Fusion of Multi-Source Data Based on Improved Bayesian Network for Power Grid Anomaly State Identification

Yuhao Xie
Key Laboratory of Smart Grid of Ministry of Education, Tianjin University
Tianjin, China
yhxie617@163.com

Yuan Zeng
Key Laboratory of Smart Grid of Ministry of Education, Tianjin University
Tianjin, China
zengyuan@tju.edu.cn

Junzhi Ren*, Chao Qin
Key Laboratory of Smart Grid of Ministry of Education, Tianjin University
Tianjin, China
rjz_elec@tju.edu.cn
qinchao@tju.edu.cn

Lingxu Guo
State Grid Tianjin Electric Power Company
Tianjin, China
lingxuguo@126.com

***Abstract*—This paper proposes an integrated framework that employs an Improved Bayesian Network (BN) for physically consistent multi-source data fusion and contrastive representation learning for enhancing representation separability, enabling anomaly state identification in modern power grids. The Improved BN module unifies heterogeneous data sources—including high-frequency PMU, low-frequency SCADA, topological, and meteorological information—through hierarchical Bayesian inference and spatiotemporal feature encoding, establishing a coherent representation of grid dynamics. Principal component analysis and unsupervised clustering are employed to validate the separability of fused features under normal and abnormal conditions. Building upon this foundation, a contrastive learning module extracts deep discriminative representations, explicitly enlarging the separation between normal and anomalous operating states with improved robustness. Experimental validation on the IEEE 39-bus system confirms that the proposed framework effectively enhances fusion interpretability and the reliability of anomaly detection, providing a practical foundation for intelligent grid monitoring and early warning applications.**

***Index Terms* — Anomaly Detection, Contrastive Learning, Improved Bayesian Network, Multi-Source Data Fusion, Power Grid Monitoring**

I. INTRODUCTION

The transformation of modern power systems, driven by large-scale renewable integration and extensive use of power electronic devices, has increased complexity and uncertainty. These conditions make grids more vulnerable to anomalies and disturbances, posing challenges to the effectiveness of traditional monitoring and protection schemes [1].

In recent years, multi-source data fusion has emerged as a critical solution to these challenges. Bayesian inference methods have been applied to distribution system state estimation, enabling probabilistic reasoning under uncertainty [1], while Bayesian networks have been successfully used in transformer condition assessment to integrate diagnostic information [2]. Kalman filtering has been improved to handle heterogeneous data sources such as PMU and SCADA for dynamic state estimation [3]. Probabilistic Graphical Models (PGMs) further advance this by combining operational and environmental data for outage location [4]. Hierarchical fusion frameworks have also been proposed for fault diagnosis by integrating switching information and electrical measurements [5]. However, most existing fusion frameworks in practice focus primarily on data-level integration, which lacks the formal embedding of underlying physical constraints and often fails to maintain consistency across asynchronous heterogeneous sources.

At the same time, data-driven and intelligent methods have expanded the toolkit for anomaly analysis. Computational approaches for smart grid fault detection and localization [6], fuzzy neural systems for incomplete alarm information [7], and CNN-LSTM networks for spatiotemporal meteorological forecasting [8] have been demonstrated. Furthermore, studies on distribution networks have emphasized real-time early warning through multi-source data integration [9]. Wide-Area Measurement Systems (WAMS) have also become essential infrastructures, supporting islanding detection [10], co-optimal placement of devices and communication infrastructures [11], and robust information-centric network architectures [12]. Despite these efforts, pure data-driven models often suffer from boundary ambiguity in latent spaces, particularly in large-scale grid scenarios, where the trade-off between detection reliability and computational scalability remains challenging to address.

To address these gaps, this paper presents a physics-guided sequential architecture that couples an Improved Bayesian Network (BN) with contrastive representation learning (CL), which we term the BN-CL framework, for power grid anomaly state identification. Unlike purely data-driven or black-box fusion approaches, the Improved BN functions as a "Physical Filter" that integrates PMU, SCADA, topology, and meteorological measurements within a hierarchical architecture to establish a physically consistent representation. In this design, the Improved BN provides a structured foundation for contrastive learning to "sharpen" the latent manifolds, thereby expanding decision margins between normal and anomalous

979-8-3315-8138-1

states in a physically meaningful manner. This methodology achieves Space-Time Synergy while maintaining near-linear computational complexity with respect to grid scale, offering a scalable and interpretable alternative to high-dimensional models in existing literature.

The main contributions of this paper are as follows:

- An integrated BN-CL framework synergistically coupling physics-guided data fusion with representation learning to mitigate the trade-off between detection reliability and scalability.
- An Improved Bayesian Network module acting as a "Physical Filter" that integrates topology and meteorological stress to ensure physically consistent feature representation through Space-Time Synergy.
- A contrastive feature learning scheme designed to "sharpen" physics-based manifolds, which enhances identification robustness while maintaining near-linear computational complexity with respect to grid scale for large-scale grid deployment.

II. INTRODUCTION OF ALGORITHM THEORIES

A. *Improved Bayesian Network for Multi-Source Data Integration*

Bayesian inference combines measurement likelihoods with prior knowledge to integrate uncertain information. In power systems, however, data are highly heterogeneous: PMU provides high-frequency dynamic measurements, SCADA reflects slower steady-state variables, while topology and meteorological data introduce spatial and contextual dependencies.

The proposed Improved Bayesian Network (BN) provides a unified probabilistic formulation that formalizes the integration of heterogeneous priors at the modeling level. The joint probability distribution of the grid state X given multi-source data D is defined as:

$$P(X|D) \propto P(D|X) \cdot P(X|\mathcal{G}) \cdot P(X|E(t,e)) \qquad (1)$$

where $P(D|X)$ is the measurement likelihood, $\mathcal{G}$ denotes the grid topology as a spatial constraint, and $E(t,e)$ represents meteorological stress as an environmental prior. This formulation embodies Space-Time Synergy by incorporating topological connectivity and environmental stressors into the posterior inference.

In contrast to standard Dynamic Bayesian Networks (DBNs), which primarily focus on temporal state transitions $P(X_t|X_{t-1})$, this model ensures physical consistency across asynchronous layers via $\mathcal{G}$ and $E(t,e)$. Furthermore, it differs from hierarchical BNs that emphasize multi-level component dependencies by providing a structured feature space for cross-domain data fusion. This formalization captures interactions that are typically difficult to represent in conventional dynamic or hierarchical models found in the literature.

B. *Principal Component Analysis and Clustering for Fusion Validation*

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies dominant variance directions, projecting high-dimensional fused features into a lower-dimensional subspace. Its mathematical formulation is based on the eigen-decomposition of the estimated covariance matrix:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top \qquad (2)$$

where x_i represents the fused feature vector and μ denotes the mean vector. The leading eigenvectors capture the principal variance structure of the fused features.

In the reduced subspace, clustering methods such as K-means are employed to assess whether normal and abnormal states form separable groups. Distinct clustering structures indicate improved feature discriminability, whereas overlaps suggest limitations in representation quality. This process provides a model-agnostic, auxiliary validation of feature separability for the representations produced by the Improved Bayesian Network.

C. *Contrastive Learning for Anomaly Feature Extraction*

Contrastive learning is a representation learning approach that regularizes the embedding space by constructing positive and negative pairs. Samples from the same operating state form positive pairs to be mapped closer, whereas samples from different states form negative pairs to be pushed apart. This enforces intra-class compactness and inter-class separation, which are essential for effective anomaly recognition.

The training objective adopts the NT-Xent (InfoNCE) loss:

$$L = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(z_i, z_k)/\tau)} \qquad (3)$$

where z_i and z_j represent the embeddings of a positive pair, $sim(\cdot)$ denotes cosine similarity, and τ is a temperature parameter. By optimizing this objective, the fused feature representation becomes both anomaly-sensitive and robust, enabling reliable detection across diverse operating conditions.

III. FRAMEWORK IMPLEMENTATION OF MULTI-SOURCE FUSION BASED ON IMPROVED BAYESIAN NETWORK FOR ANOMALY IDENTIFICATION

The proposed anomaly identification framework operates in three stages: (i) multi-source data fusion based on the Improved BN, (ii) validation of fused features through PCA and clustering, and (iii) anomaly feature extraction via contrastive learning. The overall process is shown in Fig. 1, and the constituent modules are detailed as follows.

A. *Multi-Source Data Preprocessing and Alignment*

Heterogeneous sources are first aligned. SCADA (2 s) data are interpolated to the PMU (20 ms) timeline for temporal consistency using least-squares fitting:

$$\hat{x}(t) = \arg\min_{x(t)} \sum_{k=1}^{n} (y_k - x(t_k))^2 \qquad (4)$$

Data outliers are detected using 3σ or IQR rules, then repaired via LOESS regression:

$$\hat{y}(t_k) = \arg\min_{\beta_0,\beta_1} \sum_{j\in\mathcal{N}(k)} w_{kj}\big(y_j - \beta_0 - \beta_1(t_j - t_k)\big)^2 \qquad (5)$$

Meteorological data are synchronized by spline interpolation, and topology is represented as graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

979-8-3315-8138-1

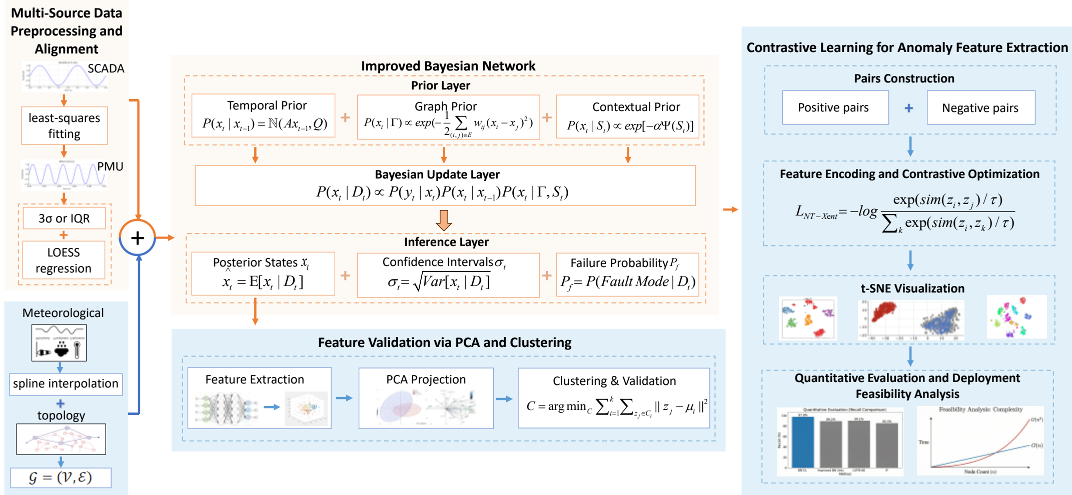

Fig. 1. Overall framework of anomaly identification mode

B. Improved Bayesian Network for Multi-source Data Fusion

The Improved BN fuses PMU, SCADA, topology, and meteorology within a probabilistic framework. State estimation follows Bayesian recursion:

$$P(x_t|D_{1:t}) = \frac{P(y_t|x_t)\int P(x_t|x_{t-1})P(x_{t-1}|D_{1:t-1})dx_{t-1}}{P(y_t|D_{1:t-1})} \tag{6}$$

Meteorological time series are processed via CNN-LSTM to compute stress indices $E(t,e)$. Together with topology, these factors form a Bayesian graphical model, where Gibbs sampling is utilized to approximate branch failure probabilities:

$$P(f_e = 1|D) \sim \text{Gibbs}(P(X|D)) \tag{7}$$

An Interacting Multiple Model (IMM) filter is employed to characterize transitions across Normal, Storm, and Restoration regimes.

C. PCA and Clustering Validation of Fused Features

Dimensionality reduction is applied to the fused feature matrix via PCA. Clustering quality is assessed with the Silhouette coefficient $s(i)$ and the Davies–Bouldin Index (DBI) to quantify the separability of fused features.

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \tag{8}$$

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i} \parallel x - c_i \parallel \tag{9}$$

$$M_{ij} = \parallel c_i - c_j \parallel \tag{10}$$

$$R_{ij} = \frac{S_i+S_j}{M_{ij}} \tag{11}$$

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j\neq i} R_{ij} \tag{12}$$

where $a(i)$ denotes the average intra-cluster distance and $b(i)$ represents the minimum average distance to the nearest neighboring cluster; C_i denotes the sample set of cluster i with cardinality$|C_i|$; c_i is the centroid of cluster i; S_i measures the intra-cluster dispersion; M_{ij} is the inter-centroid distance between clusters i and j; and R_{ij} denotes the separability ratio.

D. Contrastive Learning for Anomaly Feature Extraction

Contrastive learning extracts anomaly-sensitive representations by leveraging augmented views of the same sequence as positive pairs to enforce invariance, while treating sequences from different categories as negative pairs to drive the learning of discriminative features.

Through training, embeddings form distinct clusters of normal and abnormal states, further confirmed by t-SNE visualization. Subsequently, quantitative performance metrics together with computational complexity and scalability analysis are employed to verify the effectiveness and practical feasibility for deployment of the proposed framework.

IV. CASE STUDIES AND RESULTS ANALYSIS

A. Data Description

1) Data sources and structure

The experiments are conducted on the IEEE 39-bus system, extended with renewable generation and external disturbances, to generate three distinct data categories: operational data (PMU and SCADA) representing dynamic and steady-state behaviors, topological data reflecting grid coupling, and meteorological data characterizing environmental influences.

2) Fusion-level definition

To evaluate the Improved BN module, three datasets representing increasing fusion levels are defined: (i) raw single-source (individual PMU or SCADA measurements); (ii) partially fused (PMU and SCADA aligned via interpolation and Kalman filtering); and (iii) multi-source fusion (PMU, SCADA, topology, and meteorology fused via the Improved BN).

Each dataset includes normal and abnormal states simulated through line contingencies, providing well-separated feature distributions to support anomaly identification and quantitative evaluation.

B. PCA-Based Clustering Analysis

1) Visualization method

PCA is applied to examine the separability of fused features across fusion levels. The analysis extracts the first two principal components (PC1 and PC2) to perform dimensionality reduction, providing a compact representation of dominant feature variance between normal and abnormal states.

2) Fusion-level comparison

To evaluate the clustering quality under different fusion configurations, both qualitative visualization and quantitative metrics were employed. Fig. 2 shows the PCA projections for three fusion levels, and Table I lists the corresponding Silhouette coefficient $s(i)$ and the Davies–Bouldin Index (DBI). A higher $s(i)$ and a lower DBI generally signify enhanced intra-cluster cohesion and inter-cluster separation.

a) Raw single-source data: Normal and abnormal samples are almost indistinguishable, indicating that single-sensor data lack discriminative capacity. Although the clustering indices appear acceptable, they are not meaningful due to the absence of abnormal clusters.

b) Partially fused data: After PMU–SCADA fusion, class boundaries begin to emerge, and separability improves. The moderate $s(i)$ (0.35) and DBI (1.12) indicate partial enhancement of cohesion but limited structural consistency.

c) Multi-source data fusion: The fusion based on the proposed Improved BN achieves clear class separation and relatively compact intra-class clustering. The highest $s(i)$ (0.71) and lowest DBI (0.58) confirm enhanced structural coherence and strong feature discriminability.

TABLE I. QUANTITATIVE CLUSTERING METRICS UNDER DIFFERENT FUSION LEVELS

Fusion Level	$s(i)$	DBI
Single-source	0.99	0.01
Partially fused	0.35	1.12
Multi-source data fusion	0.71	0.58

3) Observations

The comparison across fusion levels demonstrates that the multi-source fusion based on the proposed Improved BN produces the most coherent and discriminative feature space. The fused features maintain compact intra-cluster cohesion and pronounced inter-class margins, confirming that the Improved BN provides a robust and interpretable foundation for subsequent anomaly state identification.

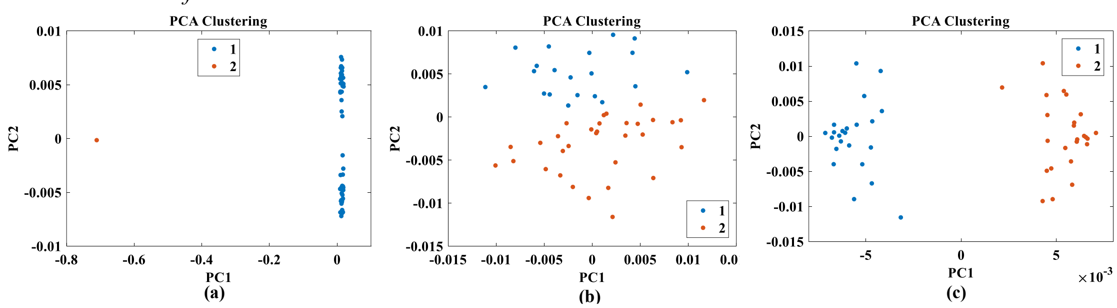

Fig. 2. PCA projections for three fusion levels. (a) single-source data; (b) partially fused PMU–SCADA data; (c) multi-source fusion based on the Improved Bayesian Network.

C. Contrastive Representation Learning for Enhanced Representation Separability

1) Training performance

The contrastive learning module is trained on multi-source data, where the NT-Xent loss function optimizes latent feature representations by pulling semantically similar samples closer and pushing dissimilar ones apart, thereby distinguishing normal and abnormal operating states.

As shown in Fig. 3, the loss value decreases steadily over training epochs, indicating stable convergence and effective optimization. This smooth decline demonstrates that the model progressively enhances its ability to capture intrinsic correlations among multi-source inputs, producing discriminative feature embeddings that generalize well to unseen grid conditions.

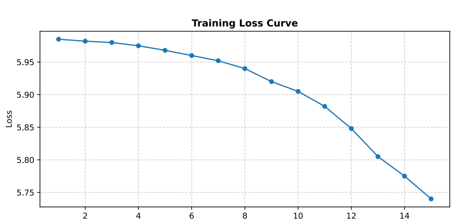

Fig. 3. Trend of the NT-Xent loss function during contrastive training

2) Feature distribution visualization

After training, the deep feature embeddings obtained from the encoder are projected into a two-dimensional space using t-SNE for visualization.

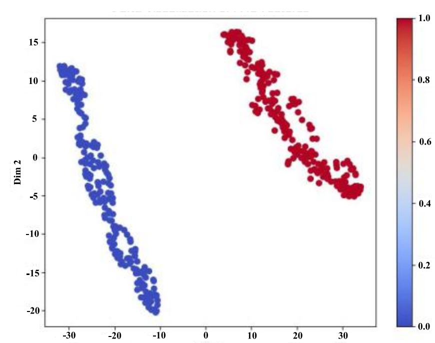

Fig. 4. t-SNE visualization of learned feature representations

As illustrated in Fig. 4, the learned features form two distinct manifolds corresponding to normal (blue) and abnormal (red) states. The clear separation and compact intra-class distributions indicate that the model successfully learns consistent patterns within each operating condition, while maintaining large margins between normal and faulted states. Under the same clustering evaluation protocol as the PCA-based analysis in Section IV-B, quantitative evaluation shows that the $s(i)$ increases from 0.71 to 0.91, accompanied by a

979-8-3315-8138-1

reduced DBI, indicating decreased intra-class dispersion and enlarged inter-class separation in the fused latent space. Such a structured representation provides a stable geometric basis for subsequent anomaly identification.

D. *Anomaly Identification and Performance Evaluation*

To conduct anomaly state identification, a distance-based anomaly identification mechanism is applied in the fused latent space. Based on this mechanism, the effectiveness of the proposed BN-CL framework is quantitatively evaluated, and Table II summarizes a performance comparison with Isolation Forest and LSTM-AE benchmarks. Based on simulations of the IEEE 39-bus system, the proposed framework attains a Recall of 97.8% and suppresses the False Positive Rate (FPR) to 1.1%, significantly outperforming the baseline data-driven methods.

Furthermore, an ablation study was conducted to isolate the incremental contribution of contrastive representation learning beyond physical data fusion. In this study, the “Improved BN Only” case represents a baseline where the same distance-based anomaly identification mechanism is applied directly to the fused feature vectors, without contrastive learning. While this physical fusion provides a solid baseline Recall of 89.2%, the integration of Contrastive Learning refines class manifolds and drives the performance to 97.8%. These results empirically demonstrate that the representation separability established by contrastive representation learning can be effectively translated into reliable anomaly detection performance.

TABLE II. PERFORMANCE COMPARISON OF ANOMALY DETECTION METHODS

Method	Accuracy	Precision	Recall	FPR
Isolation Forest	86.4%	84.1%	85.3%	5.2%
LSTM-AE	91.2%	89.5%	90.1%	3.4%
Improved BN Only	90.5%	88.2%	89.2%	2.8%
Full BN-CL Framework	98.2%	97.5%	97.8%	1.1%

E. *Computational Complexity and Scalability Analysis*

To evaluate the practical feasibility for industrial deployment, we perform a qualitative analysis of the framework's computational complexity and scalability. Due to the factorized probabilistic structure of the Improved BN and the sample-wise forward computation in contrastive learning, the probabilistic inference and representation learning stages exhibit near-linear computational complexity with respect to the number of nodes on the order of $O(n)$, where n denotes the number of nodes. This near-linear property provides strong evidence for the scalability of the proposed framework, as computational costs do not experience a non-linear explosion when the node count increases, thereby mitigating the risk of rapid computational growth observed in fully connected models. Based on the synergy of high scalability, near-linear computational efficiency, and functional reliability, the results indicate that the framework is practically feasible for real-world deployment. The optimized complexity enables the system to support near real-time responses for grid operators even in large-scale system environments.

V. CONCLUSION

This paper presents a comprehensive BN-CL framework for anomaly state identification in power grid operation, integrating the Improved BN for multi-source data fusion and contrastive learning for feature-level anomaly recognition. The proposed Improved BN model effectively unifies heterogeneous data sources—including PMU, SCADA, topological, and meteorological information—through hierarchical Bayesian inference and spatiotemporal modeling, enabling coherent representation of grid dynamics. Experimental studies based on the IEEE 39-bus system demonstrate that the fused data significantly enhance the feature separability between normal and abnormal states. Furthermore, the contrastive learning module sharpens latent representations and strengthens the robustness of anomaly detection. Overall, the proposed approach provides an interpretable and scalable solution for intelligent grid monitoring and early warning. Future research will prioritize large-scale validation on real-world utility datasets to further assess its practical industrial viability. Additionally, we plan to incorporate adaptive learning mechanisms to improve online generalization and cross-scenario transferability, to further improve robustness under diverse and evolving grid operating conditions.

ACKNOWLEDGMENT

This work is supported by Project of State Grid Corporation of China (No. DK-KJ XM 2025-12).

REFERENCES

[1] Massignan J A D, London J B A, Bessani M, et al. Bayesian inference approach for information fusion in distribution system state estimation[J]. IEEE transactions on smart grid, 2021, 13(1): 526-540.

[2] Cui Y, Ma H, Saha T. Multi-source information fusion for power transformer condition assessment[C]//2016 IEEE Power and Energy Society General Meeting (PESGM). IEEE, 2016: 1-5.

[3] Li M, Tan J, Zhang J, et al. Design of Multi-Source Heterogeneous Data Fusion Algorithm for Distribution Networks Based on Improved Kalman Filter[J]. IEEE Access, 2025.

[4] Yuan Y, Dehghanpour K, Wang Z, et al. Multisource data fusion outage location in distribution systems via probabilistic graphical models[J]. IEEE Transactions on Smart Grid, 2021, 13(2): 1357-1371.

[5] Wang S P, Zhao D M. A hierarchical power grid fault diagnosis method using multi-source information[J]. IEEE Transactions on Smart Grid, 2019, 11(3): 2067-2079.

[6] Jiang H, Zhang J J, Gao W, et al. Fault detection, identification, and location in smart grid based on data-driven computational methods[J]. IEEE Transactions on Smart Grid, 2014, 5(6): 2947-2956.

[7] Peng H, Wang J, Ming J, et al. Fault diagnosis of power systems using intuitionistic fuzzy spiking neural P systems[J]. IEEE Transactions on Smart Grid, 2017, 9(5): 4777-4784.

[8] Zhao Jianwei, Lin Yuchang, Chen Sheng, et al. Meteorological disaster fault prediction method for power systems based on CNN-LSTM algorithm [J]. Advanced Technology of Electrical Engineering and Energy, 2025, 44(03): 109-118.

[9] Deng Lijuan, Xie Yonghua, Wu Qilin, et al. Research on real-time fault early warning method for distribution network based on multi-source data fusion [J]. Digital Communication World, 2025, (02): 64-66.

[10] Lin Z, Xia T, Ye Y, et al. Application of wide area measurement systems to islanding detection of bulk power systems[J]. IEEE Transactions on Power Systems, 2013, 28(2): 2006-2015.

[11] Shahraeini M, Ghazizadeh M S, Javidi M H. Co-optimal placement of measurement devices and their related communication infrastructure in wide area measurement systems[J]. IEEE Transactions on Smart Grid, 2012, 3(2): 684-691.

[12] Ravikumar G, Ameme D, Misra S, et al. iCASM: An information-centric network architecture for wide area measurement systems[J]. IEEE Transactions on Smart Grid, 2020, 11(4): 3418-3427.

979-8-3315-8138-1