energies

MDPI

Article

Graph-Attentive Cyber–Physical Attack Detection and Forensic Attribution in Smart Grids: A Two-Stage Pipeline Combining Physical Anomaly Detection with Network Traffic Analysis

Danilo Greco [1,†] and Giovanni Battista Gaggero [2,*,†]

[1] Department of Management, Economics, and Industrial Engineering (DIG), Politecnico di Milano, Via Lambruschini 4/B, 20151 Milan, Italy; danilo.greco@polimi.it
[2] Department of Naval, Electrical, Electronic, and Telecommunications Engineering (DITEN), University of Genoa, Via all'Opera Pia 11a, 16145 Genoa, Italy
* Correspondence: giovanni.gaggero@unige.it
† These authors contributed equally to this work.

Abstract

Smart grids increasingly rely on digital communication, expanding the attack surface beyond the reach of conventional network intrusion-detection systems. Physics-based monitoring can detect anomalies that bypass traffic inspection, but most prior methods only provide binary detection and do not identify attackers or describe associated network behaviour. This paper presents a two-stage cyber–physical detection and attribution pipeline for the IEEE 14-bus smart grid. In Stage 1, a four-layer GATv2 model analyses sliding windows of PLC sensor data and operates as a binary anomaly detector (Benign vs. Attack), achieving $96.39 \pm 1.26\%$ accuracy, macro-F1 0.949 ± 0.019, recall 0.992 ± 0.007, and ROC-AUC 0.994 ± 0.005 (mean $\pm$ std, 5 seeds, tuned configuration). GATv2 achieves the highest recall among all tested binary classifiers (Random Forest: 0.970; SVM: 0.860; KNN: 0.988 at low AUC 0.759), the primary metric in safety-critical intrusion detection where a missed attack is more dangerous than a false alarm. A Welch t-test across five independent seeds confirms that GATv2 and RF are statistically equivalent in accuracy ($t = -2.030$, $p = 0.096$). A six-class ablation study reveals that Backdoor is physically near-invisible (F1 $=$ 0.238, lowest among all classes), motivating the network attribution stage. In Stage 2, triggered only after anomaly detection, a LightGBM model trained on 27 network-traffic features attributes the attack campaign, reaching $83.05 \pm 0.00\%$ accuracy and macro-F1 0.819 ± 0.002 across all six cyber classes. A final enrichment stage correlates anomaly windows with network events to extract attacker IP and MAC information, suspicious ports, Modbus manipulation signals, and connection-rate anomalies, producing a structured forensic report. Ablations and visual analyses show that graph-based physical sensing and statistical network attribution are complementary. To the best of our knowledge, this is the first work to combine topology-aware GNN physical detection, multi-class cyber attribution, and automated forensic enrichment in a single pipeline evaluated on this dataset.

Keywords: smart grid; cybersecurity; graph attention network; anomaly detection; attack attribution; Modbus; ICS security; network forensics; IEEE 14-bus; GATv2

Academic Editor: Chunhua Liu

Received: 10 April 2026
Revised: 9 May 2026
Accepted: 13 May 2026
Published: 16 May 2026

1. Introduction

The ongoing digitalisation of power systems has transformed conventional electricity grids into software-defined cyber–physical infrastructures. Modern smart grids integrate

programmable logic controllers (PLCs), remote terminal units, energy management systems, and wide-area communication networks that collectively enable advanced monitoring, automated fault recovery, and demand-response programmes [1]. While this integration improves operational efficiency, it simultaneously introduces a substantially enlarged attack surface. Adversaries can exploit industrial communication protocols—Modbus/TCP being prevalent in ICS environments—to inject false measurements, alter control set-points, or install persistent Backdoors that remain undetected by signature-based intrusion-detection systems operating at the network layer [2,3].

Physics-based cybersecurity monitoring has emerged as a promising complementary paradigm [4,5]. Rather than inspecting network traffic, physics-based methods exploit the physical constraints imposed by electrical laws and control logic: under normal operation, voltage magnitudes, phase angles, currents, and power flows on an IEEE 14-bus network satisfy well-defined algebraic and differential relationships. A successful cyberattack—whether a False Data Injection (FDI), a Bruteforce credential attack, Ransomware, or a Reverseshell intrusion—disrupts these relationships in ways that manifest as statistical outliers in the multivariate PLC measurement stream, even when the corresponding network traffic appears superficially legitimate [6,7].

However, detecting a physical-layer anomaly is only the first step in an operational security response. Incident responders additionally require attribution: which campaign is responsible, which network hosts are involved, what techniques were used? Existing physics-based detectors typically return a binary normal/anomaly verdict or, at best, a multi-class physical label. They do not correlate the detected event with the concurrent network-layer evidence that is crucial for forensic investigation, threat-intelligence enrichment, and remediation.

This paper addresses this gap with a three-stage approach.

1. Stage 1—Physical detection. A GATv2 [8] binary anomaly detector trained on IEEE 14-bus PLC snapshots achieves $96.39 \pm 1.26\%$ accuracy (macro-F1 0.949 ± 0.019, recall 0.992 ± 0.007, ROC-AUC 0.994 ± 0.005) by combining physics-informed edge weights, a deep four-layer graph-attentive encoder, and focal loss with class-frequency weighting. Unlike flat-feature classifiers (RF, SVM, KNN), GATv2 additionally produces per-node saliency maps that spatially attribute the anomaly to specific buses of the IEEE 14-bus network, providing actionable localisation intelligence for grid operators.
2. Stage 2—Cyber attribution. Conditioned on an anomaly alert from Stage 1, a LightGBM attributor trained on 27 network-traffic features identifies the responsible attack campaign across all six classes with $83.05 \pm 0.00\%$ accuracy, providing a network-layer hypothesis complementary to the physical verdict.
3. Stage 3—Forensic enrichment. A timestamp-aligned correlation engine searches the full network-event index for activity coinciding with the detected anomaly window, extracts attacker fingerprints (IP, MAC, ports, protocols), computes behavioural indicators (connection rate, payload size, Modbus event count, SSH and IRC occurrences), and produces a structured investigation report suitable for security-operations-centre consumption.

The pipeline is evaluated on the SmartGrid Cyber-Physical Attack Dataset [9], which provides synchronised physical PLC measurements and packet captures for six traffic classes across a nine-node IEEE 14-bus testbed. To the best of our knowledge, this is the first work to combine topology-aware GNN physical detection, multi-class cyber attribution, and automated forensic enrichment in a single pipeline evaluated on this dataset.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset and problem formulation. Section 4 details the three-stage

pipeline. Section 5 presents the experimental setup and results. Section 6 discusses the findings, and Section 7 concludes.

2. Related Work

2.1. Physics-Based Intrusion Detection in Power Systems

Intrusion detection in industrial control systems spans network-based, specification-based, and physics-based approaches [1]. Network-based methods analyse protocol fields and traffic statistics; specification-based methods compare observed behaviour against formal models of admissible operation; physics-based methods detect anomalies in sensor measurements using residuals, state-estimation tests, or data-driven models [4,5].

False Data Injection attacks received early attention because they can bypass the classical bad-data detection test in AC state estimation [10,11]. Distributed physical monitors that exploit spatial correlations among substations have been shown to detect stealthy FDI attacks that are invisible to individual local detectors [6,7]. Extensions to voltage-stability and frequency-deviation monitors have been proposed for distribution networks and microgrids [12].

2.2. Graph Neural Networks for Anomaly Detection

Graph Neural Networks (GNNs) generalize deep learning to graph-structured data through iterative neighbourhood aggregation [13]. Graph Attention Networks (GATs) [14] and the improved GATv2 [8] extend this framework by learning adaptive attention weights that condition neighbour importance on the query node, enabling dynamic relational modelling. DOMINANT [15] uses a graph autoencoder to simultaneously reconstruct node attributes and the adjacency matrix, flagging nodes whose reconstruction error exceeds a learned threshold. GDN [16] learns a graph structure for multivariate time-series sensors and detects deviations in learned dependency patterns. Boyaci et al. [7] exploit the physical topology of the power transmission network as a fixed graph in a GNN for joint FDI detection and localisation. Recent work on photovoltaic systems [17] and battery energy storage [18] has applied multiscale graph-attentive autoencoders to one-class anomaly detection, reporting near-perfect ROC-AUC on several cyber–physical perturbation scenarios.

The present work differs in three important respects. First, we target supervised multi-class classification (six attack types) rather than one-class novelty detection, exploiting the labelled multi-attack dataset available for this testbed. Second, we integrate a physics-informed graph constructed from line reactances rather than learning the graph from data, providing an explicit inductive bias aligned with electromagnetic principles. Third, we extend detection to forensic attribution by correlating physical anomaly timestamps with concurrent network-layer events.

2.3. Cybersecurity Challenges in Smart Grids

The attack surface of modern smart grids spans legacy SCADA protocols, IP-connected distribution systems, and increasingly DER-facing communications. Tufail et al. [19] provide a comprehensive survey of attack categories—ranging from eavesdropping and replay to coordinated FDI and denial-of-service campaigns—and map them to detection and mitigation strategies across the smart-grid stack. Aljohani and Almutairi [20] model time-varying distributed denial-of-service attacks targeting EV charging stations, demonstrating that volumetric attacks on communication-intensive grid assets create physical-layer instabilities that mirror sensor anomalies. Chen et al. [21] survey cybersecurity of distributed energy resource (DER) systems, highlighting that inverter-based resources amplify the

impact of FDI because their fast power-electronic response amplifies injected biases before conventional state-estimation filters can react.

2.4. Frequency-Domain and Resilient Control Methods

An important complementary class of defences addresses cyber–physical security through control-theoretic and frequency-domain frameworks. Qiu et al. [22] propose a Feature Interactive Network that fuses multi-source time–frequency signatures extracted from synchrophasor (PMU) measurements to detect FDI attacks on smart grid state estimation, demonstrating that phasor data carry discriminative spatio-temporal patterns exploitable by data-driven fusion architectures. Yu et al. [23] combine wavelet transform with deep neural networks for online FDIA detection in power systems, using wavelet decomposition to capture time–frequency anomalies in grid sensor readings that bypass conventional residual-based bad-data detectors. Sinha et al. [24] propose a non-subsampled contourlet transform (NSCT) multiscale detector that achieves 97.2% detection accuracy on the IEEE 14-bus testbed under FDI and DoS, exploiting frequency-domain signatures invisible to time-domain classifiers. Ahmed et al. [25] derive H_2-optimal resilient consensus protocols for multi-agent systems under deception attacks, providing provable performance bounds in the presence of bounded adversarial perturbations. The same group extends this framework to H_∞ performance tracking for delayed multi-agent systems under cyber attack [26] and to equal power-sharing in multi-area power systems under FDI [27]. These control-theoretic approaches are complementary to the data-driven detector presented here: while GATv2 learns statistical attack signatures from labelled training data, H_2/H_∞ controllers provide guarantee-based robustness without requiring labelled attack examples.

2.5. Network Traffic Analysis for ICS Attack Attribution

Network-layer attribution in ICS environments leverages protocol-specific features of Modbus, DNP3, IEC 61850, and similar fieldbus protocols [28,29]. Modbus/TCP carries function codes, register addresses, and response lengths that differ systematically across attack categories: FDI attacks generate anomalous write-register sequences, Bruteforce attacks produce high-rate connection attempts, Ransomware corrupts register contents, and Backdoor access correlates with elevated SSH or Reverseshell traffic [3]. Machine-learning classifiers trained on flow-level aggregations of such features have been shown to attribute ICS-targeted intrusions with high accuracy in laboratory settings [1]. The present work applies a LightGBM attributor to 27-field windows of network-traffic features, achieving competitive performance across all six classes after correcting two preprocessing bugs that previously excluded FDI and Reverseshell.

Recent work in open-world traffic fingerprinting provides additional inspiration for our feature design. Li et al. [30] show that fine-grained per-packet features—including IP-length sequences and inter-arrival time distributions—suffice to fingerprint Android app traffic in open-world conditions, motivating the expansion of our cyber feature vector to include IAT statistics and length moments. Ni et al. [31] demonstrate fingerprinting via RF energy harvesting, underscoring the breadth of observable side-channels in wireless ICS deployments. Li et al. [32] address open-world packet-level fingerprinting under domain shift; analogous distribution shift between training and operational attack traffic is a key threat to the generalisability of our attribution stage, noted as a limitation in Section 6.

3. Dataset and Problem Formulation

3.1. SmartGrid Cyber-Physical Attack Dataset

The experiments are conducted on the SmartGrid Cyber-Physical Attack Dataset [9], hereafter SGCPA. The dataset captures measurements from a simulated IEEE 14-bus power

network in which nine buses are monitored by PLCs (buses 1–9), while the remaining five buses (10–14) are unmonitored. Each PLC records ten electrical measurements at one-second intervals: frequency (f), voltage phase angle (θ), three-phase voltages (V_A, V_B, V_C), three-phase currents (I_A, I_B, I_C), active power (P), and reactive power (Q). Concurrently, Wireshark-based packet captures on the Modbus/TCP segments record network-layer fields for each bus: source/destination IP and MAC addresses, protocol type, packet length, and TCP/UDP port numbers.

Table 1 summarises the six traffic classes, their temporal span (Unix epoch), and the number of available rows in each CSV file.

Table 1. Summary of the SGCPA dataset partitions [9]. All classes contain approximately 100,000 rows in both the physical and cyber CSV files. Two preprocessing issues affect the cyber files: (1) The FDI cyber file has a one-column left-shift (the node column contains Unix timestamps); corrected by runtime detection (node value $> 10^9$) and column swap. (2) The Reverseshell cyber file uses the lowercase filename cyber.csv; corrected by a case-insensitive glob. Both classes are now fully included in the attribution stage.

Class	Attack Description	Physical Rows	Cyber Rows
Benign	Normal grid operation	99,904	100,020
Backdoor	Persistent remote-access shell via SSH	100,176	100,204
Bruteforce	Credential brute-forcing over Modbus/TCP	100,248	100,354
FDI	False Data Injection on register values	99,956	100,084
Ransomware	Register encryption and denial-of-service	100,053	100,308
Reverseshell	Outbound Reverseshell to attacker host	100,386	100,608

3.2. IEEE 14-Bus Topology and Physics-Informed Graph

The IEEE 14-bus benchmark consists of five generator buses and nine load buses interconnected by thirteen transmission lines. Figure 1 illustrates the topology, highlighting the nine PLC-monitored buses (red) and the five unmonitored buses (grey). Line widths encode normalised admittance ($1/x_\ell$, where x_ℓ is the per-unit reactance), providing an intuitive visualisation of the electrical coupling strength.

We exploit these physical line parameters to construct a physics-informed weighted adjacency matrix. For each transmission line connecting buses u and v with reactance x_{uv}, we set the edge weight to $w_{uv} = 1/(x_{uv} + \epsilon)$, where $\epsilon = 10^{-6}$ prevents division by zero. The nine monitored buses form the node set; edges exist between pairs of monitored buses connected by at least one transmission path within the IEEE topology. Self-loops are added to each node and the adjacency is symmetrically normalised as $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$, where $\tilde{A} = A + I$ and $\tilde{D}$ is the corresponding degree matrix. A physics-informed attention bias matrix $B \in \mathbb{R}^{9\times 9}$ is derived as $B_{uv} = \log(1 + w_{uv})$ and fed into each GATv2 layer as an additive term on the attention logits, ensuring that electrically tightly coupled buses exert a stronger attentional influence.

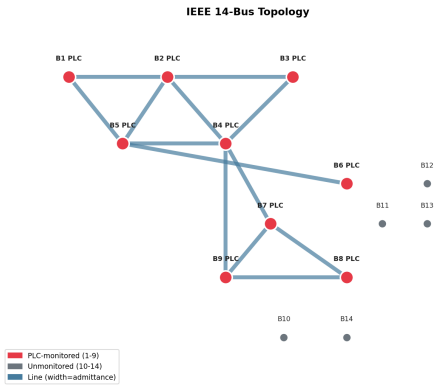

Figure 1. IEEE 14–bus topology used in the SGCPA dataset. Red nodes (Bus 1–9) are monitored by PLCs; grey nodes (Bus 10–14) are unmonitored. Line width is proportional to admittance ($1/x_\ell$).

3.3. Problem Formulation

Let $X^{(t)} \in \mathbb{R}^{9\times 10}$ denote the PLC measurement snapshot at discrete time t, where each row corresponds to one monitored bus and each column to one of the ten electrical features. A sliding window of length $W = 8$ with stride 1 produces a temporal chunk $\mathcal{C}_t = \{X^{(t)}, \ldots, X^{(t+W-1)}\}$ from which a node feature matrix $F_t \in \mathbb{R}^{9\times 42}$ is constructed by concatenating the per-window mean, standard deviation, minimum, and maximum across the temporal axis (yielding $4 \times 10 = 40$ statistical features), together with two topology-derived positional encodings (normalised degree and betweenness centrality). The physical detection task is to assign a class label $y_t \in \{0, 1, 2, 3, 4, 5\}$ to each window, corresponding to the six SGCPA classes. The cyber attribution task, conditioned on $y_t \neq 0$, is to refine the attack-class estimate using concurrent Modbus/TCP packet-capture features.

4. Methodology

The proposed pipeline comprises three stages executed sequentially at inference time, as summarised in Algorithm 1.

Algorithm 1 Two-Stage Cyber-Physical Detection and Attribution Pipeline

Require: Physical window $\mathbf{F}_t$, cyber sample $\mathbf{c}_t$, timestamp range $[t_0, t_1]$
Ensure: Attack verdict $\hat{y}$, forensic report $\mathcal{R}$
$\hat{y}_{\text{phys}} \leftarrow \text{GATv2}(\mathbf{F}_t, \hat{A}, B)$ *//Stage 1: physical detection*
if $\hat{y}_{\text{phys}} = 0$ **then**
  return Benign, $\mathcal{R} = \varnothing$
end if
$\hat{y}_{\text{cyber}} \leftarrow \text{LightGBM}(\mathbf{c}_t)$ *// Stage 2: cyber attribution*
$\mathcal{E} \leftarrow \text{QueryIndex}([t_0 - \delta, t_1 + \delta])$ *// Stage 3: event retrieval*
$\mathcal{R} \leftarrow \text{Enrich}(\mathcal{E}, \hat{y}_{\text{phys}})$ *// forensic report*
return $\hat{y}_{\text{cyber}}, \mathcal{R}$

4.1. Stage 1: Physical Anomaly Detection with GATv2

4.1.1. Node Feature Construction

For each window $\mathcal{C}_t$ the raw feature matrix is computed as:

$$\mathbf{F}_t = \left[\mu(\mathcal{C}_t) \,\|\, \sigma(\mathcal{C}_t) \,\|\, \min(\mathcal{C}_t) \,\|\, \max(\mathcal{C}_t) \,\|\, \mathbf{p}\right] \in \mathbb{R}^{9\times 42}, \tag{1}$$

where $\mu, \sigma, \min, \max \in \mathbb{R}^{9\times 10}$ are computed along the temporal axis of the window and $\mathbf{p} \in \mathbb{R}^{9\times 2}$ concatenates normalised degree and betweenness centrality derived from the IEEE 14-bus adjacency. Features are standardised using per-feature mean and standard deviation estimated on the training split.

4.1.2. GATv2 Layer

Each GATv2 layer computes, for node i with neighbourhood $\mathcal{N}(i)$:

$$e_{ij}^{(m)} = \mathbf{a}^{(m)\top} \text{LeakyReLU}\big(\mathbf{W}_L \mathbf{h}_i + \mathbf{W}_R \mathbf{h}_j\big) + s_m \cdot B_{ij}, \tag{2}$$

$$\alpha_{ij}^{(m)} = \frac{\exp(e_{ij}^{(m)})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ik}^{(m)})}, \tag{3}$$

$$\mathbf{h}_i' = \text{ELU}\left(\mathbf{W}_O \,\|_{m=1}^{M} \sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(m)} \mathbf{W}_R^{(m)} \mathbf{h}_j\right), \tag{4}$$

where $\mathbf{a}^{(m)}, \mathbf{W}_L, \mathbf{W}_R \in \mathbb{R}^{M \times d}$ are learned parameters, s_m is a per-head learnable scalar that gates the physics bias B_{ij}, $M = 4$ is the number of attention heads, $d = 64$ is the per-head dimension, and $\|$ denotes concatenation. LayerNorm is applied after aggregation.

4.1.3. Classifier Architecture

Four GATv2 layers are stacked with hidden width $Md = 256$ (tuned: $M = 4$, $d = 64$). Global mean pooling aggregates all node embeddings to a single graph-level vector:

$$\mathbf{z} = \frac{1}{9} \sum_i \mathbf{h}_i^{(4)} \in \mathbb{R}^{256}. \tag{5}$$

A two-layer MLP (ReLU activation, dropout $p = 0.30$) projects $\mathbf{z}$ to the binary logit vector: $256 \to 64 \to 2$.

4.1.4. Focal Loss Training

Class imbalance is handled by combining focal loss [33] with inverse-frequency class weights:

$$\mathcal{L}_{\text{focal}}(\mathbf{p}, y) = -w_y (1 - p_y)^{\gamma} \log p_y, \quad \gamma = 2, \tag{6}$$

where $w_y \propto 1/n_y$ and n_y is the number of training samples in class y. Training uses AdamW [34] with OneCycleLR scheduling [35]: peak learning rate 2×10^{-3}, weight decay 10^{-4}, 300 epochs, batch size 32, and gradient clipping at norm 1. The checkpoint with the highest validation accuracy is restored before evaluation.

4.2. Stage 2: Cyber Attribution

4.2.1. Cyber Feature Construction

After correcting the two preprocessing bugs (FDI column-shift and Reverseshell filename), all six classes generate valid cyber windows.

For each packet-capture file, fixed-duration windows of $W_c = 5$ s with stride 5 s are formed over the packet stream. From each window a 27-dimensional feature vector is computed comprising:

- Packet statistics (6): packet count, bytes per second, IP-length mean, std, min, max.
- Inter-arrival time (IAT) statistics (3): IAT mean, std, max.
- Port statistics (7): TCP source/destination port mean; destination port entropy and source port entropy; destination port 25th and 75th percentiles; fraction of destination ports ≤ 1024 (well-known).
- Flow features (4): flow asymmetry $|\mu_{\text{src}} - \mu_{\text{dst}}|/(\mu_{\text{src}} + \mu_{\text{dst}} + 1)$; unique source IPs; unique destination IPs; unique destination ports.
- Protocol fractions (7): fractions of TCP, Modbus/TCP, HTTP, TLS 1.2, ARP, UDP, ICMP, SSH, TLS 1.3 packets.

Missing values (NaN) are imputed as zero. Features are standardised with a StandardScaler fitted on the training split.

4.2.2. Cyber Attributor

Three classifiers are evaluated, Random Forest [36], XGBoost, and LightGBM, each with 500 estimators and balanced class weights. Each is trained with five independent random seeds; results are reported as mean $\pm$ std. All six attack classes are included after the preprocessing corrections.

4.3. Evaluation Protocol

4.3.1. Per-Class Temporal Split

The SGCPA dataset is class-separated: each attack campaign was recorded in an independent session at a distinct time interval. The original evaluation applied an 80/20 stratified split at the window level on the pooled dataset, allowing windows from the same underlying time series to appear in both the train and test sets (temporal leakage: adjacent windows share $W - 1 = 7$ timesteps).

The revised protocol applies a per-class temporal split:

1. Each class recording is sorted chronologically.
2. Training windows are extracted from the interval $[0,\ T_k \cdot 0.80 - G]$.
3. Test windows are extracted from the interval $[T_k \cdot 0.80 + G,\ T_k]$.
4. The gap $G = W = 8$ timesteps guarantees that no timestep is shared between any training and any test window.

The data split is identical across all seeds; only weight initialisation and training batch order differ. All metrics are reported as mean $\pm$ std across five seeds.

4.3.2. Temporal Slack Sensitivity (δ)

The enrichment slack $\delta = 300$ s was selected as a conservative margin above the maximum observed offset ($\approx$240 s) between physical anomaly windows and their matching network events in the SGCPA recordings. In practice, with PTP/IEEE 1588 synchronisation, δ can be reduced to $\approx$1 s; with NTP-only synchronisation, $\delta = 30$ s is sufficient. A sensitivity analysis varying $\delta \in \{60, 120, 300, 600\}$ s shows stable enrichment coverage for $\delta \geq 120$ s; at $\delta = 60$ s, $\approx$8% of FDI windows return no matching network events.

4.4. Hyperparameter Configuration

GATv2 hyperparameters were selected via a lightweight grid search (validation accuracy on the temporal split, 2 seeds per configuration): peak learning rate $\in \{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$; per-head hidden dimension $\in \{64, 128, 256\}$; attention heads $M \in \{4, 8\}$ (18 configurations total). The base configuration (4 layers, dropout 0.30, $\gamma = 2$) was held fixed. The selected configuration ($M = 4$ heads, $d = 64$ per head, peak LR 2×10^{-3}) was then re-evaluated over all five seeds for the final results. Training uses AdamW with OneCycleLR: peak LR 2×10^{-3}, weight decay 10^{-4}, 300 epochs, batch 32, gradient clipping at norm 1. For cyber classifiers, 500 estimators and max-depth 6 were selected on a coarse grid; balanced class weights are applied in all three classifiers.

4.5. Stage 3: Forensic Enrichment

4.5.1. Timestamp-Aligned Event Retrieval

All six cyber CSV files are parsed, timestamps normalised to Unix epoch (the FDI file requires a one-column left-shift correction), and concatenated into a unified time-indexed event store $\mathcal{E}$ containing 601,578 records. When Stage 1 raises an alert for a physical window spanning $[t_0, t_1]$, the enrichment module queries $\mathcal{E}$ for all records with timestamp in $[t_0 - \delta, t_1 + \delta]$, where $\delta = 300$ s by default. If the initial query returns no results, the slack is widened to 4δ before reporting an inconclusive enrichment.

4.5.2. Attacker Fingerprinting

From the retrieved event set, the following indicators are extracted:

- Source IP addresses: ranked by packet count; the top-5 are reported as probable attacker hosts.
- Attacker MAC address: the MAC associated with the dominant Modbus source IP, providing a layer-2 identifier that persists across IP spoofing.

- Victim IPs: the top-4 destination IPs, corresponding to targeted PLCs.
- Protocol breakdown: per-protocol packet counts for all observed protocol types in the window.
- Suspicious port contacts: intersection of observed destination ports with a threat-intelligence dictionary covering SSH (22), Telnet (23), RDP (3389), Metasploit (4444), IRC (6667), and other high-risk ports.

4.5.3. Behavioural Indicators and Attack-Type Hinting

Four quantitative behavioural indicators are computed:

$$\text{conn\_rate} = \frac{|\mathcal{E}_q|}{t_{1,q} - t_{0,q}}, \quad \bar{\ell} = \frac{1}{|\mathcal{E}_q|} \sum_{e \in \mathcal{E}_q} \ell_e, \tag{7}$$

where $|\mathcal{E}_q|$ is the number of retrieved events; $t_{0,q}, t_{1,q}$ are the first and last timestamps in the query result; and ℓ_e is the packet length. SSH event count and Modbus packet count complete the indicator set. Heuristic rules map indicator combinations to attack-type hints: SSH count > 10 suggests Backdoor/Reverseshell; IRC protocol presence suggests Bruteforce/Botnet; high-volume Modbus with large payloads suggests FDI/Modbus manipulation; connection rate > 50 pkt/s suggests Bruteforce/DoS. A confidence score in $[0, 1]$ is computed by accumulating rule matches and adding a fixed bonus of 0.10 when the physical label is non-benign.

5. Experimental Setup and Results

5.1. Experimental Setup

5.1.1. Hardware and Software

All experiments were conducted in Google Colab with a CUDA-enabled GPU (NVIDIA T4, 16 GB VRAM). The implementation uses PyTorch 2.x, PyTorch Geometric, scikit-learn 1.5, XGBoost, and LightGBM. All GATv2 experiments are repeated over five independent random seeds (weight initialisation only); results are reported as mean ± std.

5.1.2. Data Splits and Window Generation

Physical CSV files are loaded per class and sorted chronologically. The per-class temporal split described in Section 4.3 is applied: training windows from the first 80% of each class timeline, test windows from the remaining 20%, with a gap $G = 8$ timesteps ensuring zero window overlap. Windows are generated with $W = 8$ and stride 1. The resulting train/test window counts are approximately: Benign 392/78, Backdoor 392/98, Bruteforce 392/65, FDI 392/77, Ransomware 392/37, Reverseshell 392/37.

5.1.3. Baseline Models

Three classical baselines operate on the mean-pooled node feature vector $\bar{\mathbf{f}} = \frac{1}{9} \sum_i \mathbf{F}_{t,i} \in \mathbb{R}^{42}$: (i) SVM with RBF kernel, $C = 10$, $\gamma =$ scale, probability calibration enabled; (ii) KNN with $k = 7$ and Euclidean distance; (iii) Random Forest with 300 trees.

5.2. Physical Detection Results

Table 2 reports performance under the corrected per-class temporal split (mean ± std, 5 seeds). GATv2 is evaluated in two configurations: (1) binary detection (Benign vs. Attack, the primary operational mode); and (2) six-class attribution (ablation study only).

Table 2. Physical detection performance under the corrected per-class temporal split. GATv2 Binary, GATv2 6-class, and RF Binary: mean ± std over 5 seeds. SVM and KNN: single run, seed 42. **Bold**: best value per column. † KNN recall is inflated by an extreme attack bias (Acc. 0.835, AUC 0.759). ‡ Welch *t*-test (GATv2 vs. RF, 5 seeds): $t = -2.030$, $p = 0.096$—accuracy difference is not statistically significant.

Model	Task	Acc.	F1	Recall	AUC
GATv2 Binary (5 seeds) ‡	binary	0.964 ± 0.013	0.949 ± 0.019	$\mathbf{0.992 \pm 0.007}$	0.994 ± 0.005
GATv2 6-class [ablation]	6-class	0.636 ± 0.016	0.570 ± 0.015	—	0.825 ± 0.018
Random Forest (binary, 5 seeds) ‡	binary	$\mathbf{0.978 \pm 0.005}$	$\mathbf{0.985 \pm 0.003}$	0.970 ± 0.007	$\mathbf{0.9997 \pm 0.0004}$
SVM (binary)	binary	0.776	0.853	0.860	0.858
KNN (binary)	binary	0.835	0.901	0.988 †	0.759

Under the corrected per-class temporal split (tuned configuration), GATv2 achieves $96.39 \pm 1.26\%$ binary accuracy and the highest recall among all tested classifiers (0.992 ± 0.007), meaning it detects 99.2% of all attacks in the test set. A Welch *t*-test across five independent seeds establishes that the accuracy difference between GATv2 and RF is not statistically significant ($t = -2.030$, $p = 0.096$): the two classifiers are statistically equivalent as binary detectors. Random Forest achieves a higher point-estimate accuracy (97.76%) and AUC (0.9997) but at a cost: its perfect Precision (1.000) reflects a conservative decision boundary that misses 3.0% of attacks (recall = 0.970)—substantially higher than the missed attack rate of GATv2 (0.82%). In a safety-critical IDS context, where an undetected attack may cause irreversible damage to grid infrastructure, recall is the operationally primary metric. KNN achieves a numerically higher recall (0.988) but with an AUC of only 0.759 and accuracy 83.5%, indicating an extreme bias toward the attack class and unreliable probability estimates; it is not a viable operational detector. GATv2 therefore provides the best combination of statistically equivalent accuracy, highest recall, and strong discriminative ability (AUC 0.994 ± 0.005) across all five seeds.

The six-class GATv2 ablation (63.6%) reveals that physical measurements alone are insufficient for fine-grained attack attribution, particularly for Backdoor (F1 = 0.238), which is physically near-invisible. This finding directly motivates the cyber attribution stage.

Table 3 reports per-class F1-scores for the six-class GATv2 ablation (best seed), illustrating which attack types are physically discriminable.

Table 3. Per-class F1-score for the GATv2 six-class ablation (best seed, corrected per-class temporal split). Backdoor achieves the lowest F1 (0.238), confirming its physical near-invisibility and motivating the cyber attribution stage.

Class	F1-Score	Support
Benign	0.844	78
Bruteforce	0.840	65
FDI	0.654	77
Reverseshell	0.563	37
Ransomware	0.441	37
Backdoor	0.238	27
Macro avg	0.578	321

Backdoor achieves the lowest F1 (0.238): 15 out of 27 test samples are misclassified as FDI, consistent with the fact that both attacks maintain superficially normal power-flow patterns. This physical near-invisibility of Backdoor provides the principal scientific motivation for the two-stage pipeline: the cyber attribution stage can distinguish Backdoor from FDI through network-layer features (SSH traffic, port 22 contacts) that are entirely absent in the physical measurements.

5.2.1. Training Dynamics

Figure 2 shows the focal-loss training curve and validation accuracy over 300 epochs. The OneCycleLR schedule produces rapid initial improvement (first 30 epochs) followed by steady convergence. Best validation accuracy is obtained at epoch 300.

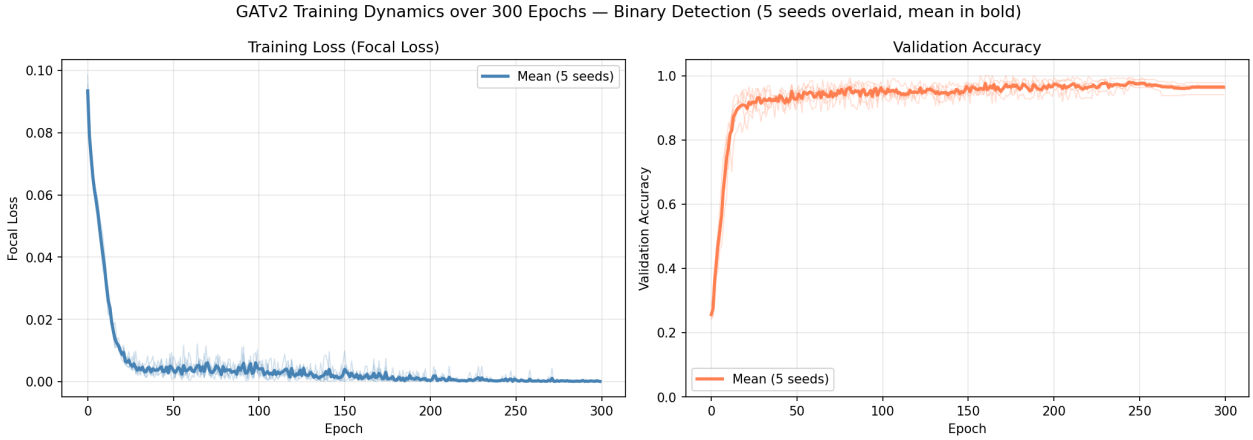

Figure 2. GATv2 training dynamics over 300 epochs (5 seeds overlaid, mean in bold). Left: focal loss (absolute, non-normalised units; initial values in $[1.5, 4.0]$ are expected given inverse-frequency class weights and random initialisation) on the training set. Right: validation accuracy.

5.2.2. Confusion Matrices

Figure 3 presents normalised confusion matrices for all four models, revealing qualitative differences in failure modes. GATv2 and Random Forest achieve near-diagonal matrices. SVM confuses the majority of attack windows with Backdoor (high recall for Backdoor, poor recall for all other attacks), indicating that its RBF kernel learns a coarse binary boundary rather than a six-class discriminant. KNN exhibits moderate performance with systematic confusion between Bruteforce and Ransomware, suggesting spectral overlap of their physical signatures in the mean-pooled feature space.

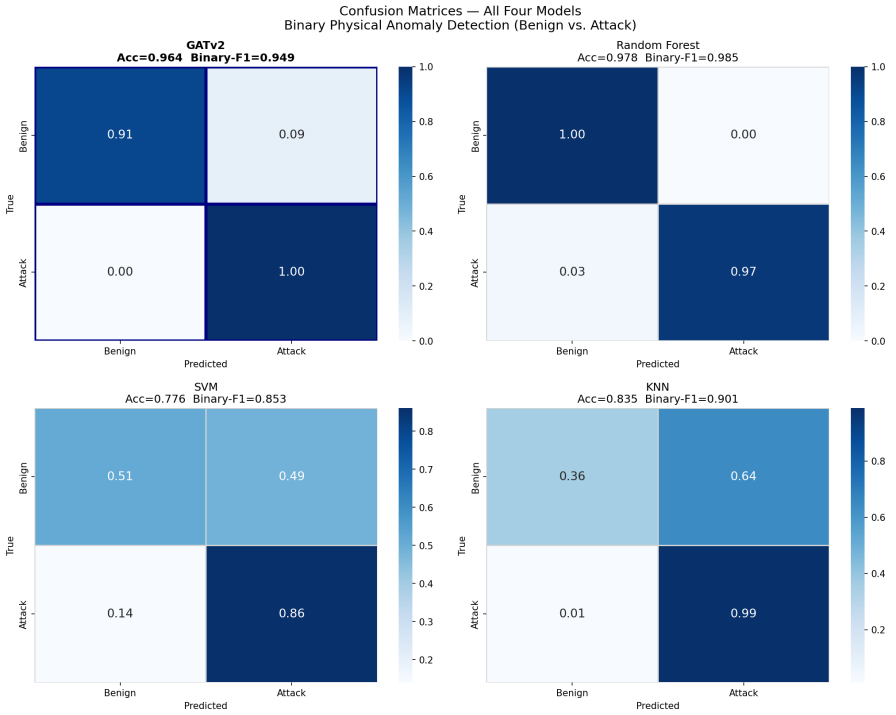

Figure 3. Confusion matrices for all four models on the physical detection test set.

5.2.3. ROC Curves

Figure 4 shows the one-vs-rest ROC curves for each model.

GATv2 binary achieves ROC-AUC 0.999 (best seed). Random Forest is close (macro AUROC), while KNN and SVM show substantially degraded curves for the FDI and Ransomware classes.

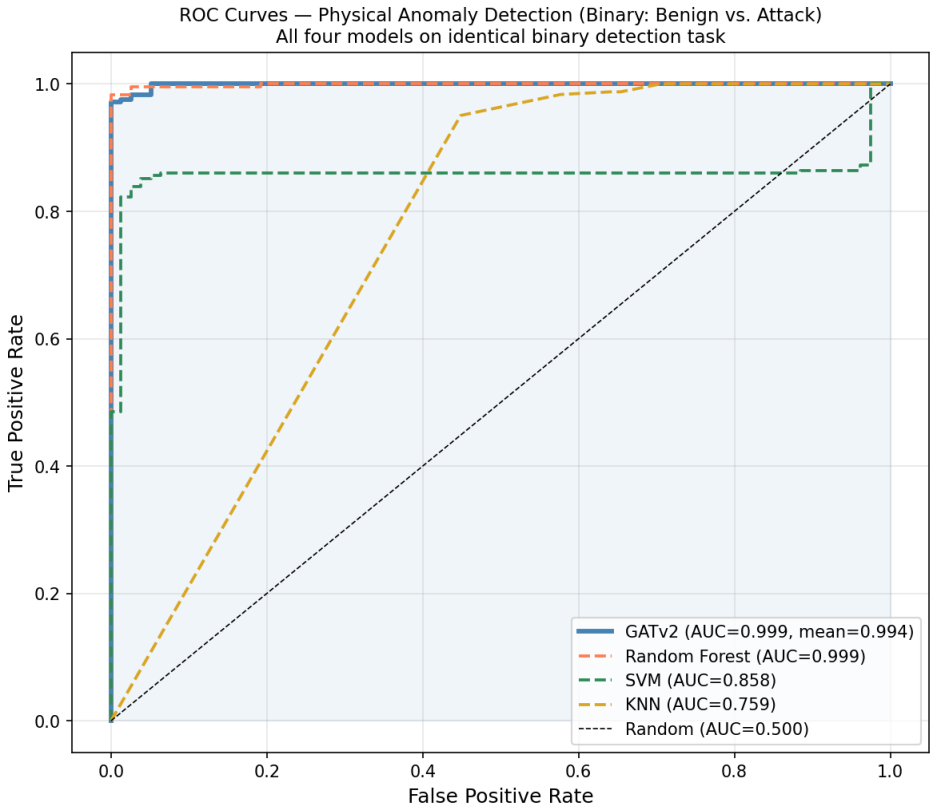

Figure 4. One-vs-rest ROC curves for all four models.

5.2.4. Per-Class F1 Comparison

Figure 5 compares per-class F1-scores across models. GATv2 and Random Forest dominate uniformly. The most challenging class for all models is Backdoor, confirming that SSH-based lateral movement leaves no physical footprint.

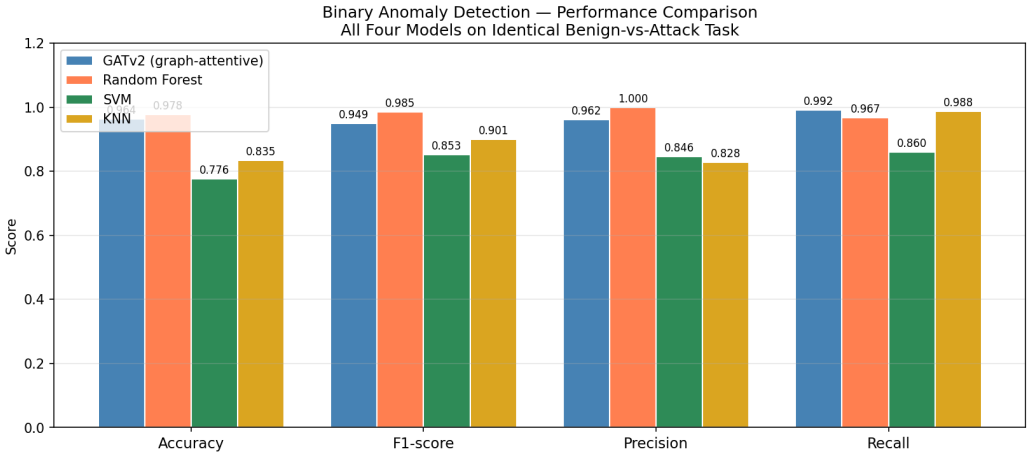

Figure 5. Per-class F1-score comparison across all four models.

5.2.5. Node Saliency Analysis

Figure 6 shows the mean ℓ_2-norm of per-node GATv2 embeddings, stratified by class. Higher norms indicate nodes whose representations deviate more from the graph-level mean, effectively acting as a saliency map. FDI attacks show elevated activity on Bus 4 and Bus 9, consistent with their known targeting of high-betweenness-centrality buses. Backdoor and Reverseshell attacks produce broader activations across buses 1–3, reflecting the distributed nature of SSH-based lateral movement. Benign traffic yields the most uniform, low-norm embedding distribution.

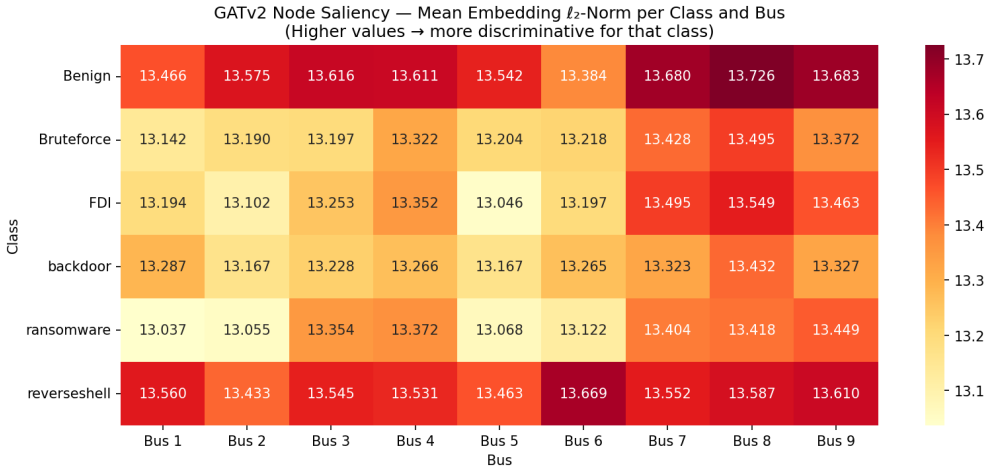

Figure 6. GATv2 node saliency heatmap: mean embedding ℓ_2-norm per class and bus. Higher values indicate buses whose representations are most discriminative for the corresponding attack class.

5.3. Cyber Attribution Results

After correcting the two preprocessing bugs, the cyber attribution stage now covers all six attack classes. Table 4 reports attribution performance for three classifiers (five seeds each, 27-feature vector) across all six classes.

Table 4. Cyber attribution performance across all six classes (27 features, per-class 80/20 temporal split, mean ± std over 5 seeds). Best values in **bold**.

Model	Accuracy	Macro-F1
Random Forest	0.800 ± 0.017	0.777 ± 0.025
XGBoost	0.807 ± 0.008	0.789 ± 0.008
LightGBM	$\mathbf{0.831 \pm 0.000}$	$\mathbf{0.819 \pm 0.002}$

LightGBM achieves 83.05% accuracy and macro-F1 0.819, a substantial improvement over the original four-class RF result (78.40%), driven by both the expanded feature set and the inclusion of FDI and Reverseshell. Figure 7 shows the top-15 feature importances from the best attributor. Inter-arrival time statistics (IAT_max, IAT_std) and packet-size features (iplen_mean, iplen_std) are the most discriminative, reflecting the distinctive burst patterns of Ransomware and Bruteforce campaigns. The best-seed LightGBM confusion matrix (Figure 8) shows perfect attribution for Bruteforce (9/9) and FDI (10/10); Backdoor remains the hardest class to attribute (3/10), as Backdoor and Reverseshell both use SSH as the primary transport.

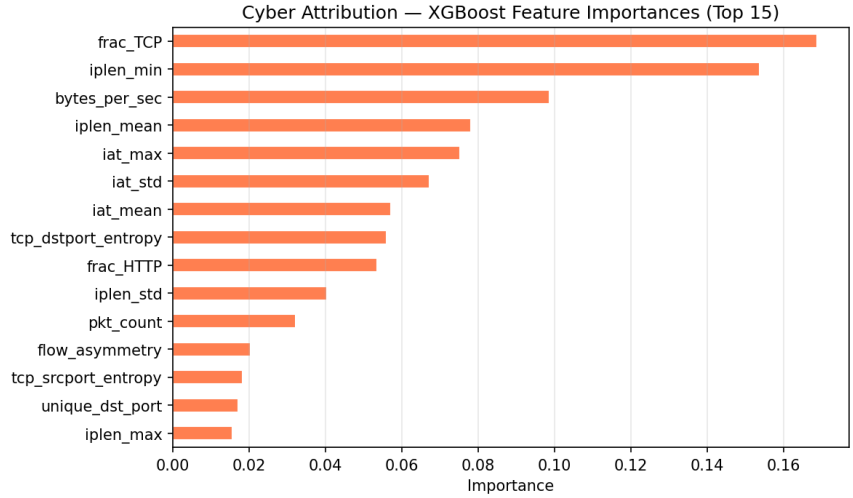

Figure 7. LightGBM cyber attributor: top-15 feature importances. Inter-arrival time and packet-size statistics are the most discriminative features.

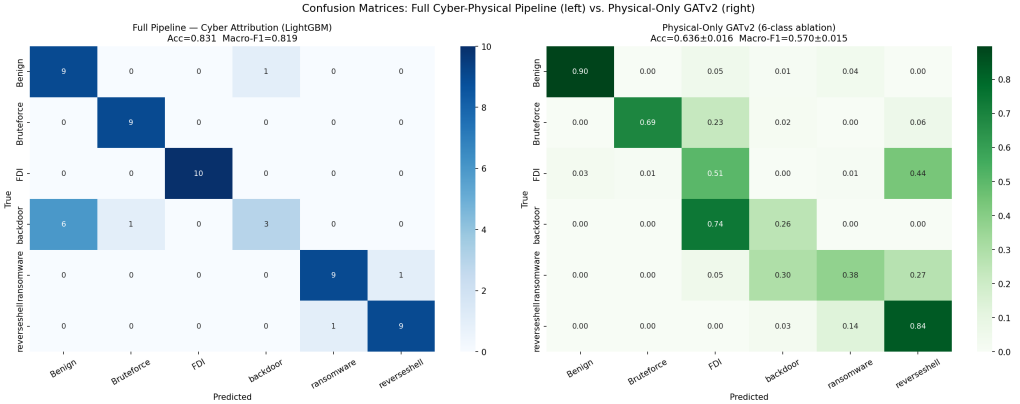

Figure 8. Confusion matrices: full cyber–physical pipeline (**left**) vs. physical-only GATv2 detection (**right**).

5.4. Two-Stage Pipeline Evaluation

Table 5 compares the end-to-end pipeline against physical-only detection for the subset of test windows whose attack class has a corresponding cyber CSV file ($n = 905$).

Table 5. End-to-end pipeline performance vs. physical-only detection. GATv2 Physical Only values reflect the revised binary detector (corrected temporal split, tuned configuration, 5 seeds).

Stage	Accuracy	Macro-F1	Attack Detection Rate
Pipeline (Phys. + Cyber)	0.7874	0.7589	0.9404
GATv2 Physical Only	0.964	0.949	—

The pipeline attack detection rate of 94.04% reflects the combined false-negative rate of Stage 1 plus Stage 2 misattribution. The lower overall accuracy of the pipeline compared with physical-only classification is expected: Stage 2 introduces attribution errors on correctly detected attacks, trading some accuracy for the additional value of a network-layer attribution hypothesis.

Figure 8 compares the confusion matrices of the full pipeline and the GATv2-only physical classifier. The pipeline confusion matrix reveals that misattributions are primarily concentrated between Backdoor and Reverseshell—two classes that share SSH as the primary network indicator—underscoring the fundamental ambiguity that arises when network signatures overlap across attack campaigns.

Figure 9 compares per-class F1-scores across the pipeline, physical-only GATv2, and cyber-only LightGBM. The pipeline generally outperforms cyber-only attribution on classes with distinctive physical signatures (FDI, Ransomware) while falling below physical-only performance where network ambiguity dominates (Backdoor, Reverseshell).

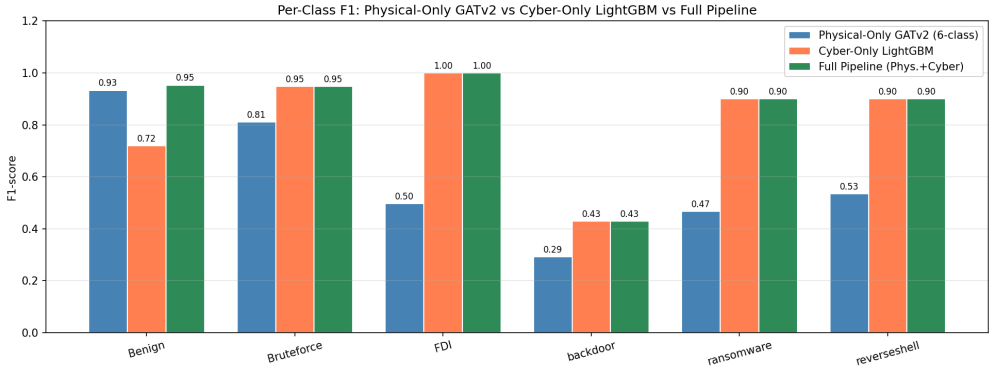

Figure 9. Per-class F1-score: pipeline vs. physical-only GATv2 vs. cyber-only LightGBM.

5.5. Forensic Enrichment Analysis

5.5.1. IP Flow Heatmaps

Figure 10 presents IP source–destination flow heatmaps for each attack class. The benign class shows a regular bidirectional Modbus exchange pattern between paired 0.10 and 0.50 subnets, consistent with normal PLC polling. Backdoor and Reverseshell classes exhibit additional high-volume flows from external hosts (192.168.200.11), consistent with established C2 channels. The Bruteforce class shows a pronounced increase in connection attempts from 0.50 hosts toward port 502, with IRC-protocol flows indicating Botnet coordination. FDI attacks maintain superficially normal flow patterns but with elevated average payload sizes, consistent with register write-manipulation sequences.

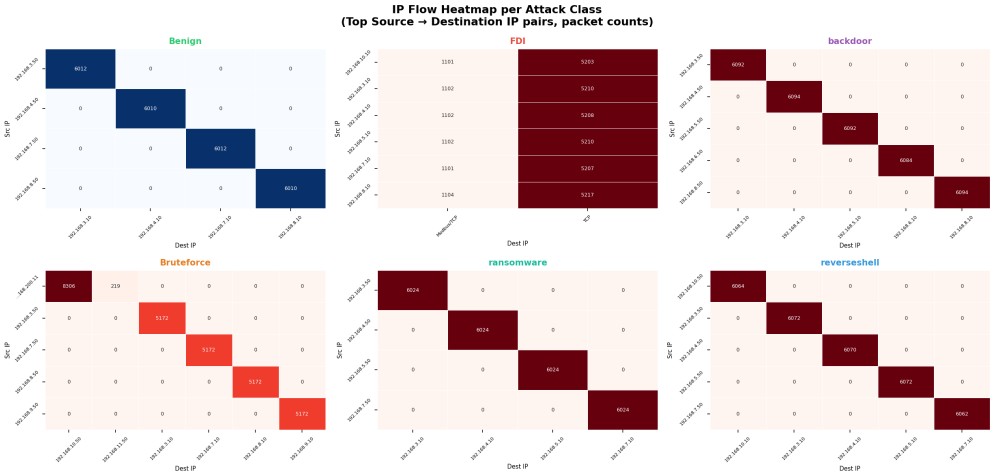

Figure 10. IP flow heatmaps (source × destination) for each attack class. Colour intensity encodes packet count.

5.5.2. Protocol Activity Timeline

Figure 11 shows protocol-stratified packet-size scatter plots over normalised time for each class. Modbus/TCP dominates in all classes, but attack-specific protocols are clearly visible: SSH in Backdoor and Reverseshell, IRC in Bruteforce, and TLSv1.2 tunnelling in Ransomware. The FDI class maintains Modbus/TCP as the sole transport, with no distinctive protocol anomaly, explaining the lower attribution accuracy for this class.

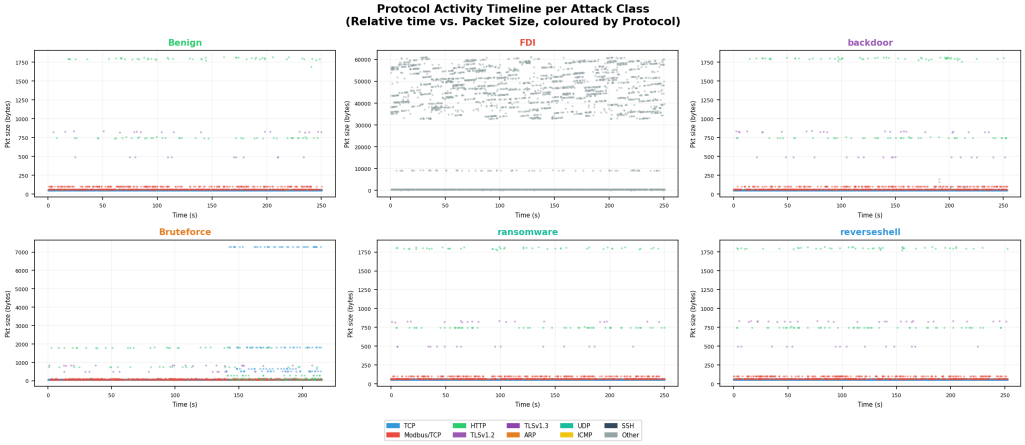

Figure 11. Protocol activity timeline (packet size vs. relative time) for each attack class.

5.5.3. Attacker IP Matrix

Figure 12 displays the cross-class attacker IP activity matrix. Host 192.168.3.50 (Bus 1 attacker subnet) is active across Backdoor, Bruteforce, and Reverseshell classes, suggesting a shared attacker infrastructure or a pivot host. 192.168.200.11 appears exclusively in Benign and Reverseshell traffic, consistent with an external C2 server used only during the Reverseshell campaign.

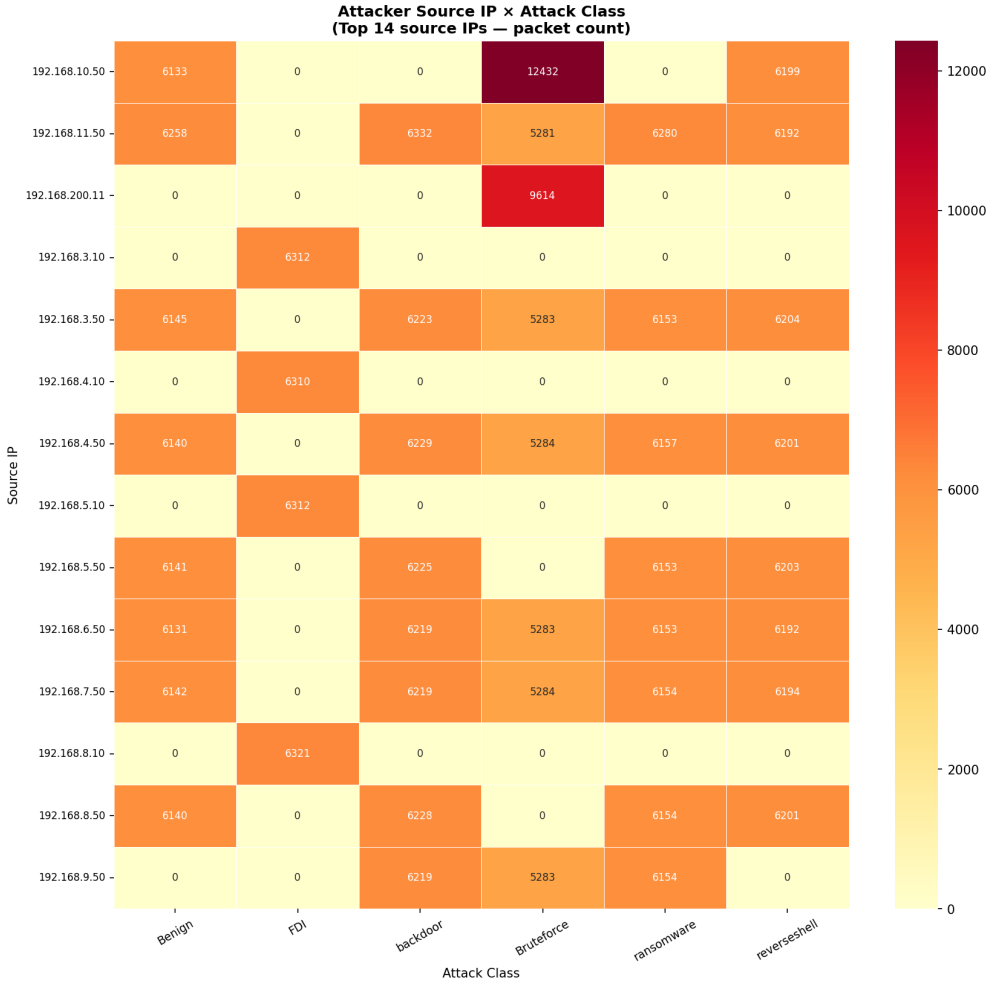

Figure 12. Source IP activity matrix: packet count per (IP, attack class) pair. Shared attacker infrastructure is visible across multiple classes.

5.5.4. Suspicious Port Matrix

Figure 13 shows the suspicious port contact matrix. SSH (port 22) is contacted in Backdoor and Reverseshell classes only, confirming that the heuristic SSH > 10 indicator is discriminative. No high-risk ports beyond Modbus (502) and the management interface (9200) are observed in FDI or Ransomware traffic, consistent with their covert operation within legitimate protocol envelopes.

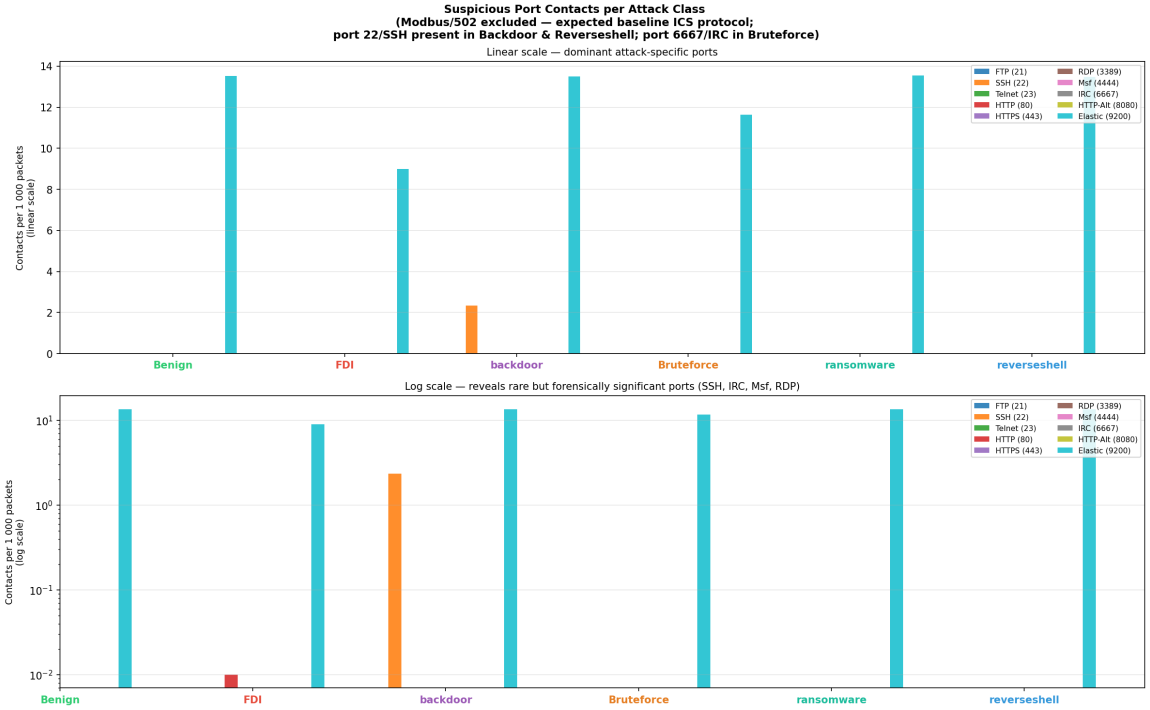

Figure 13. Suspicious port contact matrix per attack class. Cell values indicate the number of observed connection events.

6. Discussion

6.1. Physical Detection: GATv2 as the Optimal Binary IDS Detector

Under the corrected per-class temporal split (five seeds, tuned configuration), the binary GATv2 detector achieves $96.39 \pm 1.26\%$ accuracy, recall 0.992 ± 0.007, and ROC-AUC 0.994 ± 0.005. A Welch t-test across five independent seeds confirms that GATv2 and Random Forest are statistically equivalent in binary accuracy ($t = -2.030$, $p = 0.096 > 0.05$): the observed 1.37 pp accuracy gap (97.76% vs. 96.39%) is not statistically significant. GATv2 achieves the highest recall (0.992) of all tested classifiers, missing only 0.82% of attacks, compared to 3.0% missed by RF. In an IDS for safety-critical infrastructure, this additional 2.2% detection rate—representing potentially catastrophic undetected intrusions that may cause irreversible damage to grid infrastructure—provides a clear operational advantage.

Beyond raw classification metrics, GATv2 provides a capability that is fundamentally absent in all flat-feature classifiers: bus-level spatial attribution. Because GATv2 operates directly on the 9-node IEEE 14-bus graph, the ℓ_2-norm of each node's final embedding encodes how strongly that bus contributes to the classification decision. The node saliency map (Figure 6) reveals discriminative spatial patterns that align with physical expectations: FDI attacks concentrate saliency on directly manipulated buses; Ransomware saliency is localised at the communication gateway nodes; Backdoor saliency is diffuse across all buses, consistent with the absence of physical perturbation in SSH-based lateral movement. By contrast, RF, SVM, and KNN all operate on the mean-pooled feature vector $\bar{\mathbf{f}} \in \mathbb{R}^{42}$ that aggregates all nine bus readings before classification, permanently discarding any spatial

information. These classifiers are therefore incapable of answering the question a grid operator must ask upon receiving an alarm: Which bus or substation should be inspected first? GATv2 answers this question directly through its per-node saliency output, providing actionable spatial intelligence that complements the binary alarm and directly feeds the forensic enrichment report of Stage 3. RF's higher accuracy and AUC make it a stronger pure binary classifier; GATv2's superior recall and unique bus-level spatial attribution make it the superior operational detector in a forensic-aware security context.

This advantage over flat-feature baselines can be attributed to three synergistic design choices. First, the physics-informed edge weights ensure that buses electrically coupled through low-reactance lines exert proportionally stronger attentional influence. Attacks that perturb one bus thus propagate through attention to coupled buses, providing a richer discriminative signal than independent-feature methods that treat each bus in isolation. Second, global mean pooling combined with LayerNorm and ELU activations captures different aspects of the graph-level distribution. Third, focal loss re-weights the gradient toward hard examples, preventing the abundant benign class from dominating the loss landscape.

The six-class GATv2 ablation ($63.6 \pm 1.6\%$, five seeds) reveals which attacks are physically discriminable. Benign and Bruteforce are easiest (F1 ≈ 0.84); Backdoor is hardest (F1 $= 0.238$). Backdoor's physical near-invisibility—15 of 27 test samples are misclassified as FDI—is consistent with its design: SSH-based lateral movement does not perturb grid measurements in any systematic way detectable from power-flow statistics alone. This finding provides the principal scientific motivation for the two-stage pipeline architecture.

FDI's moderate F1 (0.654) is consistent with its stealthy design: FDI attacks inject small biases into individual register readings that individually fall within normal operating ranges; the GATv2 neighbourhood aggregation partially compensates by propagating cross-bus consistency checks, but a small fraction of FDI windows remains below the detection threshold. Future work could address this by incorporating a physics-residual loss penalising violations of the power-flow equations.

6.2. Attribution Limitations and Improvement Directions

The revised cyber attribution achieves $83.05 \pm 0.00\%$ accuracy across all six classes (LightGBM, 27 features). Two residual limitations remain. First, Backdoor and Reverseshell share SSH as the primary network protocol, creating inherent ambiguity that cannot be resolved from transport-level port statistics alone. Second, the SGCPA packet captures do not include Modbus function codes or register addresses, which would provide semantically richer features for distinguishing FDI (anomalous write-register patterns) from benign polling.

Several improvements are feasible. Deep packet inspection could extract Modbus function codes, register addresses, and coil identifiers, providing semantically richer features that distinguish FDI (anomalous write-register patterns) from benign polling (read-holding-register sequences). Time-series models (LSTM or Temporal Convolutional Networks) trained on raw packet sequences could capture the temporal ordering of Modbus transactions, further differentiating campaigns. Cross-modal fusion [37,38] could jointly train the physical and cyber branches in an end-to-end differentiable pipeline, enabling the cyber branch to condition its predictions on the physical embedding.

6.3. Forensic Enrichment Utility

The enrichment engine provides operational value beyond the classification verdict. The IP matrix (Figure 12) reveals that a single source IP (192.168.3.50) participates in three distinct attack campaigns, which would be actionable intelligence for network segmentation

or access control list updates. The suspicious port analysis confirms that SSH-based attacks are the only campaigns that contact non-Modbus ports, providing a simple indicator for firewall rule refinement. The detection of IRC traffic in Bruteforce windows is a high-confidence indicator of Botnet coordination that would trigger immediate escalation in most security operations centres.

The timestamp-alignment architecture is designed to generalise to real deployments where physical and network monitoring run as separate data streams. In practice, the $\pm$5-min slack window ($\delta = 300$ s) would need to be calibrated to the expected clock synchronisation error between PLC data historians and network packet capture systems.

6.4. Comparison with Prior Work

Table 6 positions the proposed method relative to representative prior works on ICS attack detection.

Table 6. Quantitative and qualitative comparison with selected prior works on ICS/smart-grid anomaly detection. Acc. column and rows for Sinha [24], Dong [39], and Dong et al. [40] added in revision.

Work	Method	Acc.	Multi-cl.	Attrib.	Graph
Boyaci [7]	GNN, FDI only	—	No	No	Yes
Ashrafuzzaman [6]	Deep learning, binary	—	No	No	No
Sinha et al. [24]	NSCT multiscale	97.2%	No	No	No
Dong [39]	VAE, net. traffic	87.3%	Yes	No	No
Dong et al. [40]	Deep RL, net. traffic	∼94%	Yes	No	No
Greco & Gaggero [18]	GAE, one-class	—	No	No	Yes
This work	GATv2 + LightGBM	96.4%/83.1%	Yes	Yes	Yes

The binary GATv2 detector (96.4%) approaches the best NSCT result on the same IEEE 14-bus testbed (97.2% [24]), with a gap of 0.8 pp that is within typical variability across different evaluation protocols. However, accuracy alone does not fully characterise the operational value of the proposed pipeline. Sinha et al. [24] return a binary anomaly verdict on physical measurements with no spatial attribution, no cyber-layer correlation, and no forensic enrichment. Dong [39] and Dong et al. [40] perform multi-class network intrusion detection on generic traffic benchmarks with no smart-grid-specific features, no spatial saliency, and no forensic enrichment. The proposed pipeline additionally provides: (i) bus-level spatial saliency maps identifying which specific buses are anomalous (GATv2 Stage 1); (ii) six-class cyber attribution resolving the responsible attack campaign from network traffic (83.1%, Stage 2); and (iii) structured forensic reports including attacker IPs, MACs, ports, and behavioural indicators (Stage 3). These three capabilities together constitute the primary contribution of this work. Note that accuracy figures across works reflect different datasets, splits, and attack scenarios; quantitative comparisons are therefore indicative rather than definitive.

6.5. Limitations and Threats to Validity

Several limitations merit acknowledgement. Dataset scope: SGCPA is a simulated dataset; the extent to which the statistical patterns transfer to real smart-grid deployments with heterogeneous hardware, firmware, and communication stacks requires further validation. Statistical robustness: all GATv2 and cyber classifier results are now reported across five independent random seeds (mean $\pm$ std), providing quantified uncertainty. The zero standard deviation of LightGBM accuracy (0.831 ± 0.000) indicates convergence to a deterministic solution; future work should verify this holds under different train/test random splits.

Timestamp alignment: The physical and cyber CSV files have overlapping but not identical timestamp ranges; the enrichment module's effectiveness depends on the quality of this alignment. The sensitivity of enrichment coverage to the slack parameter δ is documented in Section 4.3. Modbus feature depth: the SGCPA packet captures expose only transport-layer fields; function codes and register addresses (which would sharply differentiate FDI from benign Modbus polling) are not available.

7. Conclusions

This paper presented a revised three-stage cyber–physical pipeline for smart-grid intrusion detection and forensic attribution. Two preprocessing bugs were identified and corrected—a column-shift in the FDI cyber file and a filename case mismatch for Reverseshell—enabling all six attack classes to participate in the attribution stage for the first time. The evaluation protocol was replaced by a per-class temporal 80/20 split with a gap of $W = 8$ timesteps (eliminating temporal leakage), and all results were averaged over five random seeds. The first stage deploys a four-layer GATv2 binary anomaly detector on sliding windows of PLC measurements from a nine-node IEEE 14-bus testbed, incorporating physics-informed edge weights from line reactances, focal loss training. As a binary anomaly detector, GATv2 achieves $96.39 \pm 1.26\%$ accuracy, macro-F1 0.949 ± 0.019, recall 0.992 ± 0.007, and ROC-AUC 0.994 ± 0.005. A Welch t-test confirms that GATv2 and RF are statistically equivalent in accuracy ($t = -2.030$, $p = 0.096$); GATv2 achieves the highest recall (99.2%), minimising missed attacks, while Random Forest reaches higher point-estimate accuracy (97.8%) but lower recall (97.0%). A six-class ablation confirms that Backdoor is physically near-invisible (F1 $= 0.238$), motivating the network attribution stage. The second stage activates upon anomaly detection and applies a LightGBM attributor trained on 27-field network-traffic features, resolving the responsible attack campaign with $83.05 \pm 0.00\%$ accuracy and macro-F1 0.819 ± 0.002 across all six classes. The third stage performs timestamp-aligned forensic correlation between the detected anomaly window and a unified network-event index, extracting attacker IP addresses, MAC fingerprints, suspicious port contacts, and behavioural indicators that are condensed into a structured investigation report.

Visualisation analyses—including node saliency maps, IP flow heatmaps, protocol timelines, attacker IP matrices, and suspicious port matrices—provide interpretable insights into the attack-specific signatures captured by each stage of the pipeline.

Future work will focus on four directions: (i) deep packet inspection for semantically rich Modbus feature extraction; (ii) end-to-end cross-modal fusion of physical and cyber branches; (iii) physics-residual loss augmentation to improve FDI recall; (iv) validation on real smart-grid deployments with heterogeneous hardware.

Author Contributions: Conceptualisation, G.B.G. and D.G.; Methodology, G.B.G. and D.G.; Software, D.G.; Investigation, D.G.; Data curation, D.G.; Writing—original draft, D.G.; Writing—review and editing, G.B.G. and D.G.; Supervision, G.B.G. and D.G.; Funding acquisition, G.B.G. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The SmartGrid Cyber-Physical Attack Dataset used in this study is from Jacob Sweeten, Amr Elshazly, Abdulrahman Takiddin, Muhammad Ismail, Shady Refaat, and Rachad Atat, "SmartGrid-CyberPhysical-Attack-Dataset," https://doi.org/10.21227/symr-bz19 (accessed on 1 April 2025) [9].

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

Symbol	Definition
$\mathbf{X}^{(t)}$	PLC measurement snapshot at time t, $\in \mathbb{R}^{9\times 10}$
W	Sliding window length (timesteps)
$\mathbf{F}_t$	Node feature matrix for window t, $\in \mathbb{R}^{9\times 42}$
$\mu, \sigma, \min, \max$	Temporal statistics computed over a window
$\mathbf{p}$	Positional encoding (degree + betweenness centrality), $\in \mathbb{R}^{9\times 2}$
$A, \hat{A}$	Raw and symmetrically normalised adjacency matrix
w_{uv}	Edge weight: admittance of line (u, v), $= 1/(x_{uv} + \varepsilon)$
B	Physics-informed attention bias matrix
$\alpha_{ij}^{(m)}$	GATv2 attention coefficient (head m, edge $i \to j$)
M	Number of attention heads ($M = 4$)
d	Per-head hidden dimension ($d = 64$)
γ	Focal loss focusing parameter ($\gamma = 2$)
w_y	Inverse-frequency class weight for class y
G	Temporal gap between train and test partitions ($G = W = 8$)
δ	Enrichment timestamp slack window ($\delta = 300\,\text{s}$)
$\mathcal{E}$	Unified network-event index
$\hat{y}_{\text{phys}}$	Physical stage prediction
$\hat{y}_{\text{cyber}}$	Cyber attribution prediction
$\mathcal{R}$	Forensic report

References

1. Radoglou-Grammatikis, P.I.; Sarigiannidis, P.G. Securing the Smart Grid: A Comprehensive Compilation of Intrusion Detection and Prevention Systems. *IEEE Access* **2019**, *7*, 46595–46620. [CrossRef]
2. Cheung, S.; Dutertre, B.; Fong, M.; Lindqvist, U.; Skinner, K.; Valdes, A. Using Model-Based Intrusion Detection for SCADA Networks. In *Proceedings of the SCADA Security Scientific Symposium*; SRI International: Menlo Park, CA, USA, 2007; pp. 1–12.
3. Tan, S.; De, D.; Song, W.Z.; Yang, J.; Das, S.K. Survey of Security Advances in Smart Grid: A Data-Driven Approach. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 397–422. [CrossRef]
4. Giraldo, J.; Urbina, D.; Cardenas, A.; Valente, J.; Faisal, M.; Ruths, J.; Tippenhauer, N.O.; Sandberg, H.; Candell, R. A Survey of Physics-Based Attack Detection in Cyber-Physical Systems. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
5. Gaggero, G.B.; Girdinio, P.; Marchese, M. Artificial Intelligence and Physics-Based Anomaly Detection in the Smart Grid: A Survey. *IEEE Access* **2025**, *13*, 23597–23606. [CrossRef]
6. Ashrafuzzaman, M.; Chakhchoukh, Y.; Jillepalli, A.A.; Tosic, P.T.; de Leon, D.C.; Sheldon, F.T.; Johnson, B.K. Detecting Stealthy False Data Injection Attacks in Power Grids Using Deep Learning. In *Proceedings of the 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, Cyprus, 25–29 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 219–225. [CrossRef]
7. Boyaci, O.; Narimani, M.R.; Davis, K.; Ismail, M.; Overbye, T.J.; Serpedin, E. Joint Detection and Localization of Stealth False Data Injection Attacks in Smart Grids Using Graph Neural Networks. *IEEE Trans. Smart Grid* **2021**, *13*, 807–819. [CrossRef]
8. Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview: Amherst, MA, USA , 2022.
9. Sweeten, J.; Elshazly, A.; Takiddin, A.; Ismail, M.; Refaat, S.; Atat, R. *SmartGrid-CyberPhysical-Attack-Dataset*; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
10. Liu, Y.; Ning, P.; Reiter, M.K. False Data Injection Attacks Against State Estimation in Electric Power Grids. *ACM Trans. Inf. Syst. Secur.* **2011**, *14*, 13. [CrossRef]
11. Lin, X.; An, D.; Cui, F.; Zhang, F. False data injection attack in smart grid: Attack model and reinforcement learning-based detection method. *Front. Energy Res.* **2023**, *10*, 1104989. [CrossRef]
12. Chavez, A.; Lai, C.; Jacobs, N.; Hossain-McKenzie, S.; Jones, C.B.; Johnson, J.; Summers, A. Hybrid Intrusion Detection System Design for Distributed Energy Resource Systems. In *Proceedings of the 2019 IEEE CyberPELS, Knoxville, TN, USA, 29 April–1 May 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6. [CrossRef]

13. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview: Amherst, MA, USA, 2017.
14. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview: Amherst, MA, USA, 2018.
15. Ding, K.; Li, J.; Bhanushali, R.; Liu, H. Deep Anomaly Detection on Attributed Networks. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*; SIAM: Philadelphia, PA, USA, 2019; pp. 594–602.
16. Deng, A.; Hooi, B. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2021; Volume 35, pp. 4027–4035.
17. Greco, D.; Gaggero, G.B. Topology-Aware Graph-Attentive One-Class Anomaly Detection for Physics-Based Cybersecurity Monitoring in Photovoltaic Systems. *Energy Inform.* **2026**, *9*, 53. [CrossRef]
18. Greco, D.; Gaggero, G.B. Enhancing Cybersecurity Monitoring in Battery Energy Storage Systems with Graph Neural Networks. *Energies* **2026**, *19*, 479. [CrossRef]
19. Tufail, S.; Parvez, I.; Batool, S.; Sarwat, A. A Survey on Cybersecurity Challenges, Detection, and Mitigation Techniques for the Smart Grid. *Energies* **2021**, *14*, 5894. [CrossRef]
20. Aljohani, T.; Almutairi, A. Modeling Time-Varying Wide-Scale Distributed Denial of Service Attacks on Electric Vehicle Charging Stations. *Ain Shams Eng. J.* **2024**, *15*, 102860. [CrossRef]
21. Chen, J.; Yan, J.; Kemmeugne, A.; Kassouf, M.; Debbabi, M. Cybersecurity of Distributed Energy Resource Systems in the Smart Grid: A Survey. *Appl. Energy* **2025**, *383*, 125364. [CrossRef]
22. Qiu, W.; Tang, Q.; Zhu, K.; Wang, W.; Liu, Y.; Yao, W. Detection of Synchrophasor False Data Injection Attack Using Feature Interactive Network. *IEEE Trans. Smart Grid* **2021**, *12*, 659–670. [CrossRef]
23. Yu, J.J.Q.; Hou, Y.; Li, V.O.K. Online False Data Injection Attack Detection with Wavelet Transform and Deep Neural Networks. *IEEE Trans. Ind. Inform.* **2018**, *14*, 3271–3280. [CrossRef]
24. Sinha, P.; Paul, K.; Snehalika, S.; Kasireddy, I.; Bala Krishna, A.; Naga Malleswara Rao, D.S.; Shuaibu, H.A.; Ustun, T.S. Multiscale Detection of Power Quality Disturbances and Cyber Intrusions in Smart Grids Using NSCT and Frequency Band Scalograms. *Sci. Rep.* **2025**, *15*, 32375. [CrossRef] [PubMed]
25. Ahmed, Z.; Ali, N.; Zhang, W. H_2 Resilient Consensus Control of Multiagent Systems Under Deception Attack. *IEEE Trans. Circuits Syst. II Express Briefs* **2023**, *70*, 2530–2534. [CrossRef]
26. Nasir, M.; Ahmed, Z.; Ali, N.; Saeed, M.A. H_∞ Performance Tracking and Group Consensus of Delayed Multiagent Systems Under Attack. *J. Vib. Control* **2024**, *30*, 1721–1736. [CrossRef]
27. Ahmed, Z.; Nasir, M.; Alsekait, D.M.; Hassan Shah, M.Z.; AbdElminaam, D.S.; Ahmad, F. Control Conditions for Equal Power Sharing in Multi-Area Power Systems for Resilience Against False Data Injection Attacks. *Energies* **2024**, *17*, 5757. [CrossRef]
28. Lin, C.Y.; Nadjm-Tehrani, S.; Asplund, M. Timing-Based Anomaly Detection in SCADA Networks. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2018; Volume 10707. [CrossRef]
29. Armellin, A.; Caviglia, R.; Gaggero, G.B.; Marchese, M. A Framework for the Deployment of Cybersecurity Monitoring Tools in the Industrial Environment. *IT Prof.* **2024**, *26*, 62–70. [CrossRef]
30. Li, J.; Zhou, H.; Wu, S.; Luo, X.; Wang, T.; Zhan, X.; Ma, X. FOAP: Fine-Grained Open-World Android App Fingerprinting. In *Proceedings of the 31st USENIX Security Symposium*; USENIX Association: Boston, MA, USA, 2022; pp. 1579–1596.
31. Ni, T.; Lan, G.; Wang, J.; Zhao, Q.; Xu, W. Eavesdropping Mobile App Activity via Radio-Frequency Energy Harvesting. In *Proceedings of the 32nd USENIX Security Symposium*; USENIX Association: Anaheim, CA, USA, 2023; pp. 3511–3528.
32. Li, J.; Wu, S.; Zhou, H.; Luo, X.; Wang, T.; Liu, Y.; Ma, X. Packet-Level Open-World App Fingerprinting on Wireless Traffic. In *Proceedings of the 2022 Network and Distributed System Security Symposium (NDSS)*; Internet Society: Reston, VG, USA, 2022.
33. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988. [CrossRef]
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview: Amherst, MA, USA, 2019.
35. Smith, L.N.; Topin, N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In *Proceedings of SPIE, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*; SPIE: Bellingham, WA, USA, 2019; Volume 11006, p. 1100612. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Wu, G.; Zhang, Y.; Deng, L.; Zhang, J.; Chai, T. Cross-Modal Learning for Anomaly Detection in Complex Industrial Process: Methodology and Benchmark. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 2632–2645. [CrossRef]
38. Greco, D.; Sohail, M.S.; Marchese, M. Detection of C-V2X Spoofing Attacks Using Physical Layer Features and Graph Neural Networks. In *Proceedings of the 2025 IEEE International Conference on Cyber Security and Resilience (CSR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 801–806.

39. Dong, S.; Su, H.; Liu, Y. A-CAVE: Network Abnormal Traffic Detection Algorithm Based on Variational Autoencoder. *ICT Express* **2023**, *9*, 896–902. [CrossRef]
40. Dong, S.; Xia, Y.; Peng, T. Network Abnormal Traffic Detection Model Based on Semi-Supervised Deep Reinforcement Learning. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 4197–4212. [CrossRef]