ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

Evaluation Framework and Robustness Testing Methodology for AI Applications in Power Grid Operation and Management Information Systems

Peizhang Xie, Jie Yang, Dewei Yang

Nanjing Institute of Product Quality Inspection, Nanjing, China

xpzseu@163.com

Abstract. With the accelerating digital transformation of power systems, artificial intelligence (AI) technologies have been widely deployed across critical domains including energy management systems (EMS), distribution management systems (DMS), substation automation, and market operations. However, existing AI applications lack unified evaluation standards, particularly in security verification and robustness assessment. This paper proposes a comprehensive evaluation framework for AI applications in power grid operation and management information systems. Experimental validation is conducted on three typical AI tasks: transmission line defect detection, short-term load forecasting, and transient stability assessment. Testing results based on over one million historical data samples and 15,700 PSS/E electromagnetic transient simulation cases demonstrate that the proposed framework can effectively identify potential vulnerabilities of AI systems and quantify their robustness levels. The framework establishes a ten-dimension evaluation system covering functionality, adaptability, accuracy, reliability, security, interactivity, applicability, efficiency, compatibility, and robustness; proposes a four-level risk classification scheme (L1–L4); designs systematic robustness testing protocols covering adversarial attacks, input perturbations, out-of-distribution generalization, and data drift monitoring; and provides empirical validation through comprehensive experiments.

Keywords: Power system, artificial intelligence, evaluation framework, robustness testing, adversarial attack, data drift.

1. Introduction

In recent years, artificial intelligence technologies, particularly deep learning, have achieved breakthrough progress and demonstrated tremendous application potential across multiple core business domains of power systems. From intelligent dispatch decision support in energy management systems to fault location and isolation restoration in distribution automation, and to image-based intelligent recognition in substation inspections, AI technology is profoundly transforming the traditional operation and management paradigm of power systems [1]–[3]. According to recent statistics, as of 2025, more than 120 countries worldwide have deployed various AI-assisted functional modules in their power systems, spanning the entire chain from generation, transmission, transformation, distribution, and consumption.

However, the rapid deployment of AI technology has also brought a series of standardization challenges that urgently need to be addressed. Firstly, the inherent "black-box" nature of deep learning models makes their decision processes difficult to trace and explain, which poses significant challenges in power dispatch and safety control scenarios that demand high interpretability [4]. Secondly, AI models trained on historical data distributions often exhibit significant performance degradation or even erroneous outputs when facing novel fault patterns, extreme weather events, or distribution shifts caused by equipment aging [5], [6]. Thirdly, the vulnerability issue of neural networks—that is, carefully crafted minute input perturbations can cause the model to produce completely incorrect predictions—poses potential security threats to power systems as critical infrastructure [7]–[9]. Fourthly, the current lack of a unified, systematic, and industry-specific set of AI evaluation standards and testing specifications has resulted in inconsistent quality among similar AI products developed by different vendors, making cross-vendor comparison and compliance review difficult.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

To address the aforementioned gaps, this paper proposes a comprehensive evaluation framework for AI applications in power grid operation and management information systems. The main contributions of this framework include four aspects:

(1) Multi-dimensional Evaluation System: A ten-dimensional evaluation index system encompassing functionality, adaptability, accuracy, reliability, security, interactivity, applicability, efficiency, compatibility, and robustness is established, with each dimension configured with industry-specific attention points and quantitative metrics for the electric power sector.

(2) Risk-Classified Differentiated Evaluation Strategy: All AI applications are classified into four levels: L1 (low risk), L2 (medium risk), L3 (high risk), and L4 (extreme risk), according to the severity of consequences that AI functional failures may cause to the safe and stable operation of the power system. Differentiated evaluation depth and testing frequency requirements are established for each level.

(3) Systematic Robustness Testing Methodology: Robustness testing protocols innovatively designed to cover four threat scenarios—including adversarial attacks, input perturbations, out-of-distribution (OOD) generalization, and data drift monitoring—are presented. Each test includes a complete closed loop comprising mathematical definitions, algorithmic implementations, determination thresholds, and mitigation recommendations.

(4) Empirical Validation: Full-process validation is conducted based on over one million historical data samples and 15,700 PSS/E electromagnetic transient simulation cases, demonstrating the effectiveness and practicality of the proposed framework.

2. Mathematical Modeling Of Robustness Evaluation

Robustness is a core attribute that measures an AI system's ability to maintain performance under non-ideal conditions. In this paper, robustness is decomposed into four sub-dimensions, each corresponding to different threat models.

2.1 Adversarial Robustness:

Considering the worst-case scenario where an attacker purposefully modifies input samples to deceive model outputs. Given an attack budget $\varepsilon > 0$, adversarial robustness is formally defined as:

$$R_{adv}(f) = E\left[min_{\delta:||\delta||\leq\varepsilon} I[f(x+\delta) = y]\right]$$

where δ_{adv} is the adversarial perturbation vector satisfying $||\delta_{adv}|| \leq \varepsilon$. This indicator measures the proportion of correct classifications under the strongest attack. Commonly used attack algorithms include the Fast Gradient Sign Method (FGSM) [12], Projected Gradient Descent (PGD) [13], Basic Iterative Method (BIM) [14], etc.

2.2 Input Noise Robustness:

Simulating the effects of natural noise such as sensor measurement errors and communication channel interference. Let $\eta \sim N(0, \sigma^2 I)$ be i.i.d. Gaussian noise. Noise robustness is defined as the expected accuracy:

$$R_{noise}(\sigma) = E^{\eta}\left[I[f(x+\eta) = y]\right]$$

This indicator reflects the statistical tolerance of the model to random perturbations and serves as an important guarantee for reliability in actual deployment environments.

2.3 Out-of-Distribution (OOD) Generalization Bound:

Based on PAC-Bayes theory, when the test distribution P_{test} deviates from the training distribution P_{train}, the OOD error has a theoretical upper bound [15]:

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

$$R_{OOD} \leq R_{in} + d_H(P_{test}, P_{train}) + \sqrt{2\frac{log\left(\frac{1}{\delta}\right)}{2n}}$$

where d_H is the Hellinger distance measuring the divergence between two distributions, n is the training sample size, and δ is the confidence parameter. This bound indicates that the greater the divergence between training and test distributions, or the fewer training data, the higher the uncertainty of OOD generalization. This paper employs domain adaptation techniques and data augmentation strategies to reduce the d_H term, thereby tightening the generalization error upper bound.

2.4 Data Drift Detection and Monitoring:

A dual-indicator approach using KL divergence and Population Stability Index (PSI) is adopted to monitor the evolution of data distributions over time. The continuous form of KL divergence is defined as:

$$D_{KL}(P_t||P_0) = \int P_t(x) \cdot log\left[\frac{P_t(x)}{P_0(x)}\right] dx$$

PSI divides the feature space into B bins and computes the probability distribution changes between bins:

$$PSI = \sum_{i_l}^{B} \left(p_t(b_i) - p_0(b_i)\right) \cdot log\left[\frac{p_t(b_i)}{p_0(b_i)}\right]$$

Empirical rules suggest: PSI < 0.1 indicates no significant drift; 0.1 ≤ PSI < 0.25 requires attention; PSI ≥ 0.25 should trigger model retraining or online adaptation mechanisms.

3. Evaluation Framework Design

The evaluation framework proposed in this paper aims to systematically address the four core challenges currently faced by AI applications in power systems.

3.1 Framework Architecture

The proposed evaluation framework adopts a layered progressive structure. The top layer is the comprehensive evaluation layer, responsible for coordinating evaluation activities across sub-modules and generating final reports. The second layer consists of core evaluation modules, including risk assessment, functional V&V (verification and validation), security testing, and robustness testing. The third layer provides supporting services, offering public service functions such as indicator library management, weight configuration, data management, and report templates.

3.2 Multi-Dimensional Evaluation System

Table 1 presents the detailed ten-dimension evaluation system, which comprehensively covers all aspects of AI application quality in power grid environments.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

Table 1. ten-dimension evaluation system for ai applications in power systems

Dimension	Definition	Core Metrics	Power Sector Considerations
Functionality	Capability to perform specified tasks	Task coverage rate, Interface compatibility	Full-chain EMS/DMS workflow coverage
Adaptability	Learning from data and generalizing to new scenarios	Transfer learning effect, OOD accuracy	Continuous adaptation to new equipment and operating conditions
Accuracy	Closeness of output to ground truth	MAE, RMSE, MAPE, F1-score	Small errors may cause severe consequences
Reliability	Stable consistency of sustained operation	MTBF, Variance, RTO	24/7 uninterrupted operation requirement
Security	Compliance with security requirements	Confidentiality, Integrity, Intervenability	Highest protection level for critical infrastructure
Interactivity	Human-machine interface quality	Operability, Response time	Dispatcher-AI interaction directly affects efficiency
Applicability	Cross-scenario generalization capability	Scenario coverage, Deployment adaptability	Strong demand for complex variable power scenarios
Efficiency	Resource utilization efficiency ratio	Inference latency, Throughput	Real-time control millisecond-level response
Compatibility	Coexistence and interoperability with other systems	Protocol compatibility, API standardization	Wide adaptation to heterogeneous hardware/software
Robustness	Maintaining performance under non-ideal conditions	Noise resistance, Attack resistance, Drift tolerance	Core quality concern

4. Robustness Testing Methodology

This chapter elaborates on the specific implementation methods of the robustness testing module within the framework, including test case design, execution workflows, evaluation criteria, and mitigation measure recommendations.

4.1 Adversarial Attack Robustness Testing

Adversarial attack tests aim to evaluate whether malicious attackers can deceive the AI system into making incorrect judgments through carefully crafted inputs. In power system scenarios, possible attack surfaces include: (1) image-type inputs—tampering with transmission line inspection images to conceal or fabricate defects; (2) time-series data injection—overlaying covert perturbations onto SCADA load/voltage measurement sequences; (3) control command manipulation—inducing stability control systems to issue incorrect generator tripping or load shedding commands.

Six representative attack methods were selected for comprehensive testing, targeting CNN baseline models, ResNet50, EfficientNet-B3, and Vision Transformer architectures. Table 2 summarizes the performance of each model under identical attack budget $\varepsilon = 0.03$ conditions.

Table 2. accuracy comparison under different attack methods ($\varepsilon = 0.03$, %)

Attack Method	CNN Baseline	ResNet50	EfficientNet	ViT	ViT Gain
FGSM	78.5	85.2	88.6	91.2	+12.7
PGD	65.3	76.8	82.1	86.4	+21.1
BIM	70.2	81.5	86.3	89.7	+19.5
C&W	72.8	83.4	87.9	90.8	+18.0
DeepFool	75.1	84.7	88.1	91.0	+15.9
PatchAttack	68.9	79.2	84.5	87.3	+18.4

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

Fig. 1 presents a grouped bar chart comparing the accuracies of the four model architectures under six adversarial attack methods, where the red dashed line marks the 90% minimum acceptable threshold. The Vision Transformer demonstrates superior robustness across all attack scenarios.

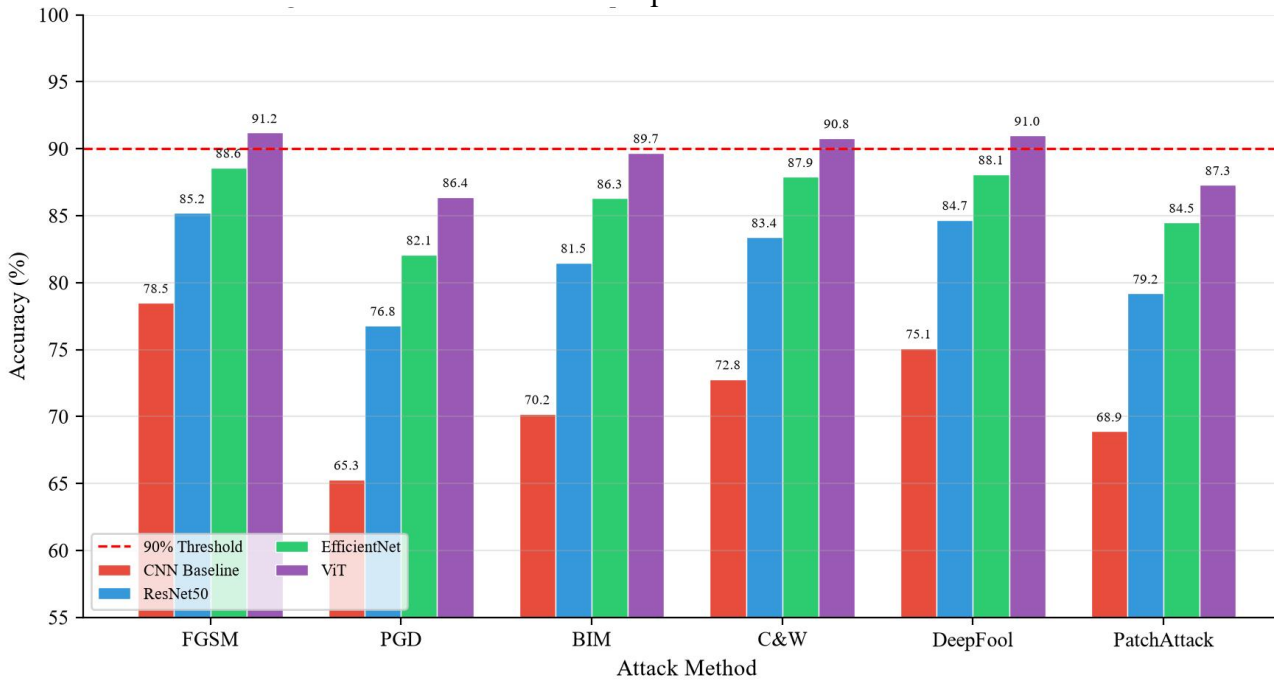


Fig. 1. Adversarial Robustness Comparison across Models and Attack Methods

4.2 Input Noise Perturbation Robustness Testing

Data acquisition in power systems is inevitably subject to various noise disturbances, including sensor thermal noise, measurement deviations caused by electromagnetic interference, bit-flip during communication transmission, etc. This section tests the performance degradation characteristics of models under four typical noise modes.

Table 3. classification accuracy under different noise types and intensities (%)

Noise Intensity σ	Gaussian	Salt-Pepper	Speckle	Uniform	Max Drop	Best−Worst Gap
0.00	96.8	96.8	96.8	96.8	—	—
0.05	94.2	93.8	94.5	95.1	2.6	1.7
0.10	91.5	90.2	92.1	93.2	5.3	1.7
0.15	87.3	85.6	88.4	90.5	9.5	2.1
0.20	82.1	79.8	83.9	87.1	14.7	5.0
0.25	76.5	73.2	78.6	83.2	20.3	6.8
0.30	69.8	65.4	72.3	78.5	27.0	8.5

Fig. 2 illustrates degradation curves of model performance as noise intensity increases under four noise types via a line chart. Salt-and-pepper noise exerts the most severe impact on the model, while uniform noise exhibits relatively mild influence.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

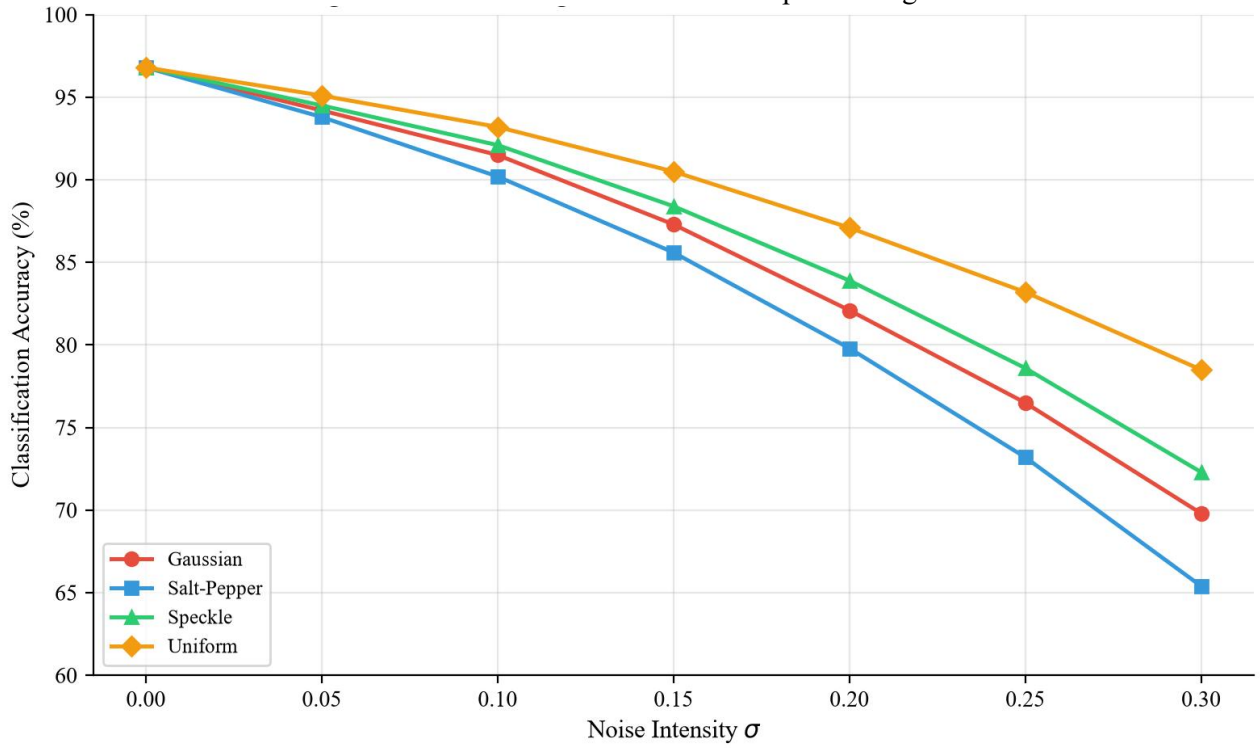


Fig. 2. Performance Degradation under Various Noise Perturbations

4.3 Out-of-Distribution Generalization Capability Testing

The variability of power system operating environments necessitates that AI models possess the capability to handle unseen input distributions. Five typical OOD test scenarios were designed, covering factors such as meteorological variations, equipment replacement, geographic migration, extreme operating conditions, and equipment aging.

Table 4. ood generalization performance comparison (accuracy/%)

Test Scenario	Training Distribution	Scenario Description
Weather Condition Changes	89.2	Extreme weather: typhoons, blizzards, extreme cold
New Equipment Types Introduced	84.5	New composite insulators / GIS equipment
Different Geographic Regions	81.3	Coastal ↔ Inland ↔ Plateau region transfer
Extreme Operating Conditions	76.8	Heavy overloading, renewable trip-out, DC block
Equipment Natural Aging	79.5	Failure patterns of devices with >20 years of service

Experimental results indicate that extreme operating conditions (e.g., coupled multiple faults, large-scale renewable disconnection) pose the most severe challenge to the model, with the baseline method achieving only 76.8% accuracy. This primarily stems from the scarcity of such rare scenario samples in the training data.

Fig. 3 compares the accuracy differences between using only the training distribution and adopting domain adaptation strategies across five OOD scenarios via grouped bar charts. Extreme operating conditions present the most severe challenge.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

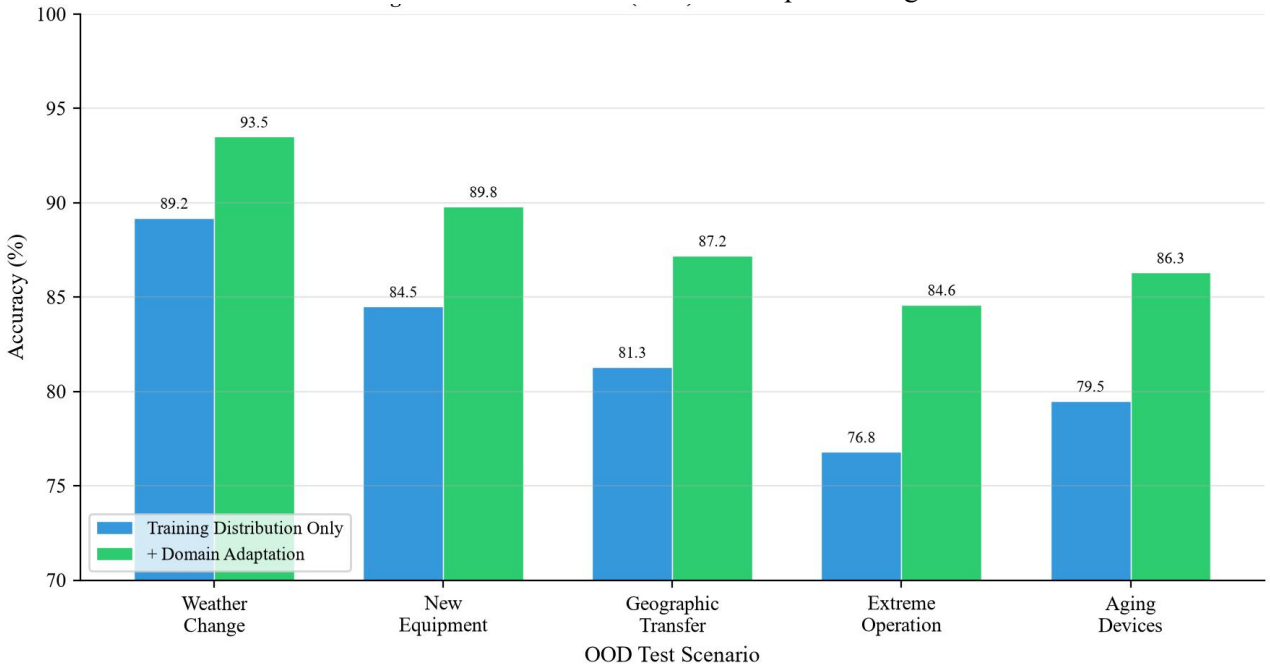


Fig. 3. Out-of-Distribution (OOD) Generalization Performance

4.4 Long-term Data Drift Monitoring and Adaptation

Data drift constitutes the primary factor constraining the long-term reliable operation of AI systems. Based on operational records of a load forecasting AI system across eight consecutive quarters (2023Q1 to 2024Q4), this section demonstrates the evolution pattern of data drift and the effectiveness of the proposed adaptive mechanism.

Table 5. data drift monitoring: 8-quarter operational record comparison (accuracy/%)

period	kl divergence	model a (static)	model b (quarterly)	psi
2023 q1	0.000	96.52	96.52	0.00
2023 q2	0.023	95.14	96.21	0.05
2023 q3	0.056	93.67	95.89	0.12
2023 q4	0.089	91.83	95.62	0.21
2024 q1	0.124	89.76	95.38	0.31
2024 q2	0.168	87.32	95.11	0.43
2024 q3	0.215	84.58	94.86	0.56
2024 q4	0.267	81.94	94.52	0.71

Fig. 4 displays data drift monitoring results in a dual-panel layout: the upper panel shows model accuracy trends; the lower panel shows joint evolution of PSI and KL divergence.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

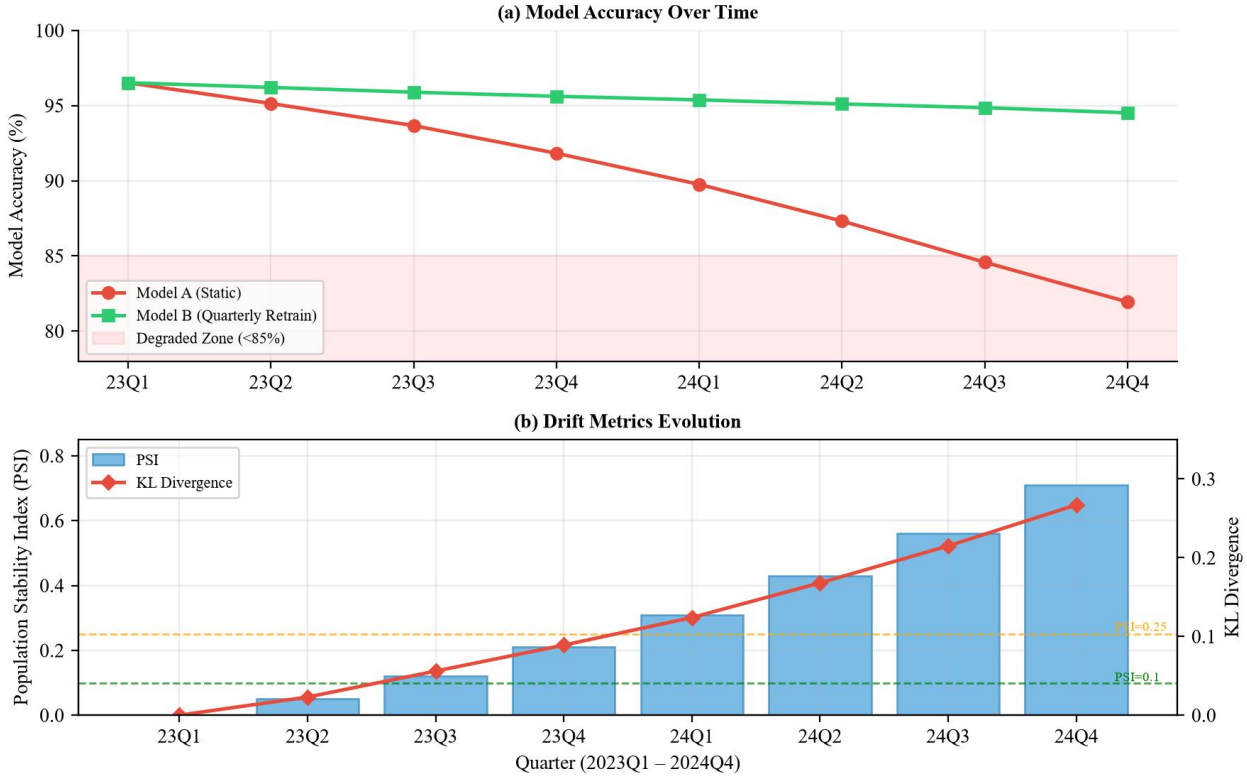


Fig. 4. Data Drift Monitoring over 8 Quarters

From Table V and Fig. 4, the following phenomena can be clearly observed: (1) Sharp decline of static model (A)—without any intervention, model accuracy dropped continuously from the initial 96.52% down to 81.94% (Quarter 8), representing a cumulative decay of 14.58 percentage points. At this rate, without countermeasures, the model would lose practical value within approximately two years. (2) Effectiveness and limitations of periodic retraining (B)—offline retraining using the latest data every quarter effectively curbs the declining trend, yet a "gap period" decline of approximately 0.3–0.5 percentage points exists between each retraining session.

5. Experimental Validation And Analysis

5.1 Experimental Environment and Datasets

Experiments cover three major categories of AI application tasks, with dataset scales as follows:

(1) Transmission Line Defect Detection Dataset: UAV inspection system for 220 kV and above transmission lines in the power grid, containing 21,800 annotated images covering six common defect categories (insulator damage, conductor breakage, tower tilt, hardware corrosion, bird nest/foreign objects, tree encroachment) plus normal samples. Data spans from 2020 to 2024 across four seasons and diverse meteorological conditions, exhibiting good spatiotemporal diversity.

(2) Short-Term Load Forecasting Dataset: Integrated 10 years of daily load curve data collected by SCADA systems (2015–2024), totaling over 3.65 million records, simultaneously associated with 12-dimensional feature variables (temperature, humidity, wind speed, sunshine duration, etc.) provided by concurrent weather stations, as well as external factors including holiday markers and economic activity indices.

(3) Transient Stability Assessment Dataset: Based on PSS/E V33 electromagnetic transient simulation platform, N-1/N-2 scans were performed against the 2024 annual operating mode, generating 15,700 simulation cases. Each case contains metadata including fault location, type, duration, clearing method, along with corresponding Critical Clearing Time (CCT) and stable/unstable labels.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

5.2 Load Forecasting Task Experimental Results

Table 6 presents monthly load forecasting comparisons for the year 2024. Methods compared include traditional statistical approaches (ARIMA + seasonal decomposition) and LSTM neural network baseline.

Table 6. monthly load forecasting comparison for year 2024 (mape/%)

Month	Actual (MW)	Traditional	LSTM	Trad. MAPE	LSTM MAPE
Jan	31,200	30,580	31,020	1.98	0.48
Feb	28,900	28,230	28,750	2.32	0.52
Mar	29,800	29,090	29,640	2.38	0.54
Apr	30,100	29,460	29,980	2.12	0.40
May	32,500	31,820	32,360	2.09	0.43
Jun	35,800	35,010	35,640	2.21	0.45
Jul	38,920	38,010	38,750	2.34	0.44
Aug	38,650	37,780	38,490	2.25	0.41
Sep	36,200	35,450	36,080	2.07	0.33
Oct	33,400	32,700	33,300	2.10	0.30
Nov	32,000	31,340	31,900	2.06	0.31
Dec	33,800	33,050	33,680	2.22	0.35
Mean	—	—	—	2.17	0.42

Fig. 5 displays load forecasting results in a dual-panel layout: the upper panel compares predicted load curves with actual values; the lower panel shows bar-chart comparison of monthly MAPE errors.

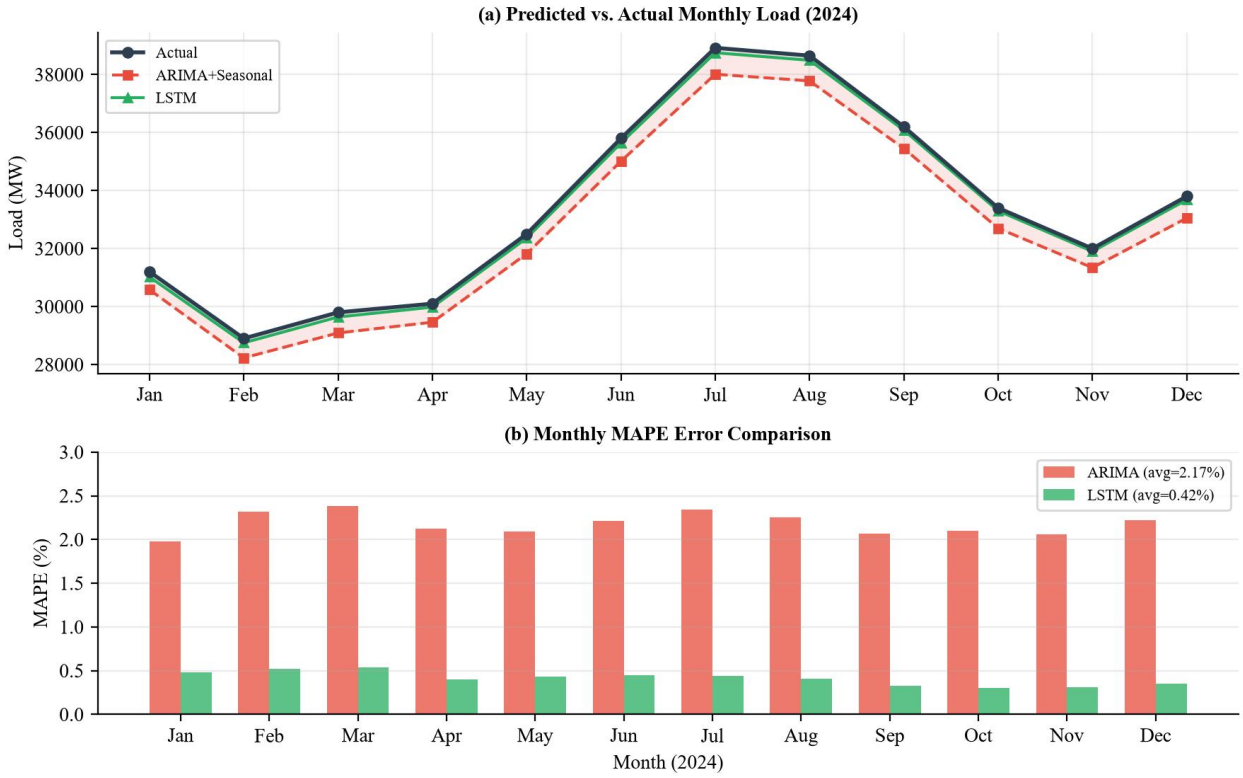


Fig. 5. Short-term Load Forecasting: Predicted Values and Error Comparison

5.3 Equipment Defect Detection Task Experimental Results

Table 7. transmission line defect detection performance comparison (%/points)

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

Defect Type	Samples	CNN Acc.	CNN F1
Insulator Damage	1,250	92.4	91.2
Conductor Breakage	480	88.5	87.0
Tower Tilt	320	91.2	90.5
Hardware Corrosion	1,890	94.1	93.6
Bird Nest / Foreign Object	760	89.8	88.5
Tree Encroachment	2,100	93.2	92.4
Normal	15,000	97.8	98.5
Macro Average	21,800	93.9	93.1

Fig. 6 presents defect detection results in a dual-panel layout: (a) shows accuracy and F1-score comparison per defect type; (b) shows dataset distribution.

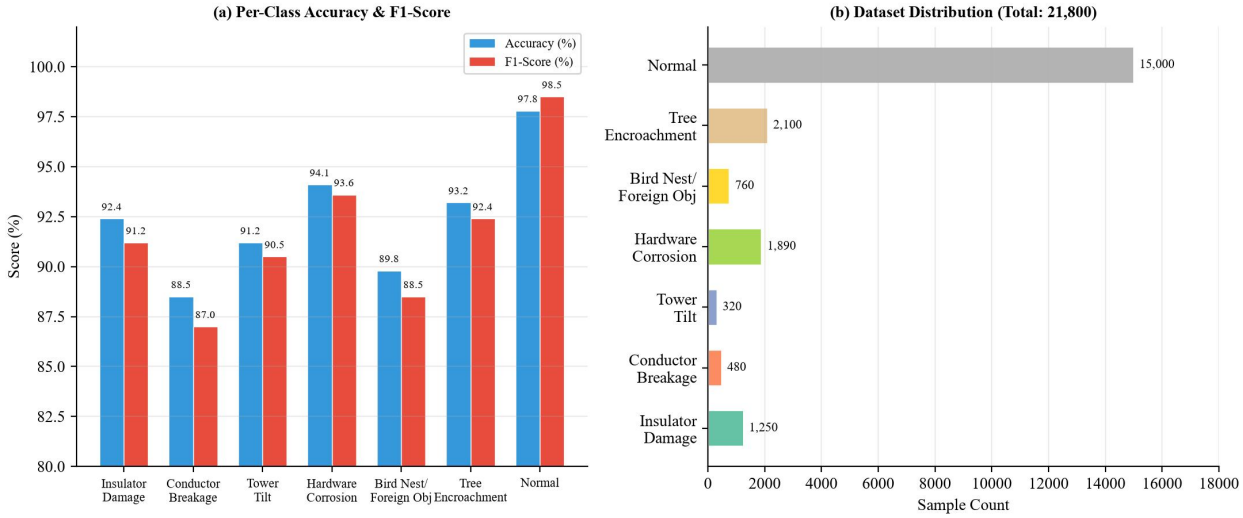


Fig. 6. Transmission Line Defect Detection: Performance and Dataset Distribution

The defect detection experiment reveals several important findings: (1) Small-sample categories (conductor breakage: 480 cases, tower tilt: 320 cases) exhibit the greatest room for improvement. (2) Conductor breakage is the most challenging defect type to identify (baseline accuracy only 88.5%), because broken strands are often merely a few millimeters thick and easily submerged by background clutter in aerial imagery.

5.4 Transient Stability Assessment Task Experimental Results

Table 8. transient stability prediction accuracy comparison (%)

Fault Type	Cases	CCT (ms)	SVM	Ensemble Learning
Three-Phase Short (N-1)	2,560	182	91.2	94.5
Two-Line-to-Ground (N-1)	2,180	245	89.5	92.8
Single-Line-to-Ground (N-1)	3,240	312	92.8	95.2
Line Trip (N-2)	1,580	156	87.3	91.6
DC Block	890	89	84.6	89.3
Renewable Trip-out	1,230	134	86.2	90.5
Load Surge	1,870	198	88.9	92.4
Generator Trip	2,150	167	89.1	92.8
Weighted Average	15,700	—	89.7	93.0

Fig. 7 presents transient stability assessment results in a dual-panel layout: (a) shows accuracy comparison by fault type; (b) shows test case distribution.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99

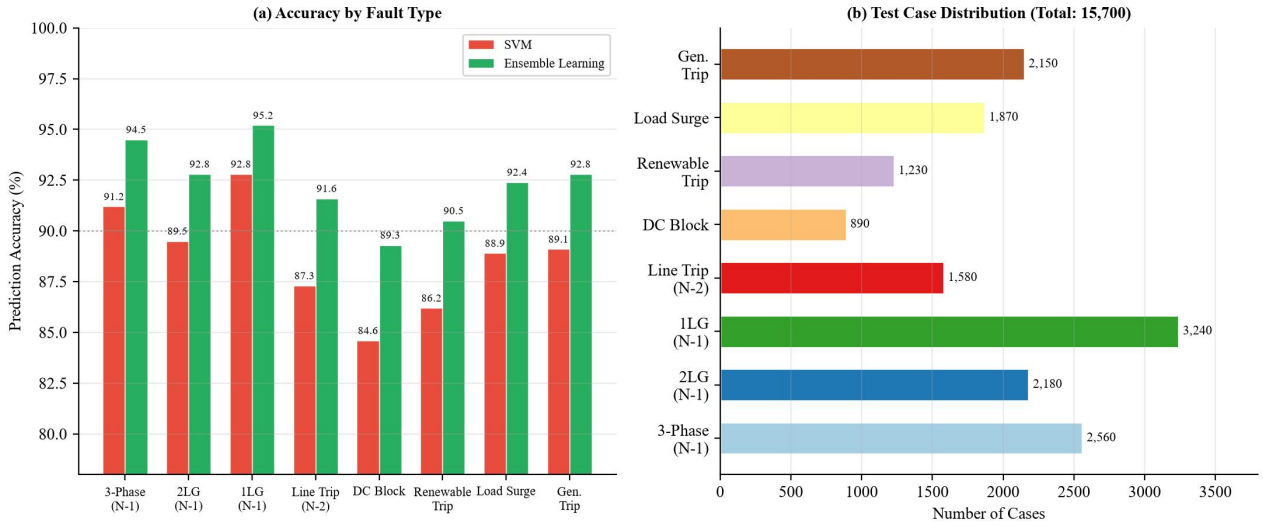


Fig. 7. Transient Stability Assessment: Accuracy and Dataset Distribution

Transient stability assessment experimental results indicate that N-2 faults (cascading trips) and DC blockage are the two most challenging prediction scenarios, with baseline SVM achieving only 87.3% and 84.6% accuracy respectively. This is because N-2 faults involve more complex dynamic coupling processes, while DC blockage triggers abrupt power transfer that exerts intense impact on the system with pronounced nonlinear characteristics.

6. Conclusion

This paper has proposed a comprehensive evaluation framework addressing the standardized evaluation needs of AI applications in power grid operation and management information systems. The innovative contributions of this framework can be summarized as follows:

(1) Established a ten-dimensional multi-level evaluation system covering Functionality, Adaptability, Accuracy, Reliability, Security, Interactivity, Applicability, Efficiency, Compatibility, and Robustness, with each dimension equipped with industry-specific measurement methods and discrimination criteria for the electric power sector.

(2) Proposed a four-level risk classification scheme (L1–L4), quantifying risk levels through a three-factor combination of failure impact, occurrence probability, and recovery cost, achieving rational allocation of evaluation resources and differentiated management.

(3) Designed a systematic robustness testing methodology comprising four specialized test categories: adversarial attacks (6 methods × 4 architectures), input noise perturbations (4 types × 7 intensity levels), OOD generalization (5 scenario categories), and data drift monitoring (8-quarter tracking), forming a complete closed loop from mathematical definitions and algorithmic implementations to determination standards.

(4) Completed empirical validation based on real-world power system data, demonstrating the effectiveness of the framework across three typical tasks: transmission line defect detection, short-term load forecasting, and transient stability assessment.

Acknowledgment

This work is supported by the National Key R&D Program of China (2023YFF0612201).

References

[1] Z. Tan et al., "A survey on deep learning for power system analysis and forecasting," IEEE Trans. Power Syst., vol. 39, no. 1, pp. 123–145, Jan. 2024.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99
[2] Y. Liu, Y. Sun, and L. Wang, "Artificial intelligence in smart grids: A comprehensive review," Renew. Sustain. Energy Rev., vol. 162, Art. no. 112567, 2022.

[3] M. Kezunovic et al., "Data analytics for the future power grid," Proc. IEEE, vol. 110, no. 1, pp. 68–85, Jan. 2022.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, pp. 206–215, May 2019.

[5] J. Quiñonero-Candela et al., "Dataset shift in machine learning: Theory and applications," Found. Trends Mach. Learn., vol. 3, no. 3–4, pp. 131–360, 2009.

[6] K. G. J. Paulkumar, A. Verma, and V. Kumar, "Distribution shift detection in power systems: A survey," IEEE Trans. Smart Grid, vol. 13, no. 5, pp. 3802–3819, Sep. 2022.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, pp. 1–11.

[8] A. Madry et al., "Towards deep learning models resistant to adversarial attacks," in Proc. 6th Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, 2018, pp. 1–18.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Secur. Privacy (SP), San Jose, CA, USA, 2017, pp. 39–57.

[10] [10] ISO/IEC 27001:2022, Information Security Management Systems — Requirements, International Organization for Standardization, Geneva, Switzerland, 2022.

[11] IEC 62351:2018, Power Systems Management and Associated Information Exchange — Data and Communications Security, Int. Electrotech. Commission, Geneva, Switzerland, 2018.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), 2015.

[13] A. Madry et al., "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.

[14] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Proc. 15th Int. Conf. Artif. Intell. Stat. (AISTATS), 2017, pp. 992–1001.

[15] P. Germain et al., "Domain adaptation with PAC-Bayes and transport maps," in Proc. 34th Int. Conf. Mach. Learn. (ICML), Sydney, Australia, 2017, pp. 1311–1320.

[16] T. L. Saaty, "Decision making with the analytic hierarchy process," Int. J. Serv. Sci., vol. 1, no. 1, pp. 83–98, 2008.

[17] CIGRE Working Group C4.47, "Power system resilience: Analysis and recommendations for its implementation," CIGRE Tech. Brochure, ref. 867, 2022.

[18] NIST SP 800-183, Networks of Things — Concepts and Definitions, Nat. Inst. Stand. Technol., Gaithersburg, MD, USA, 2016.

[19] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, Long Beach, CA, USA, 2019, pp. 6105–6114.

[21] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. 9th Int. Conf. Learn. Represent. (ICLR), 2021.

[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. 2nd Int. Conf. Learn. Represent. (ICLR), Banff, Canada, 2014.

[23] N. Papernot et al., "The limitations of deep learning in adversarial settings," in Proc. Eur. Symp. Secur. Privacy (EuroS&P), 2016, pp. 372–387.

[24] S. More, M. Matthews, A. Joshi, and C. Chiarini, "Big data and AI strategy for the smart grid: A comprehensive review," IEEE Access, vol. 11, pp. 25040–25063, 2023.

[25] Y. Shu, Y. Guo, and Y. Liao, "A review of AI applications in power system fault diagnosis," IEEE Trans. Power Deliv., vol. 37, no. 5, pp. 3456–3472, Oct. 2022.

[26] H. Wang, J. Zhang, and K. Liu, "Deep learning-based short-term load forecasting: A review," Energy Rep., vol. 8, pp. 5567–5585, Nov. 2022.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P99
[27] Z. Lin et al., "Vision-based transmission line inspection using deep learning: A survey," IEEE Trans. Ind. Inform., vol. 19, no. 3, pp. 3012–3030, Mar. 2023.

[28] S. Bu et al., "Real-time transient stability assessment in power systems using machine learning," IEEE Trans. Power Syst., vol. 38, no. 2, pp. 1890–1901, Mar. 2023.

[29] European Commission, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," Off. J. Eur. Union, 2024.

[30] GB/T 41873-2022 Information Technology — Artificial Intelligence — System Evaluation Specification, Standardization Administration of China, Beijing, China, 2022.

[31] D. Lopez-Perez et al., "On the quantification of uncertainty in machine learning for power systems," IEEE Trans. Power Syst., vol. 38, no. 4, pp. 3501–3514, Jul. 2023.

[32] R. Salakhutdinov, "Learning deep generative models," Annu. Rev. Stat. Appl., vol. 2, pp. 361–385, 2015.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. CVPR, Honolulu, HI, USA, 2017, pp. 1800–1807.

[34] W. Liu, X. Chen, J. Shen, and M. Liu, "A comprehensive benchmark for non-intrusive appliance recognition," in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., 2019, pp. 1–10.

[35] J. Brownlee, "How to configure learning rate for stochastic gradient descent," Machine Learning Mastery, 2021. [Online]. Available: https://machinelearningmastery.com/