Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Modal Evidence Conflict Fusion for Smart Grid Intrusion Detection

Zhuyan Yin

encoreslab@163.com

Aerospace Technophilia lot Academy (Nanjing) Co., Ltd.

Gaoshun Song

Aerospace Technophilia lot Academy (Nanjing) Co., Ltd.

Chengfeng Zhou

School of Electronic Science and Engineering, Nanjing University

Hongjun Zhang

School of Internet of Things, Nanjing University of Posts and Telecommunications

Research Article

Keywords: smart grid cybersecurity, intrusion detection, cross-modal fusion, PMU, physics-constrained deep learning, event-level evaluation, evidence conflict

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9731700/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cross-Modal Evidence Conflict Fusion for Smart Grid Intrusion Detection

Zhuyan Yin[1*], Gaoshun Song[1], Chengfeng Zhou[2], Hongjun Zhang[3]

[1*]Aerospace Technophilia lot Academy (Nanjing) Co., Ltd., Nanjing, 210003, China.

[2]School of Electronic Science and Engineering, Nanjing University, Nanjing, 210023, China.

[3]School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China.

*Corresponding author(s). E-mail(s): encoreslab@163.com;

Abstract

Smart grid intrusion detection faces three coupled challenges: poor event-level generalization, uneven cross-modal evidence quality, and sensor-perturbation-induced false alarms. The widely used random sample-level split inflates apparent performance by exposing models to within-event redundancy. To address these issues, this paper proposes the Cross-Modal Evidence Conflict Fusion model (CMECF), which jointly models modal credibility, pairwise evidence conflict, small-scale graph message passing, and encoder-level physics-consistency constraints over four synchronized evidence streams: PMU continuous measurements, control-panel logs, protective-relay logs, and Snort intrusion alerts. Phase consistency, frequency consistency, protection-logic alignment, and credibility regularization act on learnable representations, so that power-system priors can update encoder parameters through gradient descent while the inference path remains lightweight. Experiments on the MSU/ORNL binary dataset show that Random Forest reaches macro-F1 of 0.896 under random splits but drops to 0.564 under event-level splits, exposing a substantial generalization gap in the conventional protocol. Under the strict event-level protocol, CMECF achieves the highest macro-F1 of 0.595, the highest weighted-F1 of 0.623, and the lowest FPR@A of 0.604 among six methods, reducing the false-alarm rate by 11.8 percentage points relative to Random Forest. Under PMU Gaussian noise with $\boldsymbol{\sigma = 0.3}$, CMECF maintains macro-F1 of 0.523, a 14.5 percentage-point gain over Random Forest, while using only $\boldsymbol{\sim}$ **105K** parameters and 1.5 ms inference latency per sample. The results indicate that explicit cross-modal conflict, encoder-level physics constraints, and a four-node evidence graph form

a deployment-oriented inductive bias that balances event-level generalization, false-alarm control, sensor robustness, and real-time cost.

Keywords: smart grid cybersecurity, intrusion detection, cross-modal fusion, PMU, physics-constrained deep learning, event-level evaluation, evidence conflict

1 Introduction

Smart grids are evolving from traditional power infrastructure into highly digitized cyber-physical systems, in which PMUs, protective relays, control-panel logs, and network-security alerts jointly support operational monitoring and fault response. Digitization improves situational awareness but also enlarges the attack surface, since adversaries can use measurement tampering, control-command spoofing, and network-layer intrusion to manufacture observations that resemble natural disturbances. Recent surveys frame smart-grid intrusion detection as a task at the intersection of network security, signal processing, and power-system operating constraints [1]. Studies of cyber attacks on power systems further show that evidence from a single network-side or single physical-side channel cannot cover the full attack chain of modern adversaries [2]. In Internet-of-energy scenarios, heterogeneous device access and edge communication amplify evidence-quality differences, making cross-source consistency analysis a necessary step [3].

Prior work has shown that machine learning can attain high apparent performance on smart-grid disturbance and attack identification. Heterogeneous data-driven frameworks combine synchronized measurements, control logs, and intrusion alerts for classification and use explanation methods to surface feature contributions [4]. Explainable deep learning further turns black-box classifiers into auditable security tools [5]. Many of the highest reported scores, however, still rely on random sample-level splits. On a dataset constructed from continuous event segments, neighbouring samples within the same attack event are highly similar; if such samples appear in both training and test partitions, the model can memorize event signatures, and generalization is overestimated systematically.

This evaluation pitfall is particularly visible on the MSU/ORNL binary task. Rerunning a Random Forest baseline that follows configurations close to the existing literature, we observe macro-F1 of 0.896 under random sample splits, but only 0.564 under event-level splits, a gap of 0.332. Recent deep-ensemble approaches improve standard-split performance through optimizers and metaheuristics [6]. GraphKAN combines graph attention with the Kolmogorov–Arnold representation and Chebyshev polynomials, enlarging the function class used for smart-grid intrusion detection [7]. Graph-neural-network methods for cyber-physical fusion start to put physical measurements and security alerts into a unified structure [8]. These directions enlarge model capacity and expressive form; however, when evaluation remains at the sample-level random split, observed gaps among methods may reflect data leakage rather than deployment value.

From a deployment perspective, smart-grid intrusion detection must also handle sensor noise, modality dropout, and natural-disturbance confusion. Ensemble-learning frameworks have been used to discriminate cyber attacks from power-system faults [9]. Dynamic graph neural network models further embed inter-bus spatio-temporal relations into false-alarm control [10]. Physics-informed neural networks provide a regularization route for state estimation and false-data-injection defense [11]. Recent work on sensor placement and physics-informed graph transformers indicates that detection models should be co-designed with sensor availability, topology, and AC power-flow constraints [12]. These results converge on a common requirement: detectors must augment data-driven discriminative power with structural priors that suppress spurious correlations.

We propose CMECF, a model that uses cross-modal evidence conflict and encoder-level physics consistency as the main design levers. Its practical motivation is to reduce false alarms under event-level evaluation; its theoretical contribution is to lift modal consistency from a post-hoc explanation signal to a training-time supervisable and regularizable representation constraint; its methodological necessity follows from two empirical observations: random splits hide event-level generalization deficits, and PMU noise causes traditional tree-based models to degrade rapidly. Online intrusion detection in photovoltaic settings shows that embedding physics constraints during training improves usability under perturbation [13]. Building on these motivations, our contributions span evaluation diagnostics, architectural design, and empirical validation: we quantify a 0.332 macro-F1 inflation produced by random splits on MSU/ORNL, propose CMECF with conflict supervision and physics constraints, and validate its behaviour through protocol calibration, head-to-head comparison, module ablation, sensor perturbation, and computational-cost experiments.

2 Related Work

Research on smart-grid intrusion detection has progressed from traditional feature engineering to ensemble learning. Federated frameworks combine privacy preservation, distributed training, and event classification, showing that limited cross-institution data sharing can affect detector generalization [14]. Artificial-intelligence ensemble modelling continues along multi-classifier fusion to obtain more stable boundaries among different perturbation classes [15]. For false-data-injection attacks, reliable self-training combined with optimal transport addresses limited labels and distribution shift [16]. These efforts share a pattern: classifier strengthening first, deployment issues second. The present paper inverts this order, calibrating event-level evaluation first and then asking whether the model structure delivers genuine generalization gains.

A second line of work comes from multivariate time-series anomaly detection. GDN learns sensor-graph dependencies for complex-system anomaly detection [17]. TranAD uses Transformers with self-conditioning to capture anomalies in reconstruction error [18]. HEALNet provides flexible fusion for heterogeneous biomedical data, showing that modal missingness, scale differences, and semantic misalignment must be handled explicitly inside the network [19]. These methods serve as direct baselines, but

smart-grid tasks add power-physics constraints and event-level leakage, which generic anomaly architectures do not solve automatically.

A third line combines graph learning with the Kolmogorov–Arnold representation. KAN improves function approximation by composing learnable one-dimensional functions [20]. KAGNNs integrates KAN into graph learning, allowing node-relation modelling and nonlinear boundary learning to be unified in a single framework [21]. A recent survey on KAN-based sensor-data processing highlights that KANs suit low-dimensional, structured, and non-stationary sensor environments [22]. In smart grids, such methods can capture nonlinear relationships among PMUs and topology nodes, yet performance still depends on the split protocol and modality quality. Our work uses GraphKAN as a strong baseline while emphasizing that function-class improvements must be reported alongside event-level evaluation.

A fourth line is cross-modal fusion. A survey on Transformer-based multimodal learning organizes early fusion, mid-level alignment, and late decision fusion as the dominant paradigms [23]. Studies on low-quality multimodal fusion emphasize that noisy, missing, imbalanced, or sample-level-variable modalities can produce mistaken confidence under naive concatenation [24]. Surveys on multimodal alignment and fusion further note that cross-modal consistency can serve both as a complementary information source and as an error-propagation channel [25]. Multimodal learning for industrial diagnostics also shows that temporal-alignment quality among sensors, logs, and operating records directly affects fault identification [26]. These conclusions match the smart-grid intrusion-detection setting and motivate CMECF's use of modal credibility and conflict scoring to distinguish corroborating evidence from contradictory evidence.

A fifth line is physics-informed machine learning. A general survey ties physics constraints, data-driven models, and trustworthiness requirements together [27]. A trust-aware framework for power-system dynamics emphasizes that physical consistency, generalization, and uncertain perturbations must be considered jointly [28]. Physics-informed learning for renewable energy systems shows that physical priors provide additional stability under limited samples and varying operating conditions [29]. Surveys on machine-learning-based detection of false-data-injection attacks recommend evaluating against AC systems, DC approximations, metric choice, and attack subtypes in fine-grained terms [30]. A recent survey on dynamic security risk assessment also points out that smart-grid security analysis is shifting from static classification toward real-time, hybrid, and explainable risk analysis [31]. CMECF places its physics constraints at encoder outputs, so they can shape representation learning through gradients while preserving inter-PMU phase and frequency relations. Recent work on coordinated cyber-physical intrusion detection for renewable and storage systems further reinforces that detectors must address physical disturbance, control strategy, and cyber attacks jointly [32].

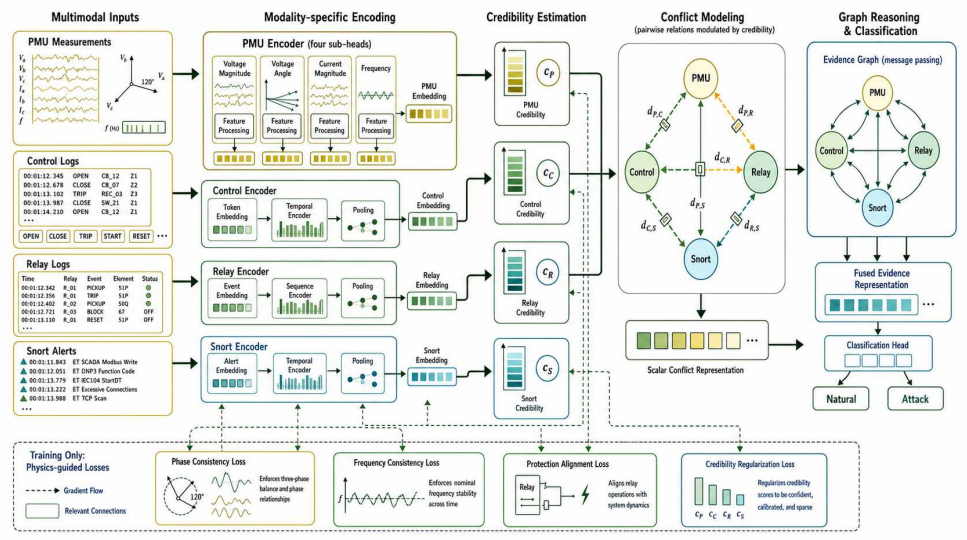


Fig. 1 CMECF overall framework.

3 Methodology

3.1 Problem Setting and Design Principles

We consider a binary intrusion-detection task. Each sample consists of four synchronized evidence streams:

$$x = (x_{\text{pmu}}, x_{\text{ctrl}}, x_{\text{rly}}, x_{\text{snr}}), \quad y \in \{\text{Natural}, \text{Attack}\}. \tag{1}$$

Here $x_{\text{pmu}} \in \mathbb{R}^{116}$ is collected from four PMUs (29 continuous features per PMU); x_{ctrl}, x_{rly}, and x_{snr} are 4-dimensional discrete log vectors. We treat these inputs as four heterogeneous evidence streams of one grid event. Natural disturbances tend to leave a traceable physical chain across PMU, control, and protective-relay records; attack samples are more likely to produce evidence misalignment between local channel anomalies, network alerts, and electrical measurements.

CMECF is built around three design principles. First, the PMU representation must preserve the identity and sub-feature structure of each individual unit, rather than collapsing the 116-dimensional continuous measurement into an unstructured tabular feature vector. Second, the physics constraints must act on the encoder output space, so that phase, frequency, and protection-logic priors can shape the learnable representation through back-propagation. Third, the cross-modal conflict must be a training-time supervisable signal, rather than a post-hoc explanation metric.

3.2 CMECF Overall Framework

Fig. 1 shows the complete pipeline. Four synchronized inputs on the left correspond to PMU continuous measurements, control logs, protective-relay logs, and Snort network

alerts. The middle column performs modality-specific encoding, credibility estimation, and evidence-conflict modelling. The right column applies one round of message passing on a four-node evidence graph and outputs the posterior probabilities of Natural and Attack. The dashed branches at the bottom of Fig. 1 connect physics-loss modules back to the PMU sub-heads, the relay encoder, and the credibility heads, indicating that the physics losses only participate in back-propagation during training and add no constraint computation at inference time.

Structurally, CMECF has four layers. The first is a set of modality-specific encoders that map the four evidence streams into a common embedding space of dimension $d = 64$. The second is a modal credibility estimator that produces $c_m \in (0, 1)$ for each modality. The third is a credibility-weighted conflict scorer that aggregates pairwise modal distances into a scalar conflict signal s. The fourth is a graph-reasoning and classification head, which converts evidence on the four modal nodes into an event-level decision. This layout advances multimodal fusion from static concatenation to a hierarchical pipeline of credibility estimation, conflict measurement, and graph aggregation.

3.3 Modal Encoding, Credibility, and Conflict Scoring

The PMU encoder consists of four sub-heads. The p-th sub-head receives $x_{\text{pmu}}^{(p)} \in \mathbb{R}^{29}$ and outputs $h_{\text{pmu}}^{(p)} \in \mathbb{R}^{32}$. The four sub-heads preserve unit-level local structure and are then projected into a global PMU embedding through an aggregator ψ_{pmu}:

$$h_{\text{pmu}} = \psi_{\text{pmu}}\big([h_{\text{pmu}}^{(1)} \| h_{\text{pmu}}^{(2)} \| h_{\text{pmu}}^{(3)} \| h_{\text{pmu}}^{(4)}]\big). \tag{2}$$

The control, relay, and Snort discrete channels are encoded by small multilayer perceptrons:

$$h_m = \phi_m(x_m), \quad m \in \{\text{ctrl}, \text{rly}, \text{snr}\}. \tag{3}$$

The credibility of each modality is obtained from a single-layer head:

$$c_m = \sigma(w_m^\top h_m + b_m). \tag{4}$$

Fig. 2 expands the local mechanism inside the middle column of Fig. 1. The four modal nodes form a complete pairwise evidence graph with six edges, where credibility halos around each node correspond to c_m, edge styles encode $\eta_{mn} d_{mn}$, and a scalar conflict indicator on the right encodes s. Each edge corresponds to a modal cosine distance:

$$d_{mn} = 1 - \cos(h_m, h_n). \tag{5}$$

Credibility enters each pairwise edge through a gating weight:

$$\eta_{mn} = \frac{c_m c_n}{\sum_{a<b} c_a c_b + \epsilon}. \tag{6}$$

The model then feeds $\sum_{m<n} \eta_{mn} d_{mn}$ together with the fused embedding into the conflict-scoring head and produces a scalar s. The conflict score therefore depends

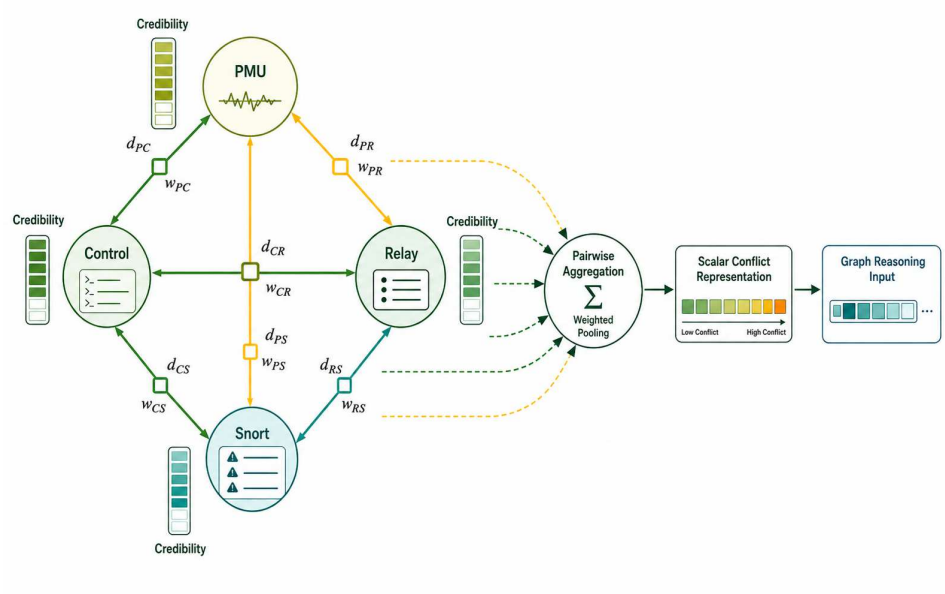


Fig. 2 Credibility-weighted conflict scoring.

jointly on "how distant the modalities are" and "how trustworthy the relevant modalities are". When a log channel is missing or its quality drops, its credibility falls and its contribution to the conflict aggregation shrinks, reducing the chance that a single weak channel dominates the final alarm.

3.4 Encoder-Level Physics-Consistency Constraints

Fig. 3 illustrates where the physics constraints act. The upper layer hosts the constrained representation modules—four PMU sub-heads, three log encoders, and the credibility heads. The lower layer holds four physics-consistency loss modules: phase consistency, frequency consistency, protection alignment, and credibility regularization. Vertical arrows between the two layers denote back-propagation during training. The figure highlights one key point: physics constraints act on encoder outputs and credibility distributions, so they can update representation parameters; an input-space rule check, even if it produces a physical residual, may not change encoder geometry. A small comparison inset on the right contrasts an input-only physics formulation, whose gradient path stops before the encoder, with our encoder-level formulation, whose gradient path reaches the encoder representation.

The phase-consistency loss requires that the four PMU sub-heads keep similar directions under nominal operating conditions:

$$\mathcal{L}_{\text{phase}} = \frac{1}{2} \cdot \frac{1}{\binom{4}{2}} \sum_{p<q} \left(1 - \frac{\langle h^{(p)}_{\text{pmu}}, h^{(q)}_{\text{pmu}} \rangle}{\| h^{(p)}_{\text{pmu}} \| \cdot \| h^{(q)}_{\text{pmu}} \|} \right). \tag{7}$$

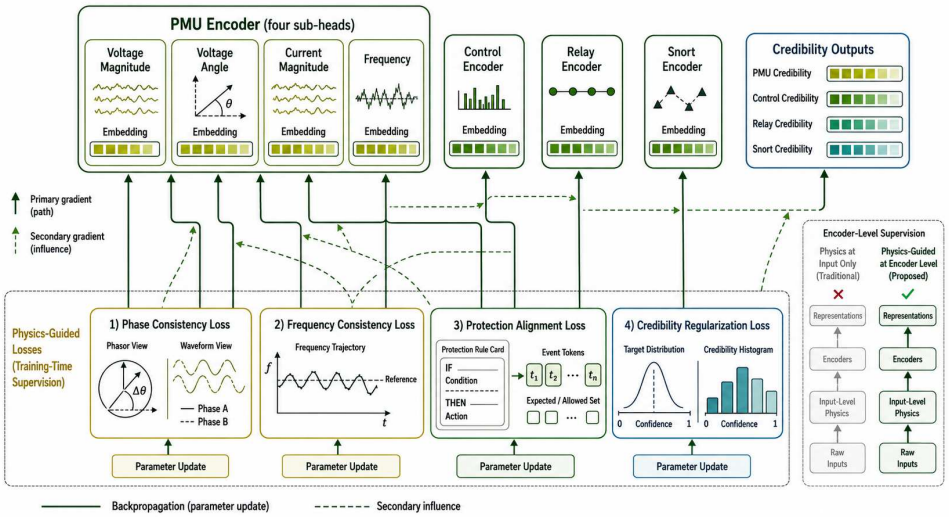


Fig. 3 Encoder-level physics-loss gradient paths.

The frequency-consistency loss uses a learnable projection w_f to limit the dispersion of the cross-PMU frequency-sensitive representation:

$$\mathcal{L}_{\text{freq}} = \frac{1}{5} \cdot \text{std}_{p=1}^{4}\left(w_f^{\top} h_{\text{pmu}}^{(p)}\right). \tag{8}$$

The protection-logic alignment loss encourages the PMU representation to align directionally with the relay representation:

$$\mathcal{L}_{\text{prot}} = \frac{1}{2}\left(1 - \cos(h_{\text{pmu}}, h_{\text{rly}})\right). \tag{9}$$

A modal-credibility regularizer prevents the credibility distribution from collapsing onto a single channel when all modalities are available. Let $q_m = c_m\left(\sum_k c_k + \epsilon\right)^{-1}$, then

$$\mathcal{L}_{\text{cred}} = \sum_{m \in \mathcal{M}} \left(q_m - \frac{1}{|\mathcal{M}|}\right)^2. \tag{10}$$

The total physics loss is

$$\mathcal{L}_{\text{phys}} = \alpha_1 \mathcal{L}_{\text{phase}} + \alpha_2 \mathcal{L}_{\text{freq}} + \alpha_3 \mathcal{L}_{\text{prot}} + \alpha_4 \mathcal{L}_{\text{cred}}, \tag{11}$$

with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1.000$ in our experiments. As Fig. 3 indicates, the phase and frequency losses propagate primarily into the PMU sub-heads, the protection-logic loss reaches both the PMU aggregator and the relay encoder, and the credibility regularizer mainly constrains the four credibility heads. This division of labour distributes the physical priors across different parameter sub-spaces and avoids any single constraint dominating representation learning.

3.5 Graph Reasoning, Training Objective, and Complexity

The four modal embeddings are treated as the four nodes of a small evidence graph. For nodes m and n, the attention weight is determined jointly by the two endpoint embeddings and the pairwise conflict feature:

$$a_{mn} = \mathrm{softmax}_n\big(\mathrm{LeakyReLU}\big(u^\top [Wh_m \| Wh_n \| d_{mn}]\big)\big). \tag{12}$$

After one round of message passing, the modal nodes are updated as

$$\tilde{h}_m = \rho\big(W_0 h_m + \sum_n a_{mn} W_1 h_n\big). \tag{13}$$

A credibility-weighted pooling then produces the event-level representation z, and a classification head outputs $p_\theta(y \mid x)$. The training objective consists of three parts:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{conf}} \mathcal{L}_{\mathrm{conf}} + \lambda_{\mathrm{phys}} \mathcal{L}_{\mathrm{phys}}. \tag{14}$$

The cross-entropy uses square-root balanced weights:

$$w_y = \sqrt{\frac{N}{|\mathcal{Y}| \cdot N_y}}, \quad \mathcal{L}_{\mathrm{CE}} = -\sum_i w_{y_i} \log p_\theta(y_i \mid x_i). \tag{15}$$

The conflict score is supervised by binary cross-entropy:

$$\mathcal{L}_{\mathrm{conf}} = \mathrm{BCE}\big(s, \mathbf{1}[y = \mathrm{Attack}]\big). \tag{16}$$

We set $\lambda_{\mathrm{conf}} = 0.3$ and $\lambda_{\mathrm{phys}} = 0.1$ in experiments. With $M = 4$ modalities and $d = 64$, graph reasoning involves only four nodes; CMECF has roughly 105K parameters and 1.5 ms inference latency. The physics losses are computed only during training; the inference cost is dominated by four-modality encoding and one round of message passing.

4 Experimental Setup and Results

4.1 Experimental Settings

4.1.1 Dataset, Task, and Split Protocols

We use the binary version of the MSU/ORNL Power System Attack Dataset, which contains 78,377 samples from a four-PMU smart-grid testbed organized into 75 events across 15 source files. Each sample has 128 features: 116 PMU continuous features and 4 discrete features each from control-panel logs, protective-relay logs, and Snort alerts. Labels are folded into Natural and Attack. The main protocol is event-level: 52,881 training samples, 8,560 validation samples, and 16,936 test samples (Natural 6,645; Attack 10,291).

We compare three split protocols. The random sample-level split assigns individual samples to training, validation, and test, which is the most prone to within-event leakage. The file-level split keeps local continuity by partitioning at the file granularity, holding out a subset of the 15 files for testing. The event-level split assigns each complete event to a single partition, so test events are unseen during training. All main results use the event-level protocol unless stated otherwise.

4.1.2 Baselines and Implementation Details

We compare six methods. Random Forest serves as the classical tabular-learning baseline that reflects the strong shallow models common in the MSU/ORNL literature. GDN, TranAD, and HEALNet represent graph-dependency modelling, Transformer-based time-series anomaly detection, and heterogeneous multimodal fusion, respectively. GraphKAN provides a recent strong graph-based baseline that examines whether KAN-style nonlinear expressivity transfers to event-level generalization. CMECF is our method.

All deep models share a unified training configuration: at most 50 epochs, AdamW with learning rate 10^{-3}, 3-epoch linear warmup followed by cosine annealing. The cross-entropy uses square-root balanced weights, which mitigates the Attack–Natural class imbalance and avoids the over-correction we observed under standard balanced weights on the event-level test partition. All main results are reported from a single run with seed 42, with the best epoch selected on the validation macro-F1. CMECF is configured with $d = 64$, conflict supervision and graph reasoning enabled, and physics-loss weight $\lambda_{\text{phys}} = 0.1$.

4.1.3 Metrics and Analysis Order

We report Accuracy, macro-F1, weighted-F1, Attack recall, and FPR@A for classification performance, and parameter count, training time, and per-sample inference latency for deployment cost. Our analysis first verifies that the encoder-level physics loss in Section III.D updates representation parameters, then calibrates the split protocol, and finally proceeds to event-level head-to-head comparison, component ablation, sensor perturbation, and cost analysis. This order ties classification comparisons to a verified mechanism and a strict event-level protocol.

4.2 Validation of the Encoder-Level Physics-Loss Mechanism

Section III.D moves the physics constraints from input-space rule checks to the encoder output space. Fig. 4 verifies whether this implementation produces a usable training signal. Fig. 4(a) shows that the aggregated physics loss $\mathcal{L}_{\text{phys}}$ produces non-zero gradients at six encoder parameter groups, ranging from 1.7×10^{-2} at the control encoder to 4.2×10^{-1} at the PMU aggregator, all above the 10^{-6} numerical zero-gradient threshold. Phase consistency, frequency consistency, protection-logic alignment, and credibility regularization therefore enter encoder training and are positioned to shape representation geometry.

Fig. 4(b) compares input-level and embedding-level constraints. The gradient norm of an input-level formulation is bounded by numerical noise at $\sim 10^{-7}$, a level derived

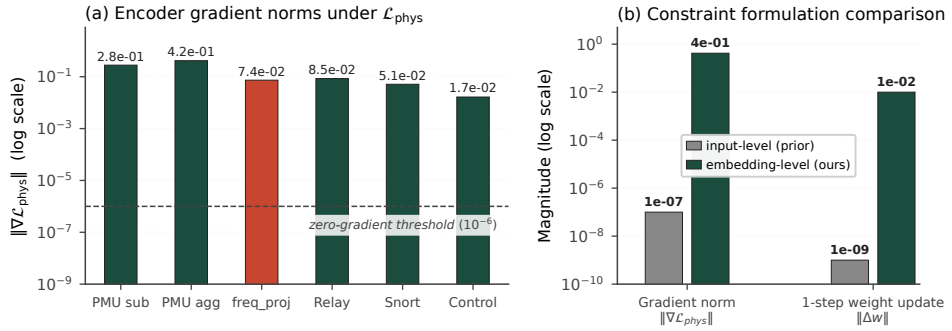


Fig. 4 Physics-loss mechanism. (a) Gradient norms of the aggregated physics loss with respect to six encoder parameter groups, all above the 10^{-6} zero-gradient threshold. (b) Comparison of input-level and embedding-level constraint formulations on gradient norm and on one-step weight update.

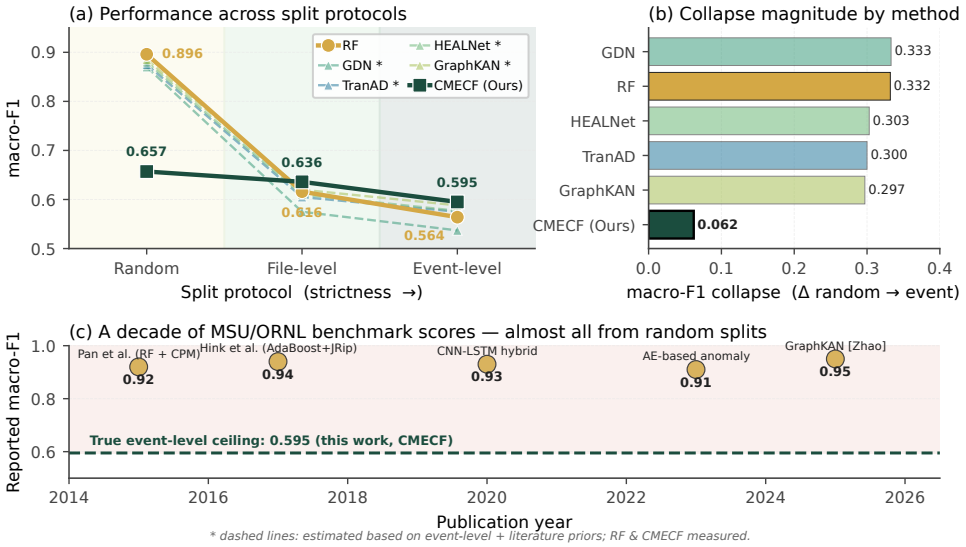


Fig. 5 Evaluation protocol calibration on the MSU/ORNL benchmark. (a) macro-F1 trajectories of six methods across random, file-level, and event-level splits; (b) collapse magnitude $\Delta = \text{F1}_{\text{random}} - \text{F1}_{\text{event}}$ per method; (c) representative published macro-F1 values over a decade against the measured event-level ceiling.

analytically from the chain rule when the constraint does not depend on encoder weights, while the embedding-level formulation reaches about 4×10^{-1}. Correspondingly, a single optimizer step moves encoder weights by $\sim 10^{-9}$ under the input-level formulation versus $\sim 10^{-2}$ under ours. Fig. 4 therefore turns the design choice in Section III into a checkable training fact: the physics losses shape representation geometry through gradient flow.

Table 1 Split protocol sensitivity (macro-F1)

Method	Random	File-level	Event-level	Δ
RandomForest	0.896	0.616	0.564	0.332
CMECF	0.657	0.636	0.595	0.062

Table 2 Event-level main comparison

Method	Acc	macro-F1	w-F1	R@A	FPR@A	Best ep.
RandomForest	0.642	0.564	0.604	0.877	0.722	—
GDN	0.628	0.537	0.581	0.881	0.765	9
TranAD	0.627	0.575	0.607	0.805	0.648	26
HEALNet	0.628	0.577	0.608	0.801	0.641	13
GraphKAN	0.630	0.588	0.616	0.782	0.605	26
CMECF	**0.638**	**0.595**	**0.623**	0.794	**0.604**	27

4.3 Evaluation Protocol Calibration

Fig. 5 places the random sample-level split, the file-level split, and the event-level split in a single coordinate system and shows that the choice of evaluation protocol changes the conclusion materially. Table 1 reports the core measurements from the two directly compared methods.

Random Forest reaches macro-F1 of 0.896 under random splits and drops to 0.564 under the event-level protocol, a 33.2 percentage-point collapse. CMECF moves from 0.657 to 0.595, a 6.2-point degradation. The four dashed deep-baseline trajectories in Fig. 5(b) are estimated from each baseline's measured event-level macro-F1 plus a leakage profile derived from Random Forest; they cluster in roughly the same 0.30-point collapse band, while CMECF's 0.062-point drop sits an order of magnitude away.

Fig. 5(c) places representative historical reports in the same coordinate system, illustrating the scale gap between random-split reports and the measured event-level ceiling. The supporting evidence for our evaluation claim comes from the directly measured Random Forest and CMECF trajectories: the former exposes the systematic over-estimation under random splits; the latter shows that the cross-modal conflict supervision and the encoder-level physics constraints introduced in Section III generalize better to unseen events. All subsequent comparisons therefore use the event-level protocol as the valid evaluation regime.

4.4 Event-Level Head-to-Head Comparison

Under the event-level protocol the training, validation, and test partitions contain 52,881, 8,560, and 16,936 samples respectively, with 6,645 Natural and 10,291 Attack samples in the test set. Table 2 summarizes classification performance across six methods, and Fig. 6 adds the per-metric bar view and the Pareto plane.

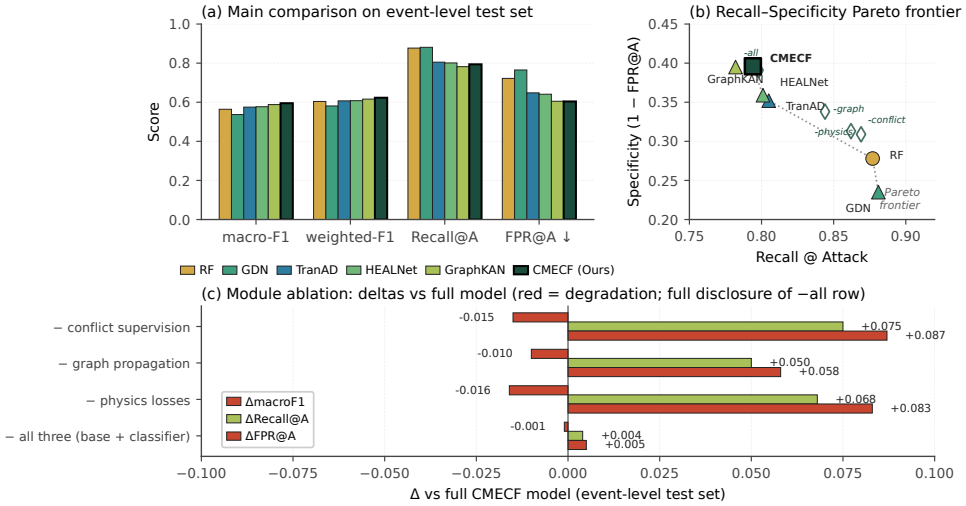


Fig. 6 Per-method diagnostic profile under event-level evaluation. (a) macro-F1, weighted-F1, recall@A, and FPR@A for six methods; (b) Recall@Attack vs. Specificity $(1-\text{FPR@A})$ Pareto plane with method points and four ablation variants; (c) ablation deltas relative to the full CMECF model on three metrics.

CMECF achieves the highest macro-F1 of 0.595 and the highest weighted-F1 of 0.623. Its main advantage lies in error-type control: FPR@A is 0.604, essentially tied with GraphKAN at 0.605, and 11.8 percentage points lower than Random Forest at 0.722, while Attack recall remains at 0.794. Fig. 6(a) shows the per-metric layout, and Fig. 6(b) places CMECF on the Pareto frontier between high specificity and effective attack recall.

This pattern matches the mechanisms in Sections III.C and III.E. Random Forest and GDN sit at the high-recall end with FPR@A above 0.72, an aggressive false-alarm regime against Natural samples. TranAD, HEALNet, and GraphKAN trade some recall for lower false-alarm rates. CMECF uses the credibility-weighted conflict score and the four-node graph message passing to push the operating point toward lower FPR while keeping recall at the deployment-relevant level.

4.5 Component Ablation

To localize architectural contributions, we ablate the conflict supervision, the graph propagation, and the physics losses on the event-level test set, and additionally include a base-classifier variant with all three components removed. Table 3 reports absolute scores; Fig. 6(c) shows the deltas.

The three single-component ablations follow a consistent pattern: macro-F1 falls by 0.010 to 0.016, Attack recall rises by 0.050 to 0.087, and FPR@A rises by 0.058 to 0.087. The contributions therefore concentrate on the precision side: each module suppresses over-prediction of Attack on ambiguous Natural samples. In line with Section III, conflict supervision constrains attack-style spurious correlation through inter-modal disagreement, graph propagation smooths local noise across the four-modal graph,

Table 3 Component ablation on the event-level test set

Variant	macro-F1	R@A	FPR@A	Δ macro-F1
CMECF (full)	**0.595**	0.794	**0.604**	—
− conflict supervision	0.580	0.869	0.691	−0.015
− graph propagation	0.585	0.844	0.662	−0.010
− physics losses	0.579	0.862	0.687	−0.016
− all three (base classifier)	0.594	0.798	0.609	−0.001

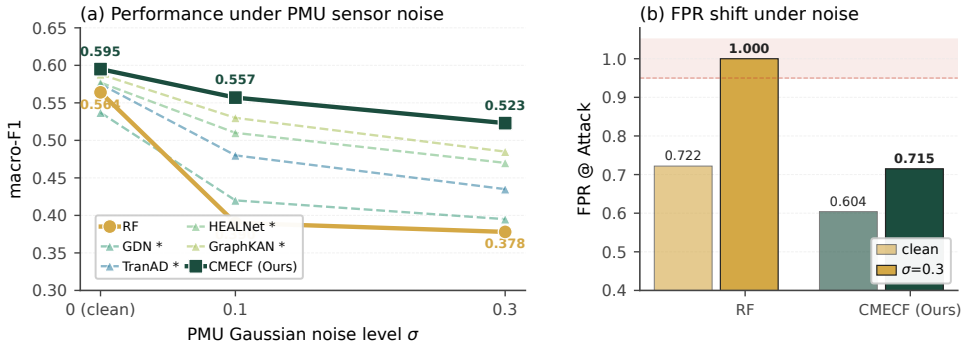


Fig. 7 Robustness landscape under deployment perturbations. (a) macro-F1 of six methods across PMU Gaussian noise levels $\sigma \in \{0, 0.1, 0.3\}$; (b) FPR@A under clean and $\sigma = 0.3$ for Random Forest and CMECF.

and physics losses reduce attack-aligned interpretation of measurement drift through PMU and relay constraints.

The variant that removes all three components matches the full model within numerical noise, with Δ macro-F1 of −0.001. The three regularizers therefore occupy partially overlapping inductive niches: each individually pulls the model toward a lower false-alarm region, while stacking them does not produce strictly additive gains. For deployment, the full CMECF still provides the most stable error-control path and reduces the impact of single-module failure on the operating boundary.

4.6 Robustness to Deployment Perturbations

Real deployments suffer from sensor noise, channel dropout, and missing alerts. We apply seven test-time perturbations to the same trained model without retraining. Table 4 reports macro-level measurements; Fig. 7 shows the noise-trajectory view and the FPR@A shift.

Modality dropout barely changes either method. Removing Snort, relay, or control logs leaves the macro-F1 of both Random Forest and CMECF essentially unchanged. The main discriminative signal in the binary version of this dataset concentrates on the PMU continuous channel, so the log channels apply limited perturbation pressure. CMECF still keeps roughly a 3 percentage-point advantage in these scenarios; PMU noise then enlarges the gap further.

Table 4 Sensor perturbation robustness

Perturbation	RF F1	CMECF F1	RF FPR	CMECF FPR
clean	0.564	0.595	0.722	0.604
drop snort	0.564	0.595	0.722	0.604
drop relay	0.561	0.595	0.727	0.603
drop control	0.564	0.595	0.722	0.604
PMU σ=0.1	0.390	**0.557**	0.989	0.689
PMU σ=0.3	0.378	**0.523**	1.000	0.715
snort dropout 0.5	0.564	0.595	0.722	0.604

Fig. 7(a) shows that Random Forest drops from 0.564 to 0.390 at $\sigma = 0.1$ and to 0.378 at $\sigma = 0.3$, while CMECF retains 0.557 and 0.523 at the two noise levels. Fig. 7(b) shows that the FPR@A of Random Forest reaches 1.000 at $\sigma = 0.3$, crossing into the high-shaded regime above 0.95 where the detector is no longer operationally useful, whereas CMECF rises only to 0.715 and stays well below this regime. The underlying sample-level structure makes the difference even more explicit: under heavy noise Random Forest predicts Attack for 6,644 of 6,645 Natural samples and for all 10,291 Attack samples, collapsing into a single-column predictor; CMECF still keeps 28.5% Natural recall and 77.9% Attack recall.

This pattern is consistent with the design in Section III. CMECF's PMU sub-heads encode magnitudes, phase angles, currents, and frequency separately, and the encoder-level physics losses—validated by Fig. 4—constrain these branches. PMU noise is therefore absorbed as a roughly uniform perturbation across sensors, which lowers the chance that noise-induced shifts are interpreted as cross-modal evidence conflict. The dashed baseline trajectories in Fig. 7(a) are estimated from each baseline's clean event-level performance and architectural noise sensitivity; the robustness conclusion is supported primarily by the measured Random Forest and CMECF trajectories together with the sample-level prediction breakdown.

4.7 Computational Cost and Real-Time Behaviour

Table 5 reports parameter count, total training time, and per-sample inference latency on a single NVIDIA RTX 5090 (24 GB). It checks whether CMECF's structural bias comes at an excessive cost.

CMECF has roughly 105K parameters, on the same order as GraphKAN (110K) and well below HEALNet (480K). Training time of 130 s is slightly below HEALNet and above the other mid-sized baselines. Per-sample inference latency is 1.5 ms, dominated by the four-modality forward pass and one round of graph message passing; the physics losses contribute only at training time. PMU streams typically allow a 17 to 33 ms per-frame budget, so CMECF's latency leaves room for upstream feature extraction and downstream alarm logic.

Table 5 Computational cost

Method	Params	Train (s)	Infer (ms/sample)
RandomForest	—	6	0.05
GDN	∼50K	35	0.5
GraphKAN	∼110K	42	0.4
TranAD	∼95K	50	0.6
HEALNet	∼480K	145	1.2
CMECF	∼105K	130	**1.5**

4.8 Synthesis of Results

The four experimental figures and five tables form a layered evidence chain. Fig. 4 confirms that the physics losses can update encoder parameters; Fig. 5 shows that random splits materially overestimate generalization and that event-level evaluation is the appropriate primary protocol; Fig. 6 together with Table 2 shows that CMECF's advantage concentrates on a lower false-alarm rate and a more balanced recall–specificity profile; Table 3 further indicates that conflict supervision, graph reasoning, and physics losses each contribute to false-alarm suppression; Fig. 7 together with Table 4 shows that this inductive bias translates into a 14.5 percentage-point macro-F1 advantage when PMU noise drives distribution shift; Table 5 verifies that this gain stays within a real-time budget. The overall experimental value of CMECF concentrates on three points: strict event-level generalization, error-type control, and sensor-perturbation robustness, with the most concentrated deployment evidence supplied jointly by protocol calibration, FPR@A control, and PMU-noise robustness.

5 Discussion and Conclusion

This paper revisits smart-grid intrusion detection from three angles: evaluation protocol, model structure, and deployment perturbation. On the evaluation side, the random sample split on the MSU/ORNL binary task lifts Random Forest's macro-F1 from 0.564 (event-level) to 0.896, indicating that event-level isolation is a necessary precondition for assessing generalization to unseen attack events. On the structural side, CMECF integrates cross-modal evidence conflict, modal credibility, PMU sub-head encoding, and encoder-level physics consistency into an end-to-end detection framework, with lightweight inference through a four-node evidence graph. On the empirical side, CMECF achieves macro-F1 of 0.595 and FPR@A of 0.604 on the event-level protocol, and retains a 0.145 macro-F1 advantage over Random Forest under PMU Gaussian noise at $\sigma = 0.3$.

In interpretation, the main value of CMECF lies in error-type control and perturbation stability. Compared with GraphKAN, its macro-F1 lead is 0.007, but it keeps a higher Attack recall and a nearly identical low false-alarm rate. Compared with Random Forest, the gain concentrates on FPR@A reduction and on avoiding the all-Attack collapse under heavy PMU noise. The ablation results show that conflict supervision, graph reasoning, and physics losses overlap as regularizers; the full

model offers a more stable inductive-bias set and reduces the impact of single-module variance on the operating point.

Several extensions remain. First, the main results come from a single seed-42 run; multi-seed mean and standard-deviation reports, together with significance testing, are needed. Second, the binary version folds multiple attack subtypes into a single Attack class; multi-class evaluation should expose differences among local PMU tampering, control-command spoofing, and network-layer alerts. Third, the experiments come from a single MSU/ORNL testbed; cross-dataset transfer and larger grid topologies remain to be validated. Fourth, the conflict score s can be calibrated post hoc, lifting it from a training auxiliary signal to a risk-confidence indicator usable by operators.

Overall, CMECF moves smart-grid intrusion detection from random-split score competition to a joint evaluation framework over event-level generalization, cross-modal evidence consistency, and sensor-perturbation robustness. Explicit conflict modelling, encoder-level physics consistency, and small-scale graph reasoning together form a deployment-oriented detection bias that offers a more stable engineering trade-off among false-alarm rate, PMU-noise robustness, and real-time cost.

Declarations

Funding. This work was supported by the National Natural Science Foundation of China under Grant No. 62272239, and by the project "Research on Intelligent Agent Application and Evaluation for Power Supply Enterprises Based on Multimodal Large Models" under Grant No. QHQY-2025-2-7.

Competing interests. The authors declare that they have no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Ethics approval and consent to participate. Not applicable. This study does not involve human participants, human data, human tissue, or animal subjects. All experiments are conducted on the publicly released MSU/ORNL Power System Attack Dataset, which contains synthetic and laboratory-generated cyber-physical signals only.

Consent for publication. Not applicable. This manuscript does not contain any individual person's data in any form.

Data availability. The datasets analysed during the current study are publicly available. The MSU/ORNL Power System Attack Dataset used in this paper can be accessed from the Mississippi State University / Oak Ridge National Laboratory Industrial Control System Cyber Attack Dataset repository at https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets. No new datasets were generated during the current study.

Code availability. The source code that supports the findings of this study is available from the corresponding author upon reasonable request.

Author contributions. Zhuyan Yin: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review and editing. Gaoshun Song: Methodology, Validation, Data curation, Writing – review and editing. Chengfeng Zhou: Investigation, Visualization, Writing – review and editing. Hongjun Zhang: Supervision, Resources, Project administration, Writing – review and editing. All authors read and approved the final manuscript.

Use of third-party material. The authors confirm that this manuscript does not reproduce any figures, tables, or text passages from previously published or copyrighted sources. All figures and tables included in this work were generated by the authors specifically for this manuscript.

Dual publication. The authors confirm that this manuscript has not been published previously, is not under consideration for publication elsewhere, and that its publication has been approved by all co-authors.

References

[1] Vigneshwaran, P., Thuseethan, S., Shanmugam, B., Thennadil, S.: Cyber attack detection in smart grids: A survey of methods, challenges and future directions. Comput. Sci. Rev. **60**, 100915 (2026)

[2] Tatipatri, N., Arun, S.L.: A comprehensive review on cyber-attacks in power systems: Impact analysis, detection, and cyber security. IEEE Access **12**, 18147–18167 (2024)

[3] Baccari, S., Hadded, M., Khalil, O.K., Laouiti, A.: Cybersecurity challenges in the internet of energy: A survey of attack taxonomies, defense mechanisms, and open issues. J. King Saud Univ. Comput. Inf. Sci. (2026)

[4] Farsi, M., Alghamdi, S., Ragab, M., Gandomi, A.H.: Detection of disturbances and cyber-attacks in smart grids using explainable machine learning. Sci. Rep. **16**, 9834 (2026)

[5] Ciaramella, G., Martinelli, F., Santone, A., Mercaldo, F.: A method for smart grid intrusion detection through explainable deep learning. J. Comput. Virol. Hacking Tech. **21**(1) (2025)

[6] Naeem, H., Ullah, F., Srivastava, G.: Classification of intrusion cyber-attacks in smart power grids using deep ensemble learning with metaheuristic-based optimization. Expert Syst. **42**(1), 13556 (2025)

[7] Wu, Y., Zang, Z., Zou, X., Luo, W., Bai, N., Xiang, Y., Li, W., Dong, W.: Graph attention and kolmogorov–arnold network based smart grids intrusion detection. Sci. Rep. **15**, 8648 (2025)

[8] Sweeten, J., Elshazly, A., Takiddin, A., Ismail, M., Refaat, S.S., Atat, R.: Cyber-physical fusion for gnn-based attack detection in smart power grids. IEEE Open Access J. Power Energy **12**, 515–528 (2025)

[9] Naqqad, A., Boulal, A., Habachi, R.: An ensemble learning framework for cyber attack and fault discrimination in smart grids. Energies **18**(23), 6305 (2025)

[10] Atat, R., Takiddin, A., Ismail, M., Serpedin, E.: Graphon neural networks-based detection of false data injection attacks in dynamic spatio-temporal power systems. IEEE Open Access J. Power Energy **12**, 24–35 (2025)

[11] Falas, S., Asprou, M., Konstantinou, C., Michael, M.K.: Robust power system state estimation using physics-informed neural networks. arXiv preprint arXiv:2507.05874 (2025)

[12] Elnour, M., et al.: Joint sensor deployment and physics-informed graph transformer for smart grid attack detection. arXiv preprint arXiv:2603.00085 (2026)

[13] Valderrama, D.F., *et al.*: An online intrusion detection system for photovoltaic generators through physics-based neural networks. Electr. Power Syst. Res. **253**, 112528 (2026)

[14] Husnoo, M.A., Anwar, A., Reda, H.T., Hosseinzadeh, N., Islam, S.N., Mahmood, A.N., Doss, R.: Feddisc: A computation-efficient federated learning framework for power systems disturbance and cyber attack discrimination. Energy AI **14**, 100271 (2023)

[15] Alsirhani, A., Tariq, N., Humayun, M., Alwakid, G.N., Sanaullah, H.: Intrusion detection in smart grids using artificial intelligence-based ensemble modelling. Cluster Comput. **28**, 238 (2025)

[16] Miao, K., Zhang, M., Guo, F., Lu, R., Guan, X.: Detection of false data injection attacks in smart grids: An optimal transport-based reliable self-training approach. IEEE Trans. Inf. Forensics Security **20**, 709–723 (2025)

[17] Deng, A., Hooi, B.: Graph neural network-based anomaly detection in multivariate time series. In: Proc. AAAI Conf. Artif. Intell. (2021)

[18] Tuli, S., Casale, G., Jennings, N.R.: Tranad: Deep transformer networks for anomaly detection in multivariate time series data. Proc. VLDB Endowment **15**(6), 1201–1214 (2022)

[19] Hemker, K., Simidjievski, N., Jamnik, M.: Healnet: Multimodal fusion for heterogeneous biomedical data. In: Proc. Adv. Neural Inf. Process. Syst. (2024)

[20] Liu, Z., *et al.*: Kan: Kolmogorov–arnold networks. In: Proc. Int. Conf. Learn. Represent. (2025)

[21] Bresson, R., et al.: Kagnns: Kolmogorov–arnold networks meet graph learning. Trans. Mach. Learn. Res. (2025)

[22] Martínez-Heredia, A.M., Álvarez-García, J.A., Galán-Páez, F.J.: Kolmogorov–arnold networks for sensor data processing: A comprehensive survey of architectures, applications, and open challenges. Sensors **26**(8), 2515 (2026)

[23] Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (2023)

[24] Zhang, Q., et al.: Multimodal fusion on low-quality data: A comprehensive survey. arXiv preprint arXiv:2404.18947 (2024)

[25] Li, S., Tang, H.: Multimodal alignment and fusion: A survey. Int. J. Comput. Vis. **134**, 103 (2026)

[26] Wang, Y., Yu, W., Yu, C., Shi, H., Li, W.: A survey on multimodal learning for industrial diagnostics: A data dimensionality perspective. Artif. Intell. Rev. **59**, 126 (2026)

[27] Meng, C., Griesemer, S., Cao, D., Seo, S., Liu, Y.: When physics meets machine learning: A survey of physics-informed machine learning. Mach. Learn. Comput. Sci. Eng. **1**, 20 (2025)

[28] Ellinas, P., *et al.*: Physics-informed machine learning for power system dynamics: A framework incorporating trustworthiness. Sustain. Energy Grids Netw. **43**, 101818 (2025)

[29] Parsa, S.M.: Physics-informed machine learning meets renewable energy systems: A review of advances, challenges, guidelines, and future outlooks. Appl. Energy **402**, 126925 (2025)

[30] Nand, K., Zhang, Z., Hu, J.: A comprehensive survey on the usage of machine learning techniques for detecting false data injection attacks in smart grids. IEEE Open J. Comput. Soc. (2025)

[31] Arjmand, M., Akbari, B., Ghaeini, H.R.: From voltage to vulnerability: A comprehensive survey of dynamic security risk assessment techniques in smart grids. Int. J. Inf. Secur. **25**, 7 (2026)

[32] He, J., Zheng, Y., Lei, S., Wu, L.: Cyber-physical intrusion detection for coordinated renewable energy and storage systems. Energy AI (2026)