following conditions must be satisfied:  $f(r_H) = 0$ ,  $r_H > 0$ ,  $M > 0$ ,  $\Lambda < 0$ , so there are the following inequalities that depend on  $\alpha$ .

$$-2r_H^3 + \alpha r_H^2 < 0 \quad (11)$$

$$r_H^4 - 2Mr_H^2 - \alpha Mr_H^2 < 0 \quad (12)$$

Based on **Equation (4)**, in the context of static spacetime, The Lagrangian  $L$  can be expressed as:

$$\mathcal{L} = \frac{1}{2}g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = \frac{1}{2}(-f(r)\dot{t}^2 + f(r)^{-1}\dot{r}^2 + r^2(d\vartheta^2 + \sin^2\vartheta d\varphi^2)) \quad (13)$$

In a spherically symmetric spacetime, we study photons moving on the equatorial plane with  $\theta = \pi/2$ . Furthermore, the metric functions do not depend on the time  $t$  and the azimuthal angle  $\varphi$ . Consequently, the equations associated with the constants of motion can be derived. These constants are:

$$E = -\frac{\partial \mathcal{L}}{\partial \dot{t}} = f(r)\dot{t}, L = \frac{\partial \mathcal{L}}{\partial \dot{\varphi}} = r^2\dot{\varphi} \quad (14)$$

The photon's energy and angular momentum are denoted by  $E$  and  $L$ , respectively. Consequently, for the given line element, the null geodesic condition  $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = 0$  can be expressed as:

$$-f(r)\dot{t}^2 + \frac{\dot{r}^2}{f(r)} + r^2\dot{\varphi}^2 = 0 \quad (15)$$

Then, according to **Equation (14)** and **Equation (15)**, we can obtain:

$$\dot{r}^2 = E^2 - \frac{L^2}{r^2}f(r) \quad (16)$$

The radial geodesic motion can be described by the following equation for the effective potential  $V_{eff}$ .

$$r^2 + V_{eff} = 0 \quad (17)$$

The quantity  $b$  denotes the impact parameter, which is defined as the ratio  $L/E$  and geometrically corresponds to the perpendicular distance from the geodesic to a radial line through the origin. Finally, we can obtain the effective potential  $V_{eff}$  as:

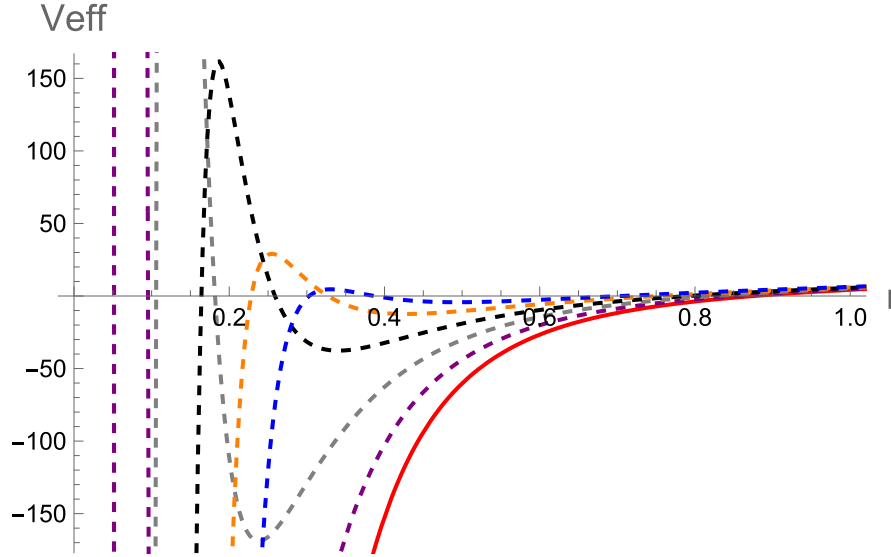
$$V_{eff} = \frac{f(r)}{r^2}L^2 - E^2 \quad (18)$$

Where <sup>[59]</sup>:

$$E^2 = -\frac{f(r)}{r}\frac{df(r)}{dr}, L^2 = l(l+1) \quad (19)$$

In **Figure 1**, we analyze the non-commutative Lorentz effective potential, and investigate its functional dependence on  $r_H$  for different values of  $\alpha$ . **Figure 1** shows the effect of the non-commutative parameter  $\alpha$  on the effective potential for a massive test particle. When  $\alpha$  is non-zero, all extrema of the non-commutative effective potential lie outside the non-commutative event horizon. Increasing  $\alpha$  lowers the maximum peak of the effective

potential and shifts it further from the horizon. The divergence of the effective potential near the event horizon arises from non-commutative geometry, which acts as a potential barrier to prevent high-energy particles from falling into the black hole.



**Figure 1.** The effective potential  $\alpha = 0$ , (Red solid line),  $\alpha = 0.1$ , (Purple dashed line),  $\alpha = 0.2$ , (Gray dashed line),  $\alpha = 0.3$  (Black Dashed line),  $\alpha = 0.4$  (Orange dashed line),  $\alpha = 0.5$  (Blue dashed line).  $P = 0.2$ ,  $M = 1$ ,  $L = 1$ .

### 3. Shadow radius of Schwarzschild-AdS black hole in non-commutative geometry

The following conditions must be satisfied for the effective potential to exist

$$V_{\text{eff}}(r) = 0, V'_{\text{eff}}(r) = 0, V''_{\text{eff}}(r) > 0 \quad (20)$$

Here, we denote the radius of the photon sphere as  $r_{\text{pH}}$ . For a given metric function, we can determine  $r_{\text{pH}}$  by solving the equation.

$$f(r_{\text{pH}}) - \frac{1}{2}r_{\text{pH}}f'(r_{\text{pH}}) = 0 \quad (21)$$

$$r_{\text{pH}} = \frac{1}{2}(3M + \sqrt{M}\sqrt{9M - 8\alpha}) \quad (22)$$

The black hole mass, which is also identified as enthalpy in the extended phase space, is given by the condition  $f(r_{\text{H}}) = 0$ .

$$M = \frac{r_{\text{H}}^2(3 + 8P\pi r_{\text{H}}^2)}{3(2r_{\text{H}} - \alpha)} \quad (23)$$

The shadow radius of a black hole is a crucial observable that provides insights into its fundamental properties. By analyzing the shadow, we can infer the black hole's mass, spin, and nature of spacetime surrounding it. Observing a black hole's shadow allows us to delve deeper into its physical characteristics and the role gravity plays in the universe. It represents the radius of the circular region projected by the black hole's event horizon as viewed by distant observers. The shadow radius, denoted by  $r_0$ , observed from a position at  $r_0$  is given by the following equation<sup>[60]</sup>.

$$r_s = r_{\text{pH}} \sqrt{\frac{f(r_0)}{f(r_{\text{pH}})}} \quad (24)$$

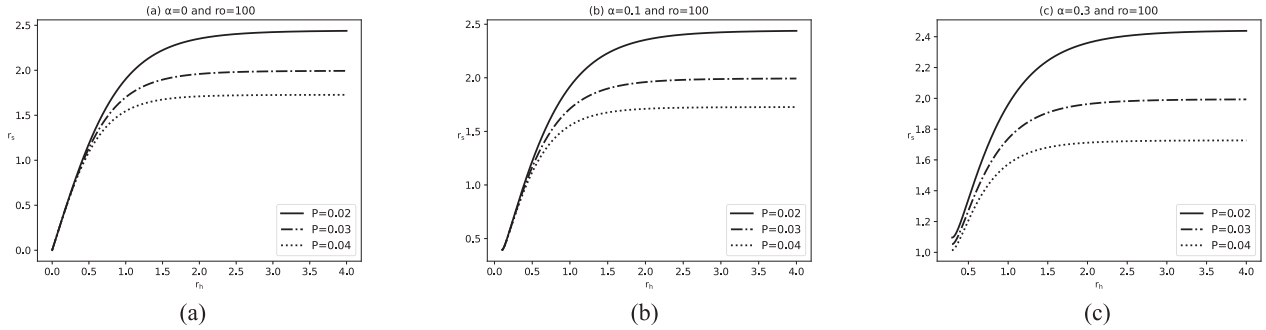
With the static observer at spatial infinity, we set the observer's radial coordinate to  $r_0 = 100$ , satisfying the condition  $f(r_0) = 1$ . Therefore, the shadow radius  $r_s$  is:

$$r_s = r_{pH} \sqrt{\frac{1}{1 + \frac{8P\pi r_{pH}^2}{r_{pH}^2} + \frac{M(-2r_{pH} + \alpha)}{r_{pH}^2}}} \quad (25)$$

$$r_{pH} = \frac{1}{2} \left( \frac{r_H^2(3+8P\pi r_H^2)}{2r_H - \alpha} + \sqrt{\frac{r_H^2(3+8P\pi r_H^2)}{6r_H - 3\alpha}} \sqrt{-8\alpha \frac{3r_H^2(3+8P\pi r_H^2)}{-2r_H + \alpha}} \right) \quad (26)$$

**Figure 2** depicts the relationship between the black hole shadow radius  $r_s$  and the event horizon radius  $r_H$  for different values of the non-commutative parameter  $\alpha$ . The results show that the shadow radius  $r_s$  increases with the event horizon radius  $r_H$ , while the growth rate gradually decreases. This demonstrates a positive correlation between  $r_H$  and  $r_s$ , suggesting that the shadow radius can serve as an indicator of the black hole's Hawking temperature. Furthermore, the shadow radius  $r_s$  decreases with increasing pressure  $P$ . For a fixed horizon radius, increasing the non-commutative parameter  $\alpha$  also leads to a decrease in  $r_s$ .

When the non-commutative parameter  $\alpha$  approaches zero, the black hole model reduces to a Schwarzschild-AdS black hole. As shown in **Figure 2(a)**, the Schwarzschild-AdS black hole also exhibits a positive correlation  $\alpha$  between  $r_H$  and  $r_s$ . Therefore, we conclude that the shadow radius can effectively describe the black hole temperature, similarly to the event horizon radius.



**Figure 2.** The relationship between shadow radius and the horizon radius. (a) Non-commutative parameter  $\alpha = 0$ ,  $r_0 = 100$  (Schwarzschild-AdS BH), (b) Non-commutative parameter  $\alpha = 0.1$ ,  $r_0 = 100$ , (c) Non-commutative parameter  $\alpha = 0.3$ ,  $r_0 = 100$ . The solid lines, segment point lines and dotted lines correspond to  $P = 0.02$ ,  $P = 0.03$  and  $P = 0.04$ , respectively.

Next, let's calculate some fundamental thermodynamic quantities. The black hole's mass is

$$M = \frac{r_H^2(3+8P\pi r_H^2)}{3(2r_H - \alpha)} \quad (27)$$

A black hole's temperature is its Hawking temperature, determined by its surface gravity,

$$T = \frac{f'(r_H)}{4\pi} = \frac{3r_H - 3\alpha - 3r_H^3\Lambda + 2r_H^2\alpha\Lambda}{12\pi r_H^2 - 6\pi r_H\alpha} \quad (28)$$

To study criticality, a specific volume,  $v = 2r_H$ , was introduced. Using this relationship, the equation of state becomes:

$$P = - \frac{3v - 6\pi T v^2 - 6\alpha + 6\pi T v\alpha}{6\pi v^3 - 8\pi v^2\alpha} \quad (29)$$

The black hole's entropy satisfies the Bekenstein-Hawking relation.

$$S = \frac{A}{4} = \pi r_H^2 \quad (30)$$

where  $A = 4\pi r_H^2$  is the surface area of the black hole's event horizon.

However, when the energy-momentum tensor  $T_0^0$  includes the black hole's mass, the traditional first law of thermodynamics,  $dM = TdS + VdP$  is violated, as shown by the following relations.

$$T \neq \left(\frac{\partial M}{\partial S}\right)_P, V \neq \left(\frac{\partial M}{\partial P}\right)_S \quad (31)$$

The modified first law of thermodynamics can be expressed as:

$$dM = WdM = TdS + VdP \quad (32)$$

Here,  $W$  represents the correction function.

$$W = 1 + \int_{r_H}^{+\infty} 4\pi r^2 \frac{\partial T_0^0}{\partial M} dr = 1 - \frac{\alpha}{2r_h} + O(\alpha^3) \quad (33)$$

Using the modified first law, we can derive the Hawking temperature.

$$T = W \left(\frac{\partial M}{\partial S}\right)_P = \frac{f'(r_H)}{4\pi} = \frac{3r_H - 3\alpha - 3r_H^3\Lambda + 2r_H^2\alpha\Lambda}{12\pi r_H^2 - 6\pi r_H\alpha} \quad (34)$$

The black hole's thermodynamic volume also can be derived.

$$V = W \left(\frac{\partial M}{\partial P}\right)_S = \frac{4\pi r_H^3}{3} \quad (35)$$

This result is consistent with those derived from the general thermodynamics of black holes. The critical point must meet:

$$\frac{\partial^2 P}{\partial (r_H)^2} = \frac{\partial P}{\partial (r_H)} \quad (36)$$

The critical condition is:

$$P_c = \frac{0.0027338}{\alpha^2}, T_c = \frac{0.036403}{\alpha}, r_c = 2.44155\alpha \quad (37)$$

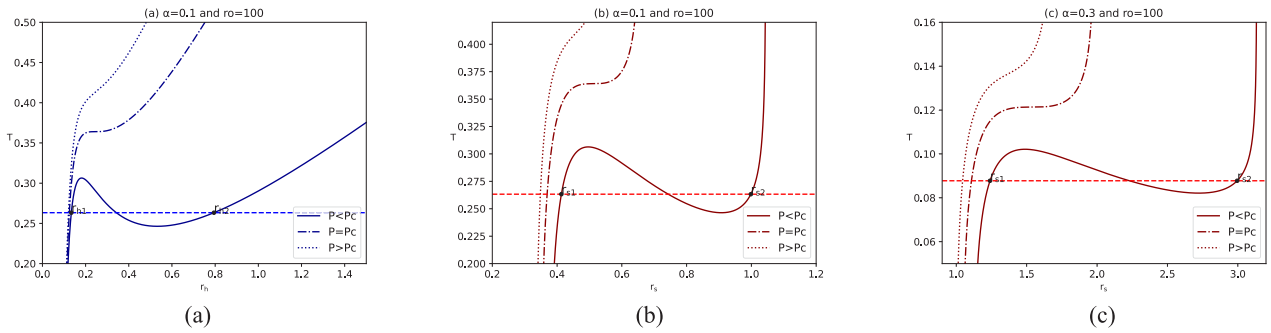
## 4. Shadow analysis of non-commutative phase transitions in Schwarzschild-AdS black holes

In this section, we first investigate the thermodynamic phase transitions of non-commutative Schwarzschild-AdS black holes with a Lorentzian potential. Based on our previously derived Hawking temperature formula, the black hole temperature varies continuously with the event horizon radius  $r_H$ . As shown in **Figure 3(a)**, the black hole temperature exhibits distinct behaviors under different pressure conditions:

- (1) When ( $P = 1.4P_c$ ), the temperature curve is a monotonically increasing smooth curve without an inflection point. The slope initially decreases and then increases, but remains always positive, indicating a supercritical phase;
- (2) When ( $P = P_c$ ), the temperature curve remains monotonically increasing, but the slope decreases to zero before increasing again, resulting in an inflection point. This curve represents the critical isobar, and the black hole is thermodynamically unstable at this point;
- (3) When the pressure ( $P = 0.4P_c$ ), the horizon radii are  $r_h^1 = 0.1338$  and  $r_h^2 = 0.7944$ , respectively. This implies that when ( $P < P_c$ ), within the range  $r_h^1 < r_H < r_h^2$ , the temperature curve exhibits non-monotonic behavior: initially increasing, then decreasing, and finally increasing again. When the pressure is below

the critical pressure, the black hole exhibits two-phase coexistence. This means that the black hole can exist in two distinct stable states: a smaller “liquid-like” state and a larger “gas-like” state, with an unstable intermediate state between them. This phase transition is analogous to the liquid-gas phase transition in van der Waals fluids.

**Figure 3(b)** and **3(c)** show the relationship between black hole temperature and shadow radius  $r_s$  for non-commutative parameters  $\alpha = 0.1$  and  $\alpha = 0.3$ , respectively. These representative values demonstrate that the shadow radius  $r_s$  can replace the event horizon radius  $r_H$  to study the phase transition process of the non-commutative Schwarzschild-AdS black hole. When the pressure  $P$  is less than the critical pressure  $P_c$ :  $r_s < r_s^1$  corresponds to a smaller stable black hole;  $r_s^1 < r_s < r_s^2$  corresponds to a thermodynamically unstable black hole; and  $r_s > r_s^2$  corresponds to a larger stable black hole. Furthermore, as the non-commutative parameter  $\alpha$  increases, the region of thermodynamic instability expands, and both the shadow radius  $r_s$  and the temperature increase.



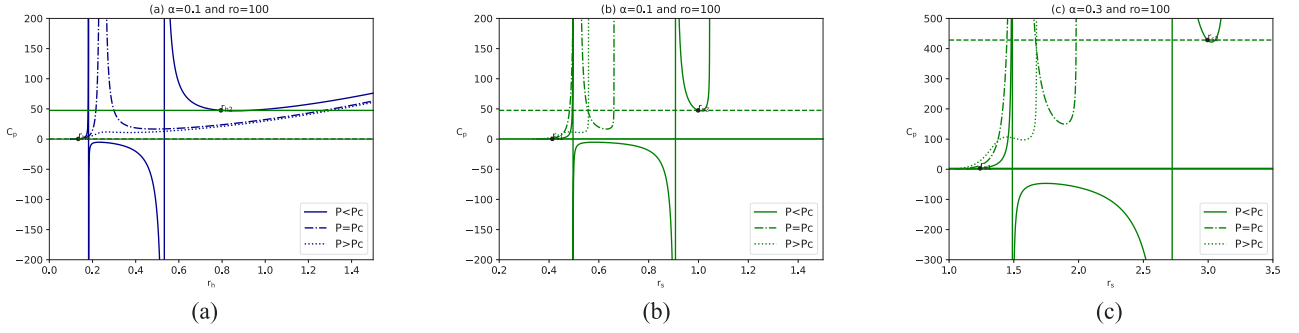
**Figure 3.** (a) Temperature as a function of  $r_H$  with  $\alpha = 0.1$ , (b) temperature as a function of  $r_s$  with  $\alpha = 0.1$ , (c) temperature as a function of  $r_s$  with  $\alpha = 0.3$ . A static observer is located at  $r_0 = 100$ .

The order of a black hole’s phase transition is closely associated with its heat capacity. In particular, a second-order phase transition at the thermodynamic critical point of a black hole can be characterized by a jump in heat capacity and a divergence in specific heat. Taking into account the relationship between entropy and area ( $S = \frac{A}{4} = \pi r_H^2$ ) and the temperature from **Equation 28**, we can derive the heat capacity of a black hole under constant pressure.

$$C_P = T \left( \frac{\partial S}{\partial T} \right)_P = \frac{8\pi r_H^2 (2r_H - \alpha) (3r_H + 24P\pi r_H^3 - 3\alpha - 16P\pi r_H^2 \alpha)}{48P\pi r_H^4 + 12r_H \alpha - 48P\pi r_H^3 \alpha - 3\alpha^2 + 2r_H^2 (-3 + 8P\pi \alpha^2)} \quad (38)$$

The divergence of a black hole’s heat capacity typically signifies system instability and suggests a possible phase transition. **Figure 4(a)** shows the relationship between the heat capacity of a non-commutative Schwarzschild AdS black hole and its event horizon radius under different pressures. **Figure 4(b)** and **4(c)** illustrate the relationship between heat capacity and shadow radius under the same conditions. The results show that the heat capacity diverges at the critical point, indicating a second-order phase transition in the black hole. The similar divergence of heat capacity at the critical point observed in **Figure 4(b)** and **4(c)** demonstrates that the shadow radius can be used to identify the order of the black hole phase transition.





**Figure 4.** (a) The functional relationship between heat capacity and  $r_H$  when  $\alpha = 0.1$ . (b) The functional relationship between heat capacity and  $r_s$  when  $\alpha = 0.1$ . (c) The functional relationship between heat capacity and  $r_s$  when  $\alpha = 0.3$ . A static observer is located at  $r_0 = 100$ .

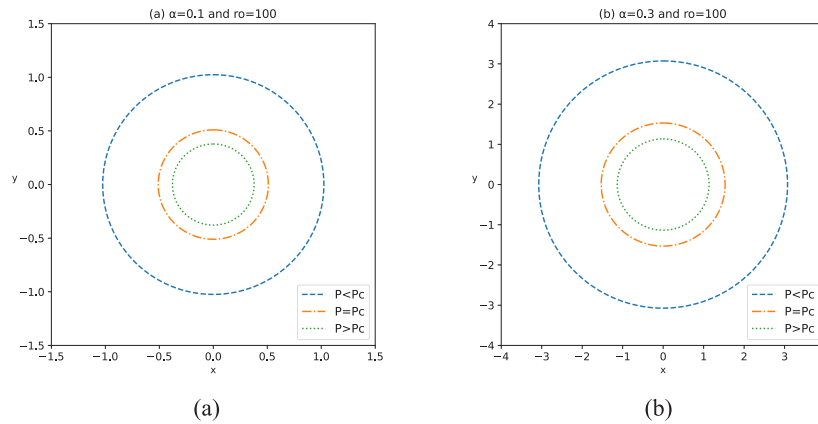
## 5. Thermal profile of the Schwarzschild-AdS BH in non-commutative geometry

This section uses the thermodynamic profile of a black hole to intuitively reveal the relationship between the black hole shadow and its phase structure. The shadow contour curve in celestial coordinates can be expressed by the formula

$$x = \lim_{r \rightarrow \infty} \left( -r^2 \sin \theta_0 \frac{d\Phi}{dr} \right) \Big|_{\theta_0 \rightarrow \pi/2} \quad (39)$$

$$y = \lim_{r \rightarrow \infty} \left( r^2 \frac{d\theta}{dr} \right) \Big|_{\theta \rightarrow \pi/2} \quad (40)$$

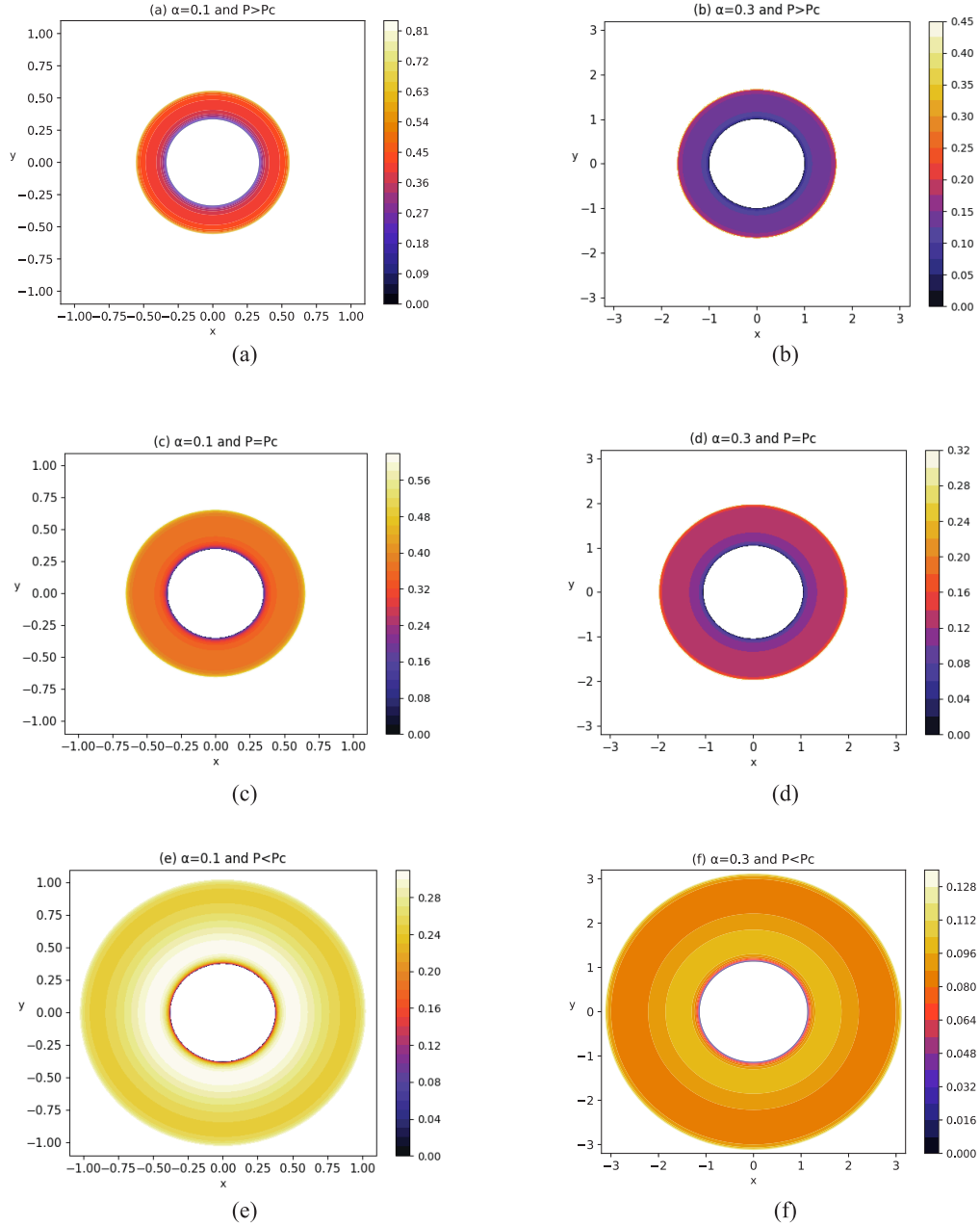
For a static observer, **Figure 5** shows the shadow contour, where the shadow radius decreases with increasing pressure. **Figure 5(a)** and **5(b)** correspond to non-commutative parameters  $\alpha = 0.1$  and  $\alpha = 0.3$ , respectively, showing that the shadow radius increases with increasing non-commutative parameter.



**Figure 5.** Shadow profiles of a non-commutative Schwarzschild-AdS black hole. (a) Non-commutative parameter  $\alpha = 0.1$ ; (b) Non-commutative parameter  $\alpha = 0.3$ . Here, the black hole mass  $M = 60$ ,  $r_0 = 100$ .

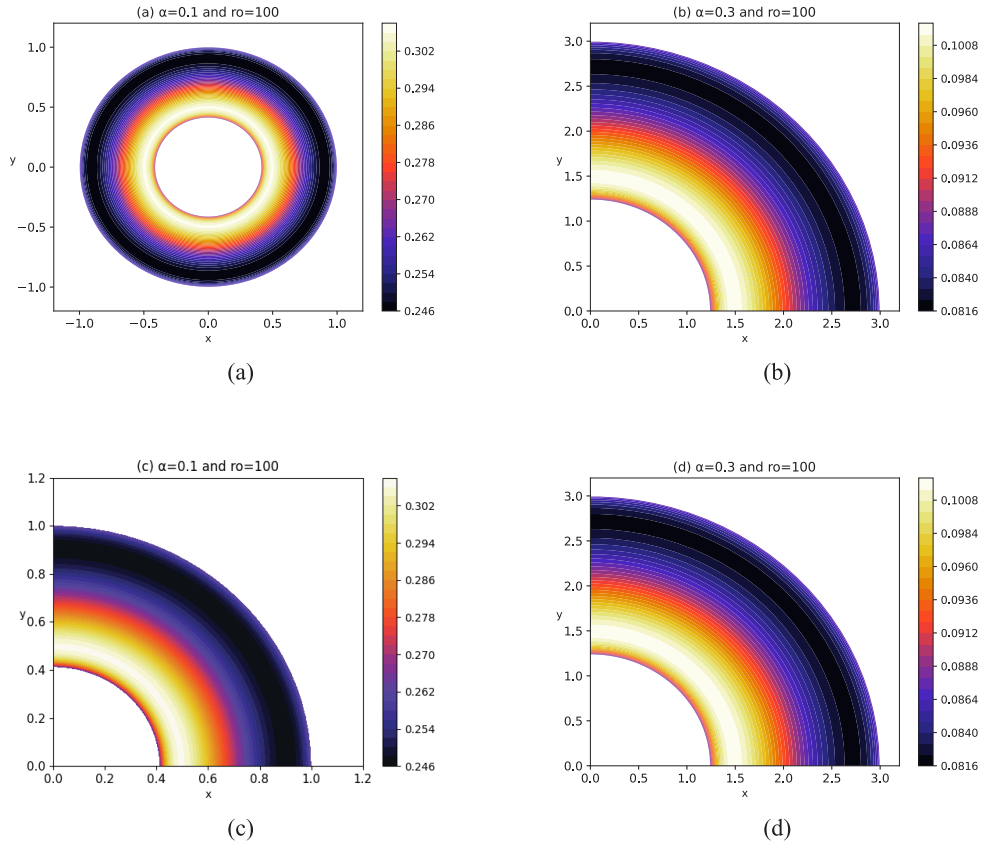
**Figure 6** presents the thermal distributions under three different pressures:  $P > P_c$  (supercritical phase),  $P = P_c$  (critical state), and  $P < P_c$ . The results are consistent with previous analyses: decreasing pressure leads to an

increased shadow radius, and increasing non-commutative parameter  $\alpha$  also leads to an increased shadow radius. **Figure 6(a)** and **6(b)** correspond to the supercritical phase ( $P > P_c$ ), where the temperature gradually increases from the center, consistent with the dashed lines in **Figure 3**. **Figure 6(c)** and **6(d)** correspond to the critical state ( $P = P_c$ ), which is thermodynamically unstable, and the temperature variation is also unstable, consistent with the dashed line segments in **Figure 3**. **Figure 6(e)** and **6(f)** correspond to the case  $P < P_c$ , where the temperature exhibits an "N-type" variation: it initially increases, then decreases, and finally increases again, consistent with the solid lines in **Figure 3**.



**Figure 6.** Thermodynamic profiles of non-commutative Schwarzschild-AdS black holes for different thermodynamic cases. (a)  $P > P_c, \alpha = 0.1$ , (b)  $P > P_c, \alpha = 0.3$ , (c)  $P = P_c, \alpha = 0.1$ , (d)  $P = P_c, \alpha = 0.3$ , (e)  $P < P_c, \alpha = 0.1$ , (f)  $P < P_c, \alpha = 0.1$ .

To more clearly show the "N-type" temperature variation, **Figure 6** restricts the shadow radius range to  $r_s^1$  to  $r_s^2$ , more prominently displaying the "N-type" change. Therefore, the shadow radius can effectively replace the event horizon radius in describing the thermodynamic properties of the black hole. The obtained results are shown in **Figure 7**. Finally, the influence of the non-commutative parameter on the phase transition of the Schwarzschild-AdS black hole can also be clearly observed from **Figure 7**, and the obtained results are consistent with those in **Figure 3**.



**Figure 7.** Thermodynamic profiles of non-commutative Schwarzschild-AdS black holes for different thermodynamic cases. (a)  $P < P_c, \alpha = 0.1$ , (b)  $P < P_c, \alpha = 0.3$ , (c)  $P < P_c, \alpha = 0.1$ , (d)  $P < P_c, \alpha = 0.3$ . (c) is a quarter of (a), (d) is a quarter of (b).

## 6. Conclusion

This paper investigates the phase transition of a Schwarzschild-AdS black hole within the framework of Non-Commutative Geometry by using the shadow radius as a substitute for the horizon radius. By utilizing the modified effective geometry incorporating the non-commutative parameter, a relationship between the shadow radius  $r_s$  and the event horizon radius  $r_h$  is established. The results indicate a positive correlation between them, implying that the black hole temperature can also be determined by the shadow radius. When  $r_0 = 100$  and the condition  $f(r_0) = 1$  is satisfied, **Equation (27)** can also be used to describe the relationship between the shadow radius and the event horizon for a Schwarzschild-AdS black hole. We present this relationship and observe that as the event horizon radius  $r_h$  increases, the shadow radius  $r_s$  also increases monotonically. Moreover, the growth trend of  $r_s$  gradually flattens out with further increase in  $r_h$ . Based on this, we argue that the shadow can be used to present

the phase transition structure of black hole thermodynamics. Based on the  $T$ – $r_h$  function, we further constructed the  $T$ – $r_s$  function and meticulously analyzed the phase transition curves of the AdS black hole. When ( $P > P_c$ ), the temperature curve is a monotonically increasing smooth curve without inflection points, indicating a supercritical phase. At the critical point ( $P = P_c$ ), an inflection point appears on the temperature curve. This curve represents the critical isobar, and the black hole is thermodynamically unstable at this inflection point. When the pressure is below the critical pressure ( $P < P_c$ ), the black hole exhibits two-phase coexistence, indicating that the black hole can exist in two distinct stable states. When the shadow radius is relatively small ( $r_s < r_{s1}$ ), it corresponds to a smaller stable black hole; when the shadow radius falls within the range  $r_{s1} < r_s < r_{s2}$ , it corresponds to a thermodynamically unstable black hole; and when the shadow radius  $r_s > r_{s2}$ , it corresponds to a larger stable black hole. The results indicate that the shadow radius  $r_s$  can be used as a substitute for the event horizon radius  $r_h$  to study the phase transition process of non-commutative Schwarzschild-AdS black holes. The results indicate that the shadow radius  $r_s$  can be used as a substitute for the event horizon radius  $r_h$  to study the phase transition process of a non-commutative Schwarzschild-AdS black hole. Subsequently, by combining the relationship between heat capacity and the event horizon radius, a relationship between heat capacity and the shadow radius is established to determine the order of the black hole phase transitions. In addition, the results also indicates that the heat capacity diverges at the critical point, implying the occurrence of a second-order phase transition in the black hole. This result is consistent with that obtained using the event horizon radius  $r_h$ . Finally, based on the  $T$ – $r_s$  function, thermal profiles of the black hole were constructed for several representative sets of non-commutative parameters. It was found that the shadow radius is closely related to the pressure, exhibiting a decrease as the pressure increases. When ( $P < P_c$ ), the temperature variation of the black hole exhibits an “increase  $\rightarrow$  decrease  $\rightarrow$  increase (N-shaped)” pattern. This indicates that the thermodynamics of the black hole can be fully captured by its thermal profile, and that the black hole shadow can indeed reflect the thermodynamic phase transition relations in the non-commutative Schwarzschild-AdS black hole. Furthermore, the influence of the non-commutative parameter on the thermodynamic phase transitions is thoroughly discussed throughout the article. Based on the above work, we conclude that the shadow can effectively serve as a substitute for the event horizon. This finding further advances further research into the thermodynamic phase transitions of black holes.

## Funding

Sichuan Province Science and Education Joint Key Project (Project No.: 25LHJJ0097)

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Abbott B, Abbott R, Abbott T, et al. (LIGO Scientific and Virgo), 2017, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral. *Phys. Rev. Lett.* 2017(119): 161101.
- [2] Abbott B, Abbott R, Abbott T, et al. (LIGO Scientific and Virgo), 2019, Search for Gravitational Waves from Intermediate-Mass Black Hole Binaries in the Second and Third Observing Runs of LIGO and Virgo. *Phys. Rev.* 2019(X9): 031040.

- [3] Gold R, et al. (Event Horizon Telescope), 2020, First M87 Event Horizon Telescope Results. V. Physical Origin of the Asymmetric Ring, *ApJ* 897, L23.
- [4] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way\*, *ApJ Lett.* 930, L12 (2022).
- [5] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. II. EHT Array, Observations, and Data Processing\*, *ApJ Lett.* 930, L13 (2022).
- [6] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. III. Imaging of the Galactic Center Supermassive Black Hole\*, *ApJ Lett.* 930, L14 (2022).
- [7] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. IV. Variability, Morphology, and Black Hole Mass\*, *ApJ Lett.* 930, L15 (2022).
- [8] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. V. Testing Astrophysical Models of the Galactic Center Black Hole\*, *ApJ Lett.* 930, L16 (2022).
- [9] Akiyama K, et al. (Event Horizon Telescope), 2022, First Sgr A Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole\*, *ApJ Lett.* 930, L17 (2022).
- [10] Bardeen J, Press W, Teukolsky S, 1972, Rotating Black Holes: Locally Nonrotating Frames, Energy Extraction, and Scalar Synchrotron Radiation, *ApJ* 178, 347 (1972).
- [11] Bozza V, 2010, Gravitational Lensing by Black Holes. *Phys. Rev. D* 82, 083005 (2010).
- [12] Ma L, Lu H, 2020, Shadow of a Rotating Black Hole in Four-Dimensional Einstein-Gauss-Bonnet Gravity. *Phys. Rev. D* 102, 066018 (2020).
- [13] Mishra A, Chakraborty S, Sarkar S, 2019, Black Hole Shadow in a Rotating Squashed Kaluza-Klein Black Hole Spacetime. *Phys. Rev. D* 100, 104054 (2019).
- [14] Gralla S, Holz D, Wald R, 2019, Black Hole Shadows, Photon Rings, and Lensed Images. *Phys. Rev. D* 100, 024012 (2019).
- [15] Jusufi K, Werner M, Banerjee A, et al., 2017, Shadow of a Rotating Five-Dimensional Black Hole in Braneworld. *Phys. Rev. D* 96, 024036 (2017).
- [16] Jusufi K, Sarkar N, Rahaman F, et al., 2018, Shadow of a Charged Rotating Black Hole in Conformal Gravity. *Phys. Rev. D* 97, 104028 (2018).
- [17] Pantig R, Yu P, Rodulfo E, et al., 2022, Shadow and Deflection Angle of Rotating Black Holes in Einstein-Maxwell-Weyl Gravity. *Phys. Rev. D* 105, 044034 (2022).
- [18] Perlick V, Tsupko O, 2022, Gravitational Lensing by Black Holes: A Review. *Phys. Rep.* 963, 1 (2022).
- [19] Synge J, 1966, The Escape of Photons from a Schwarzschild Field. *Mon. Not. R. Astron. Soc.* 131, 463 (1966).
- [20] Luminet J, 1979, Image of a Spherical Black Hole with Thin Accretion Disk. *Astron. Astrophys.* 75, 229 (1979).
- [21] Wang H, Wei S, 2022, Shadow of a Kerr-Newman Black Hole Surrounded by a Plasma. *Phys. Rev. D* 106, 024032 (2022).
- [22] Wang H, Xu Y, Wei S, 2019, Shadow of a Rotating Black Hole in a Dark Matter Halo. *Phys. Rev. D* 100, 064052 (2019).
- [23] Narayan R, Johnson M, Gammie C, 2019, Black Hole Shadows and Photon Rings: Observational Prospects. *ApJ* 884, L33 (2019).
- [24] Guo S, Huang Y, Han Y, et al., 2023, Black Hole Shadow as a Probe of Dark Matter. *ApJ* 945, L23 (2023).
- [25] Meng Y, Kuang X, Wang X, et al., 2023, Shadow of a Rotating Black Hole with a Cosmological Constant. *Phys. Rev. D* 107, 064059 (2023).

- [26] Zeng W, Ling Y, Jiang Q, et al., 2023, Shadow of a Black Hole in  $f(R)$  Gravity. *Phys. Rev. D* 107, 084076 (2023).
- [27] Zeng X, Zhang H, Zhang H, 2020, Shadow of a Rotating Black Hole in Einstein-Aether Theory. *Phys. Rev. D* 101, 044074 (2020).
- [28] Zeng X, Zhang H, 2020, Shadow of a Black Hole in the Presence of a Global Monopole. *Phys. Rev. D* 102, 064033 (2020).
- [29] He K, Guo S, Tan S, et al., 2022, Shadow of a Charged Black Hole with a Topological Defect. *Phys. Rev. D* 105, 104064 (2022).
- [30] Peng J, Guo M, Feng X, 2021, Shadow of a Rotating Black Hole in Horndeski Gravity. *Phys. Rev. D* 104, 024058 (2021).
- [31] Li G, He K, 2021, Shadow of a Black Hole in the Einstein-Maxwell-Scalar Theory. *Phys. Rev. D* 104, 024052 (2021).
- [32] Zhou X, Chen S, Jing J, 2021, Shadow of a Rotating Black Hole in the Ghost-Free Bimetric Gravity. *Phys. Rev. D* 104, 024029 (2021).
- [33] Zeng X, Yang C, Huang Y, et al., 2025, Black Hole Shadow and Thermodynamic Phase Transition. *Phys. Rev. D* 111, 044064 (2025).
- [34] Li G, He K, 2021, Shadow of a Black Hole in the Einstein-Weyl Gravity. *Phys. Rev. D* 104, 044021 (2021).
- [35] He K, Yang C, Zeng X, 2025, Correlation between Black Hole Shadow and Thermodynamic Criticality. *Phys. Rev. D* 111, 044078 (2025).
- [36] Gan Q, Wang P, Wu H, et al., 2021, Shadow of a Rotating Black Hole in the Einstein-Born-Infeld Gravity. *Phys. Rev. D* 104, 044073 (2021).
- [37] Guo S, He K, Li G, et al., 2021, Shadow of a Black Hole in the Einstein-Maxwell-Dilation Theory. *Phys. Rev. D* 104, 044024 (2021).
- [38] Zeng X, He K, Li G, et al., 2022, Black Hole Shadow in the Presence of a Magnetic Field. *Eur. Phys. J. C* 82, 764 (2022).
- [39] Hawking S, 1976, Particle Creation by Black Holes. *Commun. Math. Phys.* 43, 199 (1975), [Erratum: *Commun. Math. Phys.* 46, 206 (1976)].
- [40] Bekenstein J, 1994, Black Holes: Classical Properties, Thermodynamics, and Heuristic Quantization. *Phys. Rev. D* 49, 1912 (1994).
- [41] Kubiznak D, Mann R, 2012, P-V Criticality of Charged AdS Black Holes. *JHEP* 07, 033 (2012).
- [42] Cai R, Cao L, Li L, et al., 2013, P-V Criticality in the Extended Phase Space of Gauss-Bonnet Black Holes in AdS Space. *JHEP* 09, 005 (2013).
- [43] He K, He X, Hu X, et al., 2019, Thermodynamic Geometry and Critical Phenomena of AdS Black Holes with Nonlinear Electrodynamics. *Chin. Phys. C* 43, 125101 (2019).
- [44] Ökcü B, Aydinler E, 2024, Joule-Thomson Expansion of AdS Black Holes. *Eur. Phys. J. C* 77, 24 (2024).
- [45] Wei S, Liu Y, 2018, Thermodynamic Phase Transition of Kerr-AdS Black Holes. *Phys. Rev. D* 97, 104027 (2018).
- [46] Zhang M, Guo M, 2020, Thermodynamic Geometry of AdS Black Holes in  $f(R)$  Gravity. *Eur. Phys. J. C* 80, 790 (2020).
- [47] Belhaj A, Chakhchi L, Moumni H, et al., 2020, Thermodynamic Criticality of AdS Black Holes in Higher Dimensions. *Int. J. Mod. Phys. A* 35, 2050170 (2020).
- [48] Guo S, Li G, Liang E, 2022, Thermodynamic Phase Transition of Black Holes with a Cosmological Constant. *Phys. Rev. D* 105, 023024 (2022).
- [49] Nicolini P, 2009, Noncommutative Black Holes: A Review. *Int. J. Mod. Phys. A* 24, 1229 (2009).



- [50] Snyder H, 1947, Quantized Space-Time. *Phys. Rev.* 71, 38 (1947).
- [51] Li G, He J, Chen B, 2021, Noncommutative Black Hole Thermodynamics. *Chin. Phys. C* 45, 015111 (2021).
- [52] Aschieri P, Blohmann C, Dimitrijevic M, et al., 2005, A Gravity Theory on Noncommutative Spaces. *Class. Quant. Grav.* 22, 3511 (2005).
- [53] Aschieri P, Dimitrijevic M, Meyer F, et al., 2006, Noncommutative Geometry and Gravity. *Class. Quant. Grav.* 23, 1883 (2006).
- [54] Zeng X, Zeng G, Li G, et al., 2022,, Noncommutative Corrections to Black Hole Shadow. *Nucl. Phys. B* 974, 115639 (2022).
- [55] Zeng X, Aslam M, Saleem R, 2023, Noncommutative Black Hole Shadow and Thermodynamic Phase Transition. *Eur. Phys. J. C* 83, 129 (2023).
- [56] Nicolini P, Smailagic A, Spallucci E, 2005, Noncommutative Geometry Inspired Schwarzschild Black Hole. *Phys. Lett. B* 632, 547 (2005).
- [57] Nozari K, Mehdipour S, 2008, Noncommutative Inspired Black Holes in Extra Dimensions. *Class. Quant. Grav.* 25, 175015 (2008).
- [58] Wang R, Ma S, You L, et al., 2025,, Noncommutative Corrections to Kerr Black Hole Shadow. *Chin. Phys. C* 49, 065101 (2025).
- [59] Anacleto M, Brito F, Campos J, et al., 2023, Noncommutative Black Hole Thermodynamics and Phase Transition. *Eur. Phys. J. C* 83, 298 (2023).
- [60] Zheng H, Mou P, Chen Y, et al., 2023, Black Hole Shadow in Noncommutative Geometry. *Chin. Phys. B* 32, 080401 (2023).

**Publisher's note**

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

# Study on Performance Optimization of Radiation Protection Materials Based on Nanotechnology

Zhengyang Yuxiong

Xihua University, Chengdu 610039, Sichuan, 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 development of nuclear energy development, medical treatment and aerospace, radiation protection materials are developing towards light weight, high efficiency and high stability. In the past, traditional protective materials have been difficult to adapt to the needs of complex working conditions. Based on the application of nano-technology radiation protection materials, this paper discusses the preparation technology and engineering regulation of nano-radiation protection materials, and explores the optimization path of the performance of radiation protection materials based on nano-technology, which provides a reference for R&D and application of high-performance nano-radiation protection materials.

**Keywords:** Nanotechnology; Radiation protection materials; Performance optimization

**Online publication:** April 22, 2026

## 1. Introduction

Traditional radiation protection materials are mainly lead-based materials, conventional concrete, and common polymer substrates. Although they can achieve the basic radiation shielding effect, they cannot meet the core requirements of modern engineering for lightweight, multifunctional integration and long service life due to their common defects such as high density, poor flexibility and biological toxicity of some components. Nano-materials have ultra-high specific surface area, rich interface defect structure and special electronic configuration characteristics, which can effectively improve the interaction probability between materials and radiation particles, strengthen the attenuation and dissipation efficiency of radiation energy, and at the same time, through the precise design of microstructure, the mechanical properties and thermal properties of materials can be simultaneously optimized <sup>[1]</sup>. Compared with traditional micron-sized fillers, nano-sized functional components can achieve the same benefits with a lower doping amount, and even show better shielding effect, which can effectively improve the stability of the structure while ensuring the light weight of the material. This feature can make it a core technical direction to break through the performance bottleneck of traditional protective materials. Based on this, it is of great practical significance to systematically sort out the preparation process of nano-radiation protection materials and explore the optimization path of radiation protection materials based on nanotechnology.

## **2. Preparation process and engineering control of nano radiation protection materials**

In the application of materials engineering, the preparation technology of nano-radiation protection materials can be divided into three categories: controllable synthesis of nano-functional fillers, modification of matrix materials and molding of composite systems. Different processes correspond to different performance control objectives, and suitable preparation paths can be selected according to actual application requirements. The preparation process of engineering protective materials not only needs to focus on the material properties of experimental scale, but also needs to take into account the stability of large-scale production, effective control of economic costs and adaptability of processing and application, so as to avoid the excellent laboratory performance, but it is difficult to achieve efficient landing production.

### **2.1. Controllable synthesis technology of nano-functional fillers**

Liquid-phase precipitation method is a common method to prepare metal oxide nano-fillers, which can be suitable for the synthesis of lead oxide, ferro-ferric oxide, titanium dioxide and other nano-particles. By accurately adjusting the temperature and pH value of the reaction system and the relative ratio of reactant feeding, the particle size can be effectively controlled at 5–50nm, and nano-particles with good dispersibility, high purity and industrial mass production can be prepared. The practical application of this process can effectively inhibit the agglomeration effect of particles by adding dispersant, and can significantly improve the uniformity of products, making this preparation technology more suitable for large-scale preparation of basic nano-shielding fillers, and can be widely used in the production of industrial-grade radiation protection plates.

Hydrothermal and solvothermal methods are mostly used in the preparation of high crystallinity nano-materials, which can synthesize rare earth nanoparticles such as samarium oxide and cerium oxide and special structural materials such as boron nitride nanotubes. The products have regular crystals, low internal defects and excellent radiation shielding performance, which is suitable for the material preparation requirements of high-end radiation protection scenes. Rare earth nanoparticles can play a unique role in the protection of mixed radiation sites due to their high atomic number and moderate neutron capture cross section, but owing to their high cost of synthesis and production, they are mostly used in aerospace, nuclear medicine and other fields with high performance requirements.

Electrospinning technology can be used to prepare nanofiber protective materials, such as boron-doped polyvinyl alcohol nanofibers, carbon-based nanofibers, etc. The fiber diameter can be controlled at 100–500nm, forming a continuous three-dimensional network structure, which makes the materials have high-efficiency radiation shielding ability and good flexibility <sup>[2]</sup>. This three-dimensional network structure can provide more scattering and absorption sites for radiation particles, and at the same time, it can significantly improve the flexibility and tear resistance of the material, making it more suitable for the production and application of wearable radiation protection equipment.

### **2.2. Forming and processing technology of nano-composite materials**

Solution blending method can obtain flexible composite protective materials by dispersing nano-functional fillers in polymer solution and curing. This process is simple and low in production cost, and can realize uniform dispersion of nano-fillers, which is more suitable for composite modification of rubber, epoxy resin and other substrates. In practical production application, this process can effectively combine the ultrasonic dispersion with

high-speed stirring technology to further reduce the agglomeration probability of nanoparticles, thus effectively improving the uniformity of the overall properties of composites.

Melt extrusion is mainly aimed at thermoplastic substrates such as polypropylene and polyethylene. With the help of twin-screw extrusion equipment, nano-fillers can be uniformly dispersed in the substrate, and shaped components such as plates and pipes can be continuously produced to meet the large-scale production requirements of industrial radiation protection components. In practical application, this technology is more suitable for the application scene of automatic production line. With the support of this technology, the dispersion effect of filler can be further optimized by adjusting the rotation speed and extrusion temperature of the screw, and it is also the mainstream technology for industrial mass production of nano-composite protective materials at present.

In-situ synthesis method can effectively avoid the agglomeration problem in the process of early dispersion of filler by directly synthesizing nanoparticles in the matrix. For example, in-situ synthesis of cerium-tungsten composite nanoparticles in the matrix of regenerated collagen fibers can significantly improve the interfacial bonding force between filler and matrix and ensure the uniformity of radiation shielding performance. The practical application of in-situ synthesis method can further promote the chemical bonding between nanoparticles and matrix, help to greatly improve the transmission efficiency of interfacial stress, and effectively solve the problems of easy delamination and poor mechanical properties of traditional composites.

### **2.3. Engineering design and control of nanostructures**

Gradient nanostructures, layered nanostructures and porous nanostructures are important structural design ideas to optimize the comprehensive properties of materials. Gradient nano-metallic materials can build a high-density stacking fault and dislocation network through surface nano-crystallization, and with the help of adaptive martensite transformation, the rapid annihilation of irradiation defects can be realized. At a high irradiation dose of 155 dpa, the internal defect density is only 3.8% of that of traditional coarse-grained materials. This structural design can make the surface layer of nano-materials have more excellent radiation resistance, further maintain the good mechanical strength of the core, and promote the synergy between the function and structure of materials.

Layered nanocomposites are designed with “brick-mortar” structure, and combined with the radiation shielding effect of nano-lamellae and the energy dissipation characteristics of the interface, the mechanical properties and shielding properties of the materials are improved synergistically. By adjusting the thickness and component ratio of each layer, the graded shielding of different energy rays can be realized through customization, which enables the material adapt to a more complicated radiation environment.

Porous nanostructures can not only reduce the overall density of materials, but also increase the scattering sites of radiation particles, thus achieving efficient radiation protection under the premise of lightweight. The design and application of porous structure can also significantly improve the thermal insulation and permeability of materials, which has significant application advantages in the preparation of wearable protective equipment.

## **3. Optimization path of radiation protection materials based on nanotechnology**

### **3.1. Optimization strategy of radiation shielding performance**

Doping nano-sized components with high atomic number is the core way to improve  $\gamma$ -ray shielding ability. Doping nano-sized particles such as lead oxide, bismuth oxide and tungsten trioxide into polymer or glass matrix can effectively improve the effective atomic number and linear attenuation coefficient of materials. Relevant

experimental data show that the linear attenuation coefficient of 0.059 keV  $\gamma$ -ray doped with 25 wt% nano-lead oxide in polystyrene matrix is increased by 26.7% and the half-value layer thickness is decreased by 37% compared with that of micron-sized lead oxide composites with the same proportion. The small size effect of nanoparticles can greatly increase the interaction frequency with  $\gamma$  photons, which makes them significantly improve the radiation shielding efficiency under the same doping amount, and even much higher than micron-sized fillers. At the same time, they can effectively reduce the material density.

The optimization of neutron shielding performance depends on boron-based and rare-earth-based nano-materials. Materials such as boron nitride nanotubes and nano-samarium oxide have high neutron capture cross sections. Nano-treatment can improve the dispersion uniformity of functional elements and enhance neutron absorption efficiency. Compared with traditional materials, the orderly arrangement of boron nitride nanotube films significantly increases the density by 3 times, and the neutron shielding performance can be effectively improved by 3.7 times, which is more suitable for the lightweight protection requirements in the aerospace field. Among them, boron has a strong ability to capture thermal neutrons. When it is applied to the optimization of radiation shielding performance of materials, nano-crystallization can significantly avoid the segregation of components and effectively ensure the uniformity of neutron shielding performance, which makes it more suitable for spacecraft cabins, nuclear reactor peripheral protection and other scenes.

Multi-component nano-composite system integrates the advantages of high atomic number nanoparticles, neutron absorbing nano-components and carbon-based nano-materials, which can simultaneously achieve high-efficiency shielding of gamma rays and neutrons, make up for the short comings of radiation absorption of a single material, and is suitable for the protection scene of mixed radiation field of nuclear facilities. By using the collaborative design of multi-elements, the integrated protection against wide-spectrum radiation can be further realized, and the limitation of single shielding of traditional materials can be effectively solved.

### 3.2. Optimization method of mechanical properties

One-dimensional and two-dimensional nano-materials can be used as reinforcing phases to improve the mechanical properties of composites. Carbon nanofibers, graphene and boron nitride nano-materials can form a continuous stress transfer network in the collective, effectively improving the tensile strength and impact toughness of materials. Relevant experiments show that adding 0.5 wt% carbon nanofibers to samarium oxide/epoxy resin composites can improve the tensile strength of the materials by 18% and the impact toughness by 22%. The length-diameter ratio advantage of one-dimensional nano-materials can effectively transmit external stress, and the lamellar structure of two-dimensional nano-materials can effectively hinder the crack propagation. The cooperative application of the two materials can significantly improve the structural stability of materials.

Interface modification is an important means to improve the bonding state between filler and matrix. Surface treatment of nano-filler with silane coupling agent, oleic acid and other reagents can reduce the surface energy of particles, improve wettability with matrix, reduce interface defects and avoid material cracking caused by stress concentration. The optimized way of interface modification can effectively eliminate the gap between filler and matrix, significantly improve the bonding strength of interface structure, and make composite materials less prone to performance attenuation under long-term irradiation and mechanical load, further ensuring the performance stability.

Gradient and layered nanostructures improve the radiation deformation resistance of materials through the interface slip and energy dissipation mechanism, and solve the problems of high brittleness and easy fracture of



traditional protective materials <sup>[3]</sup>. The optimal design of mechanical properties of this structure can make the material dissipate energy through the slip between layers when it is impacted, and at the same time effectively guarantee the integrity of the whole material.

### **3.3. Optimization of thermal stability and radiation aging resistance**

Modification of nano-ceramic particles can effectively improve the thermal stability of polymer matrix. Nano-alumina, silica and other particles can effectively improve the thermal decomposition temperature and thermal oxygen stability of matrix, further inhibit the fracture and degradation of polymer molecular chains under irradiation environment, and promote the extension of service life of materials. Nano-ceramic particles can be used as a physical barrier of protective materials, which can effectively hinder the heat transfer and the diffusion of oxygen molecules, thus reducing the aging rate of polymers under high temperature irradiation.

The design of defect annihilation structure is the key link to improve the radiation resistance of metal-based nano-materials. The high-density stacking faults and dislocation networks in nano-metals can quickly annihilate defects such as vacancies and interstitial atoms during irradiation, effectively avoiding the problems of swelling and embrittlement of materials during irradiation. The design and application of high-density defect network can further realize the self-repair of irradiation damage, so that metal materials can still maintain good plasticity and strength in extreme irradiation environment.

The core-shell structure nanoparticles protect the functional components of the inner core through the thermal resistance and radiation resistance of the outer shell, and improve the stability of the material in the dual extreme environment of high temperature and high radiation. The application of this structure in the optimization of thermal stability and radiation aging resistance can realize the precise protection of functional components, effectively avoid the inactivation of core barrier elements in complex environments, and significantly improve the environmental adaptability of materials.

## **4. Environmental adaptability and optimization of green engineering application**

In addition to the performance optimization system, the environmental adaptability and green development of nano-radiation protection materials are also crucial to its engineering development. At present, most of the related research focuses on the performance test in the laboratory standard environment, and the research on the complex scene of practical application is shallow. In addition to the laboratory variables, the actual scene is often accompanied by complex environments such as low temperature alternation, high humidity, salt spray and mechanical vibration. At this time, the long-term stability of materials still needs systematic verification and consideration. By introducing weather-resistant modified components and interface stabilization treatment, nano-composite materials can maintain stable shielding effectiveness and mechanical properties in the range of -40°C to 120°C, and further reduce the performance attenuation caused by environmental factors.

At the same time, green and non-toxic is also an important direction of industry development, and the traditional nano-fillers with lead components have potential biological toxicity and environmental risks. In order to solve this safety dedication, lead-free components such as nano-tungsten, nano-bismuth and rare earth oxides can be used to replace lead-based fillers. In this optimization design, shielding efficiency can be guaranteed, and biosafety targets in medical and wearing scenes can be better met, which is more in line with the industry orientation of environmental protection and sustainable development. With the application of industrialization, the



evaluation of the whole life cycle of materials and cost control cannot be ignored. The preparation cost of high-end nano-fillers is high, and the dispersion uniformity in large-scale production is difficult to control effectively, which is a realistic bottleneck hindering the development of the industry. Developing low-energy synthesis technology and high-efficiency dispersion technology can effectively promote the nano-recycling of waste protective materials and further reduce production costs and resource consumption.

## 5. Conclusion

To sum up, nanotechnology can provide a feasible path for the performance breakthrough of radiation protection materials from the microstructure level. Through technical means such as nano-component compounding, nano-structure construction and interface modification, the radiation protection, mechanical and thermal stability of materials can be cooperatively optimized, which further meets the high-end application requirements of protection materials in nuclear energy, medical care, aerospace and other fields. At present, in the actual research and development of nano radiation protection materials, there are still practical problems such as uneven dispersion of fillers in large-scale production, high economic cost of high-end nano raw materials and long-term service performance to be verified in extreme environments. In the future, it is necessary to carry out more in-depth research and exploration around engineering and adaptability. On the one hand, it is necessary to further optimize the preparation process, improve the controllability of large-scale production costs, and promote the effective transformation of laboratory results into industrial-grade products; On the other hand, it is necessary to combine relevant application scenarios to carry out multi-functional integrated optimization design, and further endow materials with additional functions such as radiation sensing, self-repair and antibacterial. With the continuous reform and innovation of nano-preparation technology and structural control means, it can promote the high-quality development of nano-radiation protection materials with light weight, high efficiency and multi-function, provide more reliable technical support for the safety protection of high-radiation scenes, and help the green and high-end innovative development of related industries.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Zhang Y, Zhu Y, Wang Y, et al., 2026, Research on Radiation-Hardened Multi-Bit Flip-Flop Design Technology for Nano-FinFET Process. *Modern Applied Physics*, 17(1): 154–161.
- [2] Nie F, Ren Y, Zhong Z, et al., 2025, Board-Level Nano Protection Technology, Sichuan Institute of Electronics, SMT/MPT Special Committee of Sichuan Institute of Electronics, Proceedings of the 18th China High-End SMT Academic Conference 2025, 244–253.
- [3] Zhao Z, Zhang S, 2024, Research Progress on Protective Materials for Chemical Protective Clothing. *Journal of Guangdong University of Petrochemical Technology*, 34(1): 96–101.

**Publisher’s note**

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

# Integrated Volt-Energy Storage-Lighting System for Smart Road Lighting Based on Distributed MPPT: A Review of Hardware Design and Economic Analysis

Xiangran Chen<sup>1†</sup>, Jierui Feng<sup>2†</sup>, Yuying Pan<sup>3</sup>, Mingwei Li<sup>4</sup>, Yanran Li<sup>5</sup>, Yuhao Song<sup>6</sup>

<sup>1</sup>School of Economics and Management, Chongqing Jiaotong University, Chongqing, China

<sup>2</sup>School of Civil Engineering, Chongqing Jiaotong University, Chongqing, China

<sup>3</sup>College of Chongqing Jiaotong University, Chongqing, China

<sup>4</sup>School of Architecture and Planning, Chongqing Jiaotong University, Chongqing, China

<sup>5</sup>School of Economics and Management, Chongqing Jiaotong University, Chongqing, China

<sup>6</sup>Chongqing Jiaotong University School of Economics and Management, Chongqing, China

<sup>†</sup>These authors made equal contributions to this work.

**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:** Smart street lights are the key carrier of smart cities and dual carbon goals. Aiming at the problems of high energy consumption of street lamps, extensive control, lack of dynamic adaptation capabilities of existing smart street lamps, and the vulnerability of centralized photovoltaics to shadows and module aging, this paper proposes a distributed MPPT photovoltaic-energy storage-lighting integrated system. The system hardware integrates high-precision sampling, STM32 main control, DC-DC conversion and lithium iron phosphate battery management; The software integrates traffic flow monitoring, people flow statistics and multi-factor linear regression dimming algorithm to realize perception and dynamic dimming. Through the three-layer architecture of perception-processing-cloud, the system not only improves the efficiency of photovoltaic power generation and adapts to complex lighting, but also has been preliminarily verified to have significant technical feasibility and economic value in terms of power generation efficiency, energy saving and cost reduction, and full life cycle costs.

**Keywords:** Distributed MPPT; Smart street lamps; Hardware design; Investment income analysis; Shadow inhomogeneity

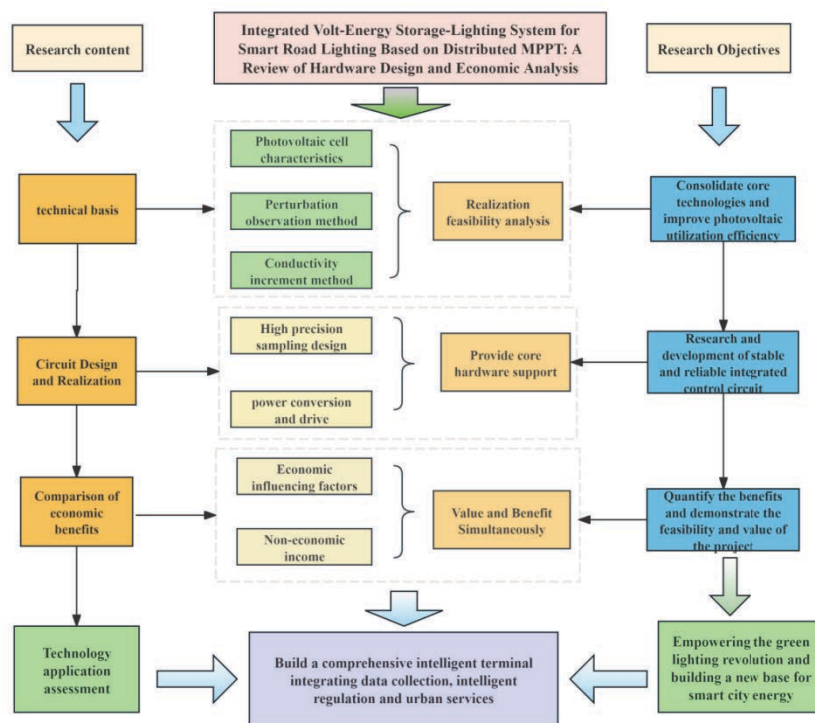
**Online publication:** April 22, 2026

## 1 . Introduction

Under the guidance of the national strategic goal of “carbon peaking in 2030 and carbon neutrality in 2060”, the construction of smart cities has put forward urgent requirements for the intelligence and energy saving of road lighting systems<sup>[1]</sup>. As a clean distributed energy source, the application of photovoltaic energy in street lighting scenarios has become an important trend. However, centralized photovoltaic power supply systems are easily affected by local shadows in complex cities, resulting in a sharp drop in overall power generation efficiency,

and widespread extensive control and unreasonable layout, problems such as high operation and maintenance costs. At present, my country's traffic street lamps have a large scale and significant energy-saving potential, but most of the research on smart street lamps is still in the initial stage, lacking an intelligent control system that integrates perception and human and vehicle flow, and fails to optimize energy utilization. In response to the above bottlenecks, this paper proposes a "photovoltaic-energy storage-lighting integration" smart street lighting system, which improves the power generation efficiency under shadow conditions through distributed MPPT technology, and combines intelligent control algorithms to achieve "on-demand lighting" <sup>[2]</sup>. At present, relevant research mostly focuses on algorithm simulation and conceptual design, and there are still obvious gaps in low-cost MPPT hardware circuits suitable for single lamps, multi-module integration adaptability, and quantitative analysis of full life cycle economic benefits in complex scenarios.

This paper will carry out research from four dimensions: theoretical basis, hardware design, economic benefits and conclusion outlook, explain the core technology of the system in turn, design the hardware circuit of the distributed MPPT controller in detail, build an economic benefit comparison model and conduct quantitative analysis, and finally summarize the technical advantages And look forward to the future direction, in order to provide support for the practical application and promotion of smart street lighting systems (**Figure 1**).



**Figure 1.** Full process flowchart.

## 2. Key technology basis of distributed MPPT system

### 2.1. Photovoltaic cell characteristics and maximum power point tracking principle

The output power of photovoltaic cells is significantly affected by factors such as irradiance, temperature and shadow, and its maximum power point (MPP) will drift dynamically with changes <sup>[3]</sup>. The higher the irradiance, the larger the corresponding power value of MPP, and the MPP will move along the power-voltage curve; The

increase in temperature will cause the open circuit voltage of the photovoltaic module to drop, make the MPP shift to the low voltage direction, and at the same time its power peak. In addition, the shadow may cause the “hot spot effect” of the photovoltaic module, resulting in multiple peaks in the output characteristic curve, making the MPP distribution scattered and unstable, and seriously affecting the overall output power of the system. Maximum power point tracking(MPPT) technology has become the key to improve the efficiency of photovoltaic utilization.

The basic working principle of MPPT is to collect the output voltage and current of the photovoltaic panel in real time through the sampling circuit, calculate the instantaneous output power, and judge the position deviation between the current working point and the real-time maximum power point according to the specific tracking algorithm, and determine the adjustment direction of the voltage or current <sup>[4]</sup>. After that, by controlling the duty cycle of the DC-DC power converter circuit, the equivalent load of the photovoltaic panel is changed, and the working voltage and current of the photovoltaic panel are dynamically adjusted, so that the working point of the system is close to and stable near the maximum power point. The closed-loop control can realize the state that the PV panel is still close to the maximum power output when the parameters change.

## **2.2. Mainstream MPPT algorithms and feasibility analysis of hardware implementation**

### **2.2.1. Perturbation observation method**

By changing the output voltage (or current) of the photovoltaic cell periodically, the disturbance observation method detects the power change trend, so that the disturbance direction gradually approaches the MPP <sup>[5]</sup>. This algorithm is simple in principle, does not need the accurate photovoltaic parameter model, the hardware sampling link is relatively simple, and the dynamic response is fast. It is the least difficult MPPT algorithm to implement on the microcontroller side <sup>[6]</sup>. Core steps such as voltage/current sampling, power calculation, comparison, and duty cycle adjustment can be completed only by relying on MCU general-purpose peripherals, and even 8-bit or 16-bit low-cost MCUs can be competent. The algorithm runs stably, does not have the problem of integral saturation or parameter drift, and occupies low resource. However, it also has some shortcomings, such as steady-state oscillation, tracking misjudgment when the irradiance changes rapidly, and the contradiction between the disturbance step size and the tracking performance. Its core parameters (disturbance step size and sampling interval) have clear requirements for hardware sampling. If the step size is too small, high-resolution ADC is required, and if it is too large, it will cause steady-state loss; If the sampling interval is too short, unstable transient values may be collected, resulting in power judgment errors, and if the sampling interval is too long, the speed will be tracked. For single-rod PV systems (whose PV array transient response time is generally 10–50ms), the sampling interval is usually set to 50-200ms. Considering the low-cost requirements, a disturbance step size of 1–3% can generally be used with a 10-bit ADC; The sampling interval is adapted to the transient response time of the system and is set to 50–200ms. Under this requirement, the universal ADC of low-cost MCU is fully sampled rate required.

### **2.2.2. Conductance increment method**

The conductance increment method is based on the mathematical model of the P-U characteristics of the photovoltaic array <sup>[7]</sup>. Based on the theory of the differential of power to voltage at the MPP, the position of the operating point relative to the MPP is judged by calculating the relationship between the instantaneous conductance and the conductance increment. The algorithm has good steady-state performance, strong anti-



interference ability, and accurate dynamic tracking, but the principle is more complex, and the calculation amount is large, which requires high hardware sampling accuracy and MCU computing ability. It usually requires 12-bit and above high-resolution ADCs, low-drift voltage/current sensors, and precision signal conditioning circuits, and relies on 32-bit high-performance MCUs to perform floating-point operations<sup>[8]</sup>. At the same time, the algorithm has strict requirements on sampling rate and synchronization control, resulting in significant hardware cost and system design complexity, which is difficult to match the application requirements of low-cost, single-pole photovoltaic systems.

### 2.2.3. Other algorithms

Intelligent MPPT algorithms such as fuzzy control, particle swarm optimization, and neural networks, although they can effectively deal with the tracking problems of the algorithm under complex conditions, and have the advantages of strong anti-interference ability, good global optimization ability, and fast tracking speed, they are common<sup>[9–11]</sup>. There are problems such as high computing complexity, strict hardware requirements, and difficult debugging. These algorithms usually rely on high-performance MCU or even DSP support, and the hardware cost can reach 5–10 times that of the disturbance observation method. It is an excess performance and low applicability in low-cost single-rod photovoltaic systems with conventional illumination<sup>[12]</sup>. On the other hand, although the hardware implementation of simple algorithms such as the constant voltage method is simpler, they cannot adapt to the changes of irradiance and temperature, and the tracking accuracy is very low<sup>[13]</sup>. They are only suitable for ultra-low-cost simple systems, and it is difficult to meet the power output requirements of conventional photovoltaic systems.

To sum up, the perturbation-observation method and its variants are the preferred algorithms for low-cost, single-pole photovoltaic systems in hardware implementation. The main reason is that the algorithm has the lowest implementation cost, only 8-bit/16-bit general-purpose MCU, 10-bit ADC and basic voltage/current sampling link are required, no additional expensive hardware is required, and the algorithm implementation and debugging are relatively simple, the code volume is small, and the logic is clear, even entry-level embedded personnel can get started quickly, suitable for low-cost design requirements of rapid mass production<sup>[14]</sup>. The steady-state oscillation problem existing in its version can be effectively optimized through improved algorithms such as variable step size and hysteresis disturbance, so that the steady-state oscillation loss is reduced to less than 1%, and at the same time, the dynamic response speed is good. Power utilization target.

## 2.3. Selection of energy storage technology

As an energy storage unit, the new high-cycle lithium iron phosphate battery performs outstandingly in outdoor work: its charge and discharge cycles can reach 2,000times (remaining capacity rate > 80%)<sup>[15]</sup>. Excellent thermal stability, decomposition temperature above 200 °C, not easy to catch fire or short circuit in the case of overcharge, overdischarge or short circuit; The voltage platform is stable (the voltage in the mid-discharge period is maintained at 3.0V–3.2V), and the output power is stable; It supports 0.2C–1C conventional charging and discharging, has no memory effect, can be charged at any time, and can work stably in a wide temperature range of -20°C to 60°C. These characteristics make it highly suitable for the working conditions of street lamps in harsh outdoor conditions.

The system adopts a “constant current-constant voltage” charging management strategy: in the charging stage, it first charges at a constant current of 0.3C to 3.65V, and then switches to constant voltage charging

until the current drops to  $0.05C$ , so as to avoid overcharging and damage to the battery <sup>[16]</sup>. In the discharge stage, a hierarchical discharge strategy is adopted. Combined with the energy consumption demand of street lamps (the working current of 60W LED lamps is about 5A), 20% of the remaining power is set as the lower discharge limit. The system supports dual power supply access of mains power and solar energy at the same time, ensuring that the basic lighting can be maintained for no less than 8 hours after the mains power is cut off. The charging and discharging management strategy can realize closed-loop control through the timer and ADC sampling of STM32 single-chip microcomputer, which is conducive to prolonging the battery life and operation and maintenance costs.

### **3. Design and implementation of hardware circuit for distributed MPPT controller**

#### **3.1. Overall hardware architecture of the system**

The system adopts an independent single-pole power supply structure. Its core energy management and transmission path is: the DC energy output by the photovoltaic panel is first optimized by the MPPT charging controller, and then charged for the lithium iron phosphate battery energy storage unit. As the energy center of the system, the energy storage unit provides a stable power supply for subsequent LED drives and lighting loads, as well as various intelligent perception and control modules <sup>[17]</sup>. As the core of the power supply system, MPPT controller integrates key sub-modules such as high-precision sampling circuit, MCU main control unit, DC-DC power conversion circuit, battery management and protection circuit, and communication interface. These modules work together to realize the maximum power point tracking of photovoltaic electric energy, high-efficiency electric energy conversion, intelligent charging and discharging management and safety protection of lithium batteries, and communicate and interact with the upper system, so as to provide the stability, efficiency and autonomy of the whole street lamp system. Operation provides core hardware support.

#### **3.2. High-precision sampling circuit design**

##### **3.2.1. Voltage sampling circuit**

Voltage sampling objects include photovoltaic terminal voltage (12V–24V) and battery terminal voltage (10.8V–13.6V) <sup>[18]</sup>. The design uses a precision resistor voltage dividing network (ratio 10:1) to reduce the voltage. In order to reduce the influence of temperature drift, the precision metal film resistor with an error of  $\pm 0.1\%$  is selected for the voltage dividing resistor. The signal after voltage division is filtered out of high-frequency noise by an RC filter circuit ( $R = 10k\Omega$ ,  $C = 0.1\mu F$ ), and then buffered and impedance matched by a voltage follower composed of an LM324 operational amplifier to ensure that the input signal of the ADC channel of the STM32 single-chip microcomputer is stable and reliable. In order to further improve the sampling accuracy, the reference voltage of the ADC adopts the high-precision reference source REF3030 (error  $\pm 0.1\%$ ), and the final voltage sampling accuracy is better than  $\pm 0.5\%$ .

##### **3.2.2. Analysis and selection of current sampling schemes**

The accuracy and stability of current sampling directly determine the tracking performance of the disturbance-observation MPPT algorithm. In order to achieve low-cost and high-reliability system design, it is necessary to comprehensively compare the mainstream current sampling schemes. This design finally chooses the “sampling resistor high common mode rejection ratio (CMRR) operational amplifier” scheme instead of Hall current sensor, mainly based on the following four dimensions:



- (1) Accuracy and linearity: The perturbation observation method needs to accurately identify small power changes (usually to distinguish  $\geq \pm 0.5\%$  differences) to avoid tracking misjudgments. The sampling resistor scheme has excellent linearity (full range  $\leq \pm 0.1\%$ ), especially in low light and low current ( $< 5\text{A}$ ) scenarios, with high CMRR ( $\geq 80\text{dB}$ ) operational amplifiers, it can effectively suppress common-mode interference, accurately capture power changes, and prevent The algorithm oscillates near the maximum power point (MPP). In contrast, Hall sensors (especially open-loop sensors) have poor linearity ( $\pm 1\% \sim \pm 3\%$ ), and nonlinear and noise problems are more prominent when the current is small, which can easily lead to power calculation deviations and algorithm misjudgments, resulting in additional power generation losses;
- (2) Cost control: Follow the design goal of “low cost and easy implementation” of the project. The cost advantage of the sampling resistor solution is significant: the unit price of precision sampling resistor and high CMRR operational amplifier is low, and there is no need for peripheral isolation or magnetic core, and the circuit is simple. Hall sensors (especially closed-loop sensors) have a high unit price, and may require additional power supply and shielding design, which reduces the complexity and overall cost of the system, and is not conducive to the large-scale promotion of the solution;
- (3) Temperature stability: The operating temperature range of street lamps is wide ( $-20^{\circ}\text{C} \sim 60^{\circ}\text{C}$ )<sup>[19]</sup>. The sampling resistance scheme can control the sampling error in the whole temperature range within  $\pm 0.3\%$  by selecting low temperature coefficient components (such as metal foil resistance  $\leq 5\text{ppm}/^{\circ}\text{C}$ , operational amplifier temperature drift  $\leq 0.1\mu\text{V}/^{\circ}\text{C}$ ) combined with software compensation algorithm, ensuring the stability of the algorithm. The Hall sensor is greatly affected by the temperature, and its magnetic core is prone to significant drift to the temperature at extreme temperatures, which affects the sampling accuracy and may cause the MPP tracking point to shift;
- (4) Bandwidth and dynamic response: For the typical parameters set by the perturbation observation method (disturbance step size  $0.1\text{V}$ , sampling interval  $0.5\text{s}$ ), the required sampling bandwidth is about  $1\text{kHz} \sim 10\text{kHz}$ . The sampling resistor scheme cooperates with operational amplifier (bandwidth is usually  $1\text{MHz} \sim 10\text{MHz}$ ) and appropriate RC filtering (cut-off is about  $159\text{Hz}$ ), has a fast response speed (time constant is about  $1\text{ms}$ ), and can track current changes in real time. Although the bandwidth of the Hall sensor may be higher, the high-frequency noise is large. Strengthening the filtering to suppress the noise will introduce signal delay ( $\geq 5\text{ms}$ ), which may affect the timely response of the algorithm to sudden illumination changes.

To sum up, the “sampling resistance high CMRR operational amplifier” solution is more in line with the design requirements of this distributed MPPT controller in terms of accuracy, cost, temperature drift and dynamic response, and is the choice to achieve high-performance and low-cost MPPT tracking.

### 3.2.3. Temperature sampling circuit

The integrated DHT11 temperature and humidity sensor is selected for temperature sampling<sup>[20]</sup>. Its temperature measurement range is  $0^{\circ}\text{C}$  to  $50^{\circ}\text{C}$ , and the accuracy is  $\pm 2^{\circ}\text{C}$ . It communicates with the STM32 single-chip microcomputer through the I2C interface, and the sampling interval is set to 1 second. The collected temperature data serves two core functions: one is to provide temperature compensation parameters for the MPPT algorithm, and correct the change of photovoltaic cell MPP with temperature; The second is to connect to the battery management system (BMS). When the temperature exceeds the safe range (such as lower than

0°C or higher than 50°C), the charging and discharging current will be automatically or cut off to ensure battery safety.

### **3.3. Power conversion and drive circuit design**

Buck DC-DC converter is selected as the core power conversion topology. The selection is based on the actual voltage matching relationship of the system: the operating voltage of the photovoltaic terminal (usually 18V~24V, the lowest is 12V) is higher than the voltage range of the battery terminal (10.8V~13.6V) in most working conditions. Even under extreme crossover conditions (photovoltaic minimum 12V vs. cell 13.6V), the voltages of the two are basically the same, and the Buck topology can still operate in critical or extremely small duty cycle states. Therefore, there is no need to adopt a more complex and costly Buck-Boost Buck-Boost topology. With its advantages of simple structure, few power devices, high conversion efficiency, and low cost, the Buck topology fully meets the design requirements of a single-pole photovoltaic street lamp system.

The power switching device is N-channel enhanced power MOSFET. In low-power, low-voltage street lamp application scenarios, MOSFETs have the advantages of lower switching losses and simpler drive circuits than IGBTs. The drive circuit adopts a discrete device push-pull scheme with low cost and high efficiency. The circuit is composed of S8050 (NPN type) and S8550 (PNP type) transistors, which can provide fast gate charge and discharge capabilities and ensure fast switching of MOSFET. Gate series resistors (10~100Ω) are used to balance switching speed and electromagnetic interference (EMI). The 3.3V or 5V PWM signal generated by the MCU can directly drive this push-pull circuit.

## **4. Comparative evaluation of economic benefits between distributed and centralized solutions**

### **4.1. Core definition and architecture differences**

The centralized solution adopts the architecture of “central control platform unified transmission network”. All street lamp terminals (including sensors, LED light sources and controllers) are wired (such as optical fiber/cable) or wirelessly connected to a single central platform, which centrally completes data processing and command issuance. Its hardware is highly dependent on central servers and large-scale cabling systems (such as fiber optic solutions). The distributed solution adopts a hierarchical architecture of “regional local autonomous control”. The system deploys multiple areas according to geographical areas, and the street lamp terminals in each area are first connected to the local area to realize data preprocessing and real-time control (such as dimming and fault warning); The area then interacts with the cloud platform through wireless such as 5G/LoRaWAN. Its hardware is centered on modular area and wireless communication devices (such as layers in intelligent networking solutions) <sup>[21]</sup>.

### **4.2. Analysis of key economic factors**

#### **4.2.1. Component prices**

The cost of centralized components accounts for about 40–50% of the total initial investment. If the component price drops by 10%, it can only drive the total investment by 4–5%, because core costs such as central servers and large-scale wiring account for a high proportion and are not affected by component prices. The cost of distributed components accounts for 60–70% of the total initial investment (area and terminal sensors account for a relatively high proportion). A 10% drop in component prices can result in a total investment of 6–7%.

Based on a project of 100,000 street lamps, the upfront cost can be saved by 150,000 to 200,000 yuan. The decrease in module prices is more significant for improving the economy of distributed solutions, especially in photovoltaic integration projects (the decrease in the price of photovoltaic panels can directly affect the cost of distributed energy storage systems) <sup>[22]</sup>.

#### **4.2.2. Controller costs**

Centrally rely on a central controller with high computing power (the cost is about 50,000–100,000 yuan per unit). Every 10% reduction in its cost can make the total initial investment 3–5% (the proportion of central equipment is high). Distributed relies on a large number of low-cost local controllers (the cost is about 50–100 yuan per unit). For every 10% reduction in its cost, the impact on the total initial investment is only 1–2% (the impact is small after the cost is allocated by the region). The cost reduction of the controller is more critical to improve the economy of the centralized scheme. However, the initial cost base of centralized controllers is large. Even after the cost decreases, its total control cost (about 80,000 yuan for 100,000 projects) is usually still higher than that of distributed solutions (about 60,000 yuan).

#### **4.2.3. Regional sunshine and shadow conditions**

The economic difference between the two schemes is small in the area with sufficient sunshine. Centralized solutions can optimize charging through unified scheduling, but distributed solutions usually have 5–8% higher charging efficiency by virtue of local fast adaptation capabilities (charging strategies within 100ms after sudden sunshine changes). In areas with dense shadows, the unified charging strategy of the centralized solution cannot adapt to local shadows, which may lead to insufficient charging in shadowed areas and the need to rely on the grid for power replenishment and electricity expenses. Distributed areas can identify shadows in real time, and dynamic single-light charging power (such as extending charging time for shadow areas, giving priority to power storage in non-shadow areas), resulting in dependence on the power grid, making the cost of energy storage systems 15–20%. In densely shaded areas, the economic advantages of distributed solutions are more prominent. In areas with sufficient sunshine, although the economic difference between the two has narrowed, the distributed solution still has the comprehensive advantage of lower operation and maintenance costs.

### **4.3. Non-economic gains**

Distributed solutions show significant systemic advantages at the non-economic level. In terms of reliability, its “regional autonomy” architecture strictly limits the scope of failure to local areas (usually 50-100 lights), and combined with the multi-link redundant backup mechanism, the overall availability of the system is greatly improved from about 95% of the centralized solution to more than 99.9%, effectively avoiding the inherent risk of “single point of failure, paralysis of the entire network” of the centralized architecture <sup>[23]</sup>. In terms of flexibility and scalability, the modular design of the distributed solution supports “plug and play”, which makes the expansion cost of new functions or devices about 70%, shortens the deployment cycle from several months to several weeks, and can respond agilely Diversified and dynamically changing urban needs. Most importantly, in terms of asset health management, the distributed architecture relies on the real-time monitoring and precise control of the status of the single-lamp energy storage battery realized by the region, which can reduce the average annual decay rate of the battery from about 15% under the centralized unified strategy to less than 8%, thereby extending the expected service life of the battery from 3–4 years to 5–6 years <sup>[24]</sup>. This not only greatly increases the cost of battery replacement and operation and maintenance, but also reflects its deep advantages in

improving system availability and full life cycle value <sup>[25]</sup>.

## 5. Conclusion

The distributed MPPT hardware controller realizes the independent high-precision maximum power point tracking (MPPT) of each photovoltaic panel by adopting the sampling architecture of “precision voltage division filter conditioning high-precision benchmark”, the improved adaptive tracking algorithm based on the disturbance observation method, and the wide voltage dynamic management strategy for lithium iron phosphate batteries. Its sampling comprehensive error is controlled within  $\pm 0.5\%$ , and the tracking efficiency exceeds 95%, thus effectively solving the problems of uneven illumination and shadow occlusion caused by high-rise buildings and trees in the city. The power generation efficiency of a single photovoltaic panel can be increased by 15–20%. Although the initial hardware cost of this solution is slightly increased, with the significant power generation gain and system redundancy reliability, it can cover the cost and create additional energy saving benefits throughout the product life cycle. In the future, the technology will evolve in the direction of high hardware integration and low power consumption, algorithm adaptive prediction, and system IoT node, through the introduction of dedicated chips, intelligent power modules, predictive models and reinforcement learning, and deep integration with smart city platforms. Finally, it will be upgraded from a single lighting controller to a comprehensive intelligent terminal integrating data acquisition, intelligent regulation and urban services.

## Disclosure statement

The authors declare no conflict of interest.

## References

- [1] Krishna V, et al., 2026, Editorial Article (Special Issue): Measurement, Control and Security of Systems for Smart Cities. *Measurement: Sensors*, 2026(44): 101990.
- [2] Değermenci M, Yalman Y, Olcay K, 2026, MPPT Algorithms for Grid-Connected Solar Systems including Deep Learning Approaches. *Scientific Reports*.
- [3] Lal M, et al., 2026, Photogalvanics for Solar Energy Conversion and Storage: A Review on Progress of PG Cells. *Next Research*, 2026(5): 101331.
- [4] Refaat O, et al., 2026, Potentials of Upcycling Photovoltaic Panels Waste in Construction: A Comparative Review. *Developments in the Built Environment*, 2026(25): 100845.
- [5] Ni H, et al., 2025, Enhancing Robustness in Photoacoustic Detection of Dissolved Acetylene in Transformer Oil: Temperature Effects on Resonance Frequency and Suppression Using the Perturbation Observation Method. *Energies*, 18(24): 6512.
- [6] Hajar A, Ahmed G, Benachir E, Innovative Neural Network and Fuzzy Logic Control Techniques for Single-Phase Grid-Connected Photovoltaic Systems using Dual-Core DSP Microcontroller in Smart Home Applications. *Measurement and Control*, 59(2): 232–244.
- [7] Andrasto T, et al., 2024, Incremental Conductance Method in Maximum Power Point Tracking. *IOP Conference Series: Earth and Environmental Science*, 1381(1): 012023.
- [8] Maghraby A, Ahmed H, Abougindia I, 2025, An Ultra-Low Noise Dynamic Comparator for Low-Power, High-



Resolution SAR ADCs. *Circuits, Systems, and Signal Processing*, 2025(prepublish): 1–15.

- [9] Leite D, et al., 2026, Evolving Granular Fuzzy Control: Overview, Case Study on the Chaotic Hénon Map, and Research Outlook. *Applied Soft Computing*, 2026(190): 114639.
- [10] Li H, et al., 2026, Impedance Force Control for Industrial Robot based on Unified Residual Compensation Iterative Learning Control and Multi-Directional Particle Swarm Optimization, *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 240(4): 1214–1228.
- [11] Abdelrahman N, Uth S, 2026, Data-Efficient Prediction in Tableting using Word Embeddings and Empirically-Guided Neural Networks. *International Journal of Pharmaceutics*: X, 2026(11): 100458.
- [12] David J, Volling T, 2026, Multi-Objective Ergonomic–Economic Project Scheduling in Construction: The Case of Photovoltaic System Installation. *Computers & Industrial Engineering*, 2026(213): 111828.
- [13] Dita R, et al., 2025, Development of a Buck Converter for Efficient Energy Storage Integration Using Constant Voltage (CV) Methods. *Energy Engineering*, 122(6): 2355–2370.
- [14] Oleschuk V, 2025, Synchronization and Symmetrization of Base Voltages of Electronic Transformer-Based Converters in the Overmodulation Zone of Voltage Source Inverters. *Surface Engineering and Applied Electrochemistry*, 61(6): 951–959.
- [15] Ling H, et al., 2026, Realizing Rapid Energy Storage and Efficient Release in a Tesla Valve Integrated Cold Energy Storage Unit for Data Center Cooling. *Applied Thermal Engineering*, 289(P1): 129683.
- [16] Subashini M, Sumathi V, 2026, Adaptive Charging Strategies for Electric Vehicles: A Reinforcement Learning Approach to Demand Response and Resource Management. *Energy Reports*, 2026(15): 109015.
- [17] Ghana Achieves Stable Power Supply, Eyes Green Future after Major Energy Reforms, 2025, M2 Presswire.
- [18] Li J, et al., 2026, Novel Control Strategy for Voltage Control System Integrated with Photovoltaic System using Terminal Reaching Law-based Fast Integral Terminal Sliding Mode Concept. *Electric Power Systems Research*, 2026(253): 112548.
- [19] Kong X, et al., 2026, Local Structural Engineering for Simultaneous Enhancement of Piezoelectricity and Temperature Stability in BiScO<sub>3</sub>-PbTiO<sub>3</sub> Piezoceramics. *Journal of Materials Science & Technology*, 2026(268): 51–60.
- [20] Lu J, et al., 2026, Multi-Physics Coupled Simulation and Hybrid Mechanistic Data Driven Modeling for High-Accuracy Radiation Error Correction in Sounding Temperature Sensors. *Sensor Review*, 46(2): 299–310.
- [21] Bala I, Singh G, Tripathi S, *Sustainable Technologies and Devices for Next-Generation Wireless Communication*: CRC Press.
- [22] Li L, Dai C, 2024, Internal and External Factors Influencing Rural Households' Investment Intentions in Building Photovoltaic Integration Projects. *Energies*, 17(5).
- [23] De Deken J, Luigjes C, 2025, Federal Solidarity, Regional Autonomy and Institutional Moral Hazard. The Case of Belgian Unemployment Insurance and Activation Policies. *Journal of European Social Policy*, 35(5): 423–438.
- [24] Alfahdi K, Gultekin H, Summad E, 2025, A Novel Global Health Index Framework for Asset Prognostics and Health Management in the Oil and Gas Industry. *Scientific Reports*.
- [25] Li H, et al., 2013, Design on Distributed Reconfigurable Satellite Full Life Cycle Value Assessment System. *Applied Mechanics and Materials*, 482(482-482): 287–291.

**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

