

Ground-Truth Physics Residuals for Detecting Stealthy False Data Injection Attacks in Smart Grids

Surya Prakash Chaturvedula¹, Dr. Pawan Kumar Pareek²

¹Research Scholar, Department of Computer Science, Nirwan University, Jaipur, India

²Associate Professor, Department of Computer Science, Nirwan University, Jaipur, India

doi.org/10.64643/IJIRTV13I5-209166-459

Abstract—Stealthy false data injection (FDI) attacks exploit a fundamental weakness of conventional residual-based bad-data detection: a carefully structured measurement perturbation can remain statistically acceptable to the state-estimation residual while moving the reported measurements away from the electrical state represented by the actual network. This paper develops and empirically evaluates a ground-truth physics-residual methodology for identifying that discrepancy. The key idea is to compute electrical-consistency residuals directly from the defender-side network admittance matrix (Ybus), reconstructed voltage phasors, reported active/reactive power, and line-flow balance, and to use these residuals as an explicit feature stream in a spatial-temporal detector. The evaluation uses a corrected benchmark of 2,016 simulation runs spanning IEEE 39-bus and IEEE 118-bus systems, 18 balanced scenario cells, and three attacker-knowledge tiers for stealthy FDI. The primary detector combines graph attention, temporal LSTM representation, and the physics-residual stream. On the held-out corrected test split, pooled precision, recall, F1 and ROC-AUC are 0.727, 0.715, 0.721 and 0.815, respectively, compared with F1/ROC-AUC of 0.596/0.569 for a Chi-squared bad-data detector and 0.621/0.677 for a vanilla LSTM. FDI-specific F1 is 0.910, 0.902 and 0.892 for attacker tiers 1–3. Removing the physics stream reduces F1 by 34.8, 38.2 and 36.0 percentage points for the three FDI tiers. A sequence-position control achieves F1=0.0 and ROC-AUC=0.5 after correction of an earlier fixed-onset leakage defect. The results provide strong first-pass evidence that explicit electrical-consistency information materially improves stealthy-FDI detection, while also exposing an important topology-dependent false-positive problem: 0% on IEEE 39-bus versus 25.5% on IEEE 118-bus. The paper therefore presents the residual methodology as a reproducible research mechanism rather than as evidence of deployment readiness.

Index Terms—smart-grid cybersecurity; false data injection; physics residual; Ybus; bad-data detection;

physics-informed machine learning; graph attention network; LSTM; cyber-physical security; power-system state estimation.

I. INTRODUCTION

Modern electric grids are cyber-physical systems in which measurement devices, communication networks, state estimators and electrical assets interact continuously. The security of the monitoring layer is therefore not separable from the physical consistency of the grid state. False data injection attacks are particularly difficult because an adversary can manipulate measurement values without necessarily producing an obvious statistical outlier. The seminal formulation by Liu, Ning and Reiter demonstrated that structured measurement attacks can evade conventional bad-data tests under appropriate knowledge of the power-system model [1]. Subsequent work has established that the problem remains relevant under partial topology knowledge, cooperative attackers and modern learning-based defenses [2–5].

A conventional bad-data detector evaluates a residual generated by a state-estimation model. In simplified linear form, a measurement vector z can be written as $z = Hx + e$, and a structured attack may be represented as $z' = z + a$. If a lies in the measurement-model subspace, the attack can leave the classical residual essentially unchanged. This creates a central research problem: a detector that only asks whether a measurement is statistically compatible with the estimator may not ask whether the measurement is electrically compatible with the actual network.

This paper investigates a complementary question: can a detector explicitly reconstruct the electrical implications of reported measurements using the ground-truth network model and expose

inconsistencies that are not captured by conventional residual tests? The proposed approach does not replace state estimation. Instead, it computes a separate physics-residual representation using the defender-side Ybus matrix and combines that representation with learned spatial and temporal features.

The paper deliberately focuses on the residual methodology rather than presenting a broad claim that every physics-informed detector will generalize. The completed evidence is benchmark-based and single-seed. The strongest evidence is the controlled removal of the physics stream, the corrected leakage-control experiment, and the consistent FDI-tier results. The paper also reports the 118-bus false-positive rate explicitly because hiding this result would give an incomplete view of detector behavior.

1.1 Research questions

RQ1: Can ground-truth electrical-consistency residuals distinguish stealthy FDI conditions from normal measurements when classical statistical residual detection is insufficient?

RQ2: Does an explicit physics-residual feature stream materially improve a spatial-temporal detector for stealthy FDI?

RQ3: Does the observed benefit persist across attacker knowledge tiers?

RQ4: Does the corrected benchmark remove the sequence-position shortcut that existed in the original fixed-onset generator?

RQ5: Does the residual-based representation remain consistently calibrated across IEEE 39-bus and IEEE 118-bus systems?

1.2 Contributions

1. A ground-truth physics-residual formulation that compares reported electrical quantities with quantities implied by the defender-side Ybus model.
2. A dual residual representation combining AC power-consistency and line-flow/power-balance consistency.
3. Integration of the residual stream with graph-attention and temporal LSTM features without using a physics-constrained optimization loss.
4. Controlled evidence that removing the residual stream substantially reduces stealthy-FDI detection performance.

5. A leakage-corrected evaluation protocol using randomized attack onset, bounded duration and physical-state restoration.

6. A topology-stratified error analysis that exposes a 25.5% normal-operation false-positive rate on the IEEE 118-bus case.

II. BACKGROUND AND RELATED WORK

2.1 Stealthy FDI and residual-based detection

FDI attacks are designed to alter measurements used by state estimation. The original attack formulation showed that carefully constructed attack vectors can bypass existing bad-measurement detection [1]. Reviews subsequently documented the broader attack surface and the continuing limitations of residual-based defenses [2]. The implication is not that residuals are useless; rather, the residual must be defined relative to the right model and the right threat assumptions.

Recent research increasingly combines temporal learning, graph learning and physics-aware representations. Graph convolution and graph-attention approaches exploit the topology of the power network, while recurrent and temporal-convolution approaches model evolution across measurements [6–8]. These methods motivate a spatial-temporal detector, but they do not eliminate the need to distinguish statistical regularity from electrical consistency. The present study therefore treats physics as an explicit information source rather than assuming that a neural network will infer all electrical constraints implicitly.

2.2 Research gap

Three gaps motivate the present study. First, benchmark construction frequently emphasizes label balance and attack diversity without providing an independently computable electrical-consistency reference. Second, many detector evaluations report pooled accuracy without showing whether the physics representation itself contributes to the result. Third, fixed attack timing can introduce a sequence-position shortcut that makes apparently strong temporal models misleading. The present work addresses these gaps together: the residual is calculated from the known network model, its contribution is tested by ablation, and the corrected dataset randomizes attack onset and duration.

2.3 Distinction between a physics residual and a physics-constrained loss

The terminology in this paper is intentionally precise. The detector is called physics-informed because it receives an explicit feature stream derived from electrical equations. The completed experiment does not impose a physics-constrained loss, differentiable power-flow constraint, or physics-based regularization term during neural-network optimization. Claims of physics-constrained training would therefore be inaccurate. This distinction is important for reproducibility and for fair comparison with PINN-style methods.

III. PROBLEM FORMULATION AND THREAT MODEL

3.1 Network representation

Let the transmission network be represented by a graph $G = (V, E)$, where buses are nodes and transmission elements are edges. At time t , the reported electrical state includes voltage magnitude $|V_i, t|$, voltage angle $\theta_{i, t}$, active and reactive power quantities, line-flow variables and line-status information. The defender is assumed to know the correct network topology and the corresponding bus-admittance matrix Y_{bus} . The attacker does not necessarily possess this complete information.

3.2 Attack model

The principal threat is measurement-level FDI. Three attacker-knowledge tiers are evaluated. Tier 1 represents high knowledge, with complete topology/model information and AC-feasible alternate-state generation. Tier 2 represents partial local topology knowledge and topology-bounded perturbation; this is the principal partial-knowledge evaluation case. Tier 3 represents limited topology knowledge with a generic statistical perturbation mechanism. These tiers are experimental constructs for controlled comparison, not a claim that real adversaries fall into exactly three categories.

3.3 Ground-truth defender assumption

The defender-side Y_{bus} matrix is treated as trusted ground truth within the benchmark. This assumption is intentionally asymmetric: the attacker may have incomplete information, while the detector is evaluated against the actual network model. The

methodology therefore answers a specific question—whether electrical inconsistency relative to a trusted network model can reveal measurement manipulation—not the broader question of whether a compromised defender model can itself be detected.

3.4 Formal detection objective

Given X_t , G , Y_{bus} and reported electrical measurements M_t , estimate $y_t \in \{0,1\}$ such that $y_t=1$ denotes an attack, while minimizing false alarms under normal operation.

The physics representation is denoted P_t . The learned spatial representation is S_t and the temporal representation is T_t . The fused representation is $Z_t = [S_t, T_t, P_t]$, followed by a binary classification head. The residual stream is therefore a complementary representation rather than the sole decision variable.

IV. GROUND-TRUTH PHYSICS RESIDUAL FORMULATION

4.1 AC voltage reconstruction

For each bus i and time step t , the reported voltage magnitude and angle are converted to a complex phasor. Let V_t denote the vector of complex bus voltages. Using the defender-side Y_{bus} matrix, the electrically implied injected complex power is obtained from the network equation.

$$I_t = Y_{bus} V_t$$

$$S_{calc, t} = \text{diag}(V_t) \text{conj}(I_t) \text{baseMVA}$$

The calculated active and reactive components are compared with the reported values. Under a physically consistent state, the discrepancy should be close to the numerical precision of the simulation and arithmetic implementation, subject to the exact sign convention used by the dataset generator.

$$r_{P, i, t} = P_{calc, i, t} - P_{reported, i, t}$$

$$r_{Q, i, t} = Q_{calc, i, t} - Q_{reported, i, t}$$

4.2 Sign convention and implementation validation

A critical implementation issue was identified during development: the internal power-reporting convention used by the primary simulation environment can be the mathematical opposite of the standard network-injection convention. The residual implementation was therefore validated explicitly, including generator buses, before attack results were interpreted. After correction, normal and pre-attack residuals remained

near the numerical floor, while FDI intervals produced much larger residuals. This check is essential because a sign error can create artificial residuals and invalidate the interpretation of the entire physics stream.

4.3 Line-flow and power-balance residuals

A second residual family checks consistency between reported line flows and bus-level power balance. For each relevant bus, incoming and outgoing line-flow quantities are aggregated and compared with the reported net injection or demand. This creates a topology-aware invariant that is complementary to the nodal AC power reconstruction.

$$r_{\text{bal}, i, t} = P_{\text{net}, i, t} - (\sum P_{\text{in}, i, t} - \sum P_{\text{out}, i, t})$$

The final physics representation contains eight residual components spanning the AC consistency and line-flow/power-balance families. The components are normalized before fusion so that large-system scale does not automatically dominate smaller systems. The exact normalization is fixed before test evaluation and is not tuned using held-out test labels.

4.4 Why the residual is useful against stealthy FDI

A classical estimator residual and the proposed physics residual answer different questions. The former asks whether the measurements remain consistent with the estimator's statistical model. The latter asks whether the reported electrical quantities remain consistent with the trusted physical network equations. An attack can be statistically plausible while being inconsistent with the defender's physical model. Conversely, a genuine topology change or legitimate operating transition can create a physics discrepancy unless the network state and topology are represented correctly. This is why the residual must be interpreted jointly with topology, timing and operating context rather than as an unconditional attack label.

4.5 Expected residual behavior

Condition	Expected residual behavior	Interpretation
Normal / pre-attack	Near numerical precision	Reported state is consistent with trusted network equations.
Stealthy FDI	Substantial increase	Measurement values are statistically plausible but

	relative to normal floor	electrically inconsistent.
Breaker manipulation	Topology-sensitive change	Residual behavior depends on whether detector topology matches physical topology.
Load alteration	Magnitude-dependent change	Residual magnitude can overlap with legitimate operating variation.
Attack end / state restoration	Return toward normal floor	Supports correct temporal localization of the injected interval.

V. BENCHMARK DATASET AND EXPERIMENTAL DESIGN

5.1 Dataset composition

The corrected benchmark contains 2,016 runs across IEEE 39-bus and IEEE 118-bus transmission systems. The design comprises 18 balanced scenario cells, with 112 runs per cell. Each run contains 20 nominal one-second simulation steps. The nine scenario cells per network are normal operation, three stealthy-FDI tiers, coordinated breaker manipulation, and four load-altering conditions (mild/severe surge/drop).

Property	Value
Total simulation runs	2,016
Network cases	IEEE 39-bus; IEEE 118-bus
Scenario cells	18
Runs per cell	112
Time steps per run	20
Nominal time step	1 s
Attack families	Stealthy FDI; breaker manipulation; load alteration
FDI tiers	3
Primary simulator	pandapower
Cross-validation	MATPOWER/Octave; selective GridLAB-D and ns-3; HELICS orchestration/cross-checks
Dataset format	One JSON record per run + manifest index

5.2 Dataset integrity controls

Full-scale validation reported zero schema-conformance defects, zero label-integrity defects, exact target counts for all 18 scenario cells, and no retained NaN electrical states on steps flagged as

converged. Overall non-convergence was approximately 1.02% of simulation steps and was retained as an explicit validity condition rather than silently converted into normal data.

5.3 Leakage correction

The original generator used a fixed attack onset at step 10 and continued attacks to the end of the episode. This produced a serious shortcut: a position-only classifier achieved F1=0.894. The generator was corrected by randomizing attack onset over steps 5–14, bounding duration to 3–8 steps, and restoring the physical state after the attack window. The corrected position-only control achieved F1=0.0 and ROC-AUC=0.5 on the held-out test set. Only corrected results are treated as final evidence.

5.4 Train-validation-test protocol

The run-level split is stratified 60/20/20: 1,224 training runs, 396 validation runs and 396 held-out test runs. Splitting occurs at the run level rather than the individual time-step level, preventing adjacent observations from the same simulation from crossing partitions. The split seed is 42 and the final training seed is 2027. Hyperparameter selection was limited to four candidate configurations and selected using validation F1.

5.5 Model configuration

Component	Configuration
Spatial encoder	Graph Attention Network (GAT), hidden dimension 32
Temporal encoder	LSTM, hidden dimension 64
Physics stream	Eight normalized residual components
Fusion	Concatenation of spatial, temporal and physics representations
Loss	Binary cross entropy with positive weighting
Learning rate	5×10^{-4}
Early stopping	Validation F1
Window	20-time steps
Training seed	2027
Split seed	42

VI. PROPOSED DETECTION ARCHITECTURE

6.1 Spatial representation

At each time step, node and measurement features are passed through a graph-attention encoder. The graph

structure is defined by the transmission network. Attention allows the representation to aggregate information from electrically connected neighboring buses while weighting their contributions adaptively.

$$H_t = \text{GAT}(G, X_t)$$

6.2 Temporal representation

The sequence of spatial representations is passed to an LSTM. Convergence indicators and the derived frequency-deviation proxy are included as temporal channels. The LSTM captures changes that cannot be explained from a single snapshot.

$$T_{1:T} = \text{LSTM}(H_{1:T}, C_{1:T})$$

6.3 Physics representation

$$P_t = \text{Normalize}([r_{P,t}, r_{Q,t}, r_{\text{bal},t}, \dots])$$

The physics branch is deliberately lightweight. It does not learn Y_{bus} , nor does it optimize a differentiable power-flow constraint. It receives residual features calculated from the trusted network model. This design makes the role of physics auditable: a reviewer can inspect the residual values independently of the neural network.

6.4 Fusion and decision

$$Z_t = [\text{Pool}(H_t), T_t, P_t]$$

$$\hat{y}_t = \text{sigmoid}(\text{MLP}(Z_t))$$

The detector outputs an attack probability at the sequence/run level. The operational threshold used for the final evaluation is selected on the validation set and then frozen for the held-out test set. The paper does not claim that this threshold is universally optimal across utility environments.

VII. BASELINES AND EVALUATION METRICS

7.1 Baselines

Two principal baselines are used. The first is a Chi-squared bad-data detector representing a conventional statistical residual approach. The second is a vanilla LSTM using temporal data without explicit graph or physics streams. The comparison is intentionally interpretable: BDD represents classical statistical detection, LSTM represents learned temporal detection, and the proposed architecture adds explicit spatial and electrical-consistency information.

7.2 Metrics

Precision = $TP/(TP+FP)$, measuring the fraction of positive predictions that are correct.

Recall = $TP/(TP+FN)$, measuring the fraction of attacks detected.

$F1 = 2PR/(P+R)$, balancing precision and recall.

ROC-AUC, measuring threshold-independent ranking discrimination.

False-positive rate under normal operation, reported separately by network.

Time-to-detect (TTD), measured in simulation steps from attack onset to first detection within the predefined budget.

Statistical inference is deliberately conservative. The completed detector and ablation results use a single selected training seed. Accordingly, this paper does not fabricate confidence intervals or p-values. Multi-seed replication with run-level paired bootstrap or permutation testing is identified as necessary follow-up work.

VIII. RESULTS

8.1 Pooled held-out performance

Method	Precision	Recall	F1	ROC-AUC
Chi-squared BDD	—	—	0.596	0.569
Vanilla LSTM	—	—	0.621	0.677
Proposed GAT-LSTM + physics residual	0.727	0.715	0.721	0.815

On the corrected held-out test split, the proposed detector achieved $F1=0.721$ and $ROC-AUC=0.815$. The observed F1 increase relative to vanilla LSTM is 10.0 percentage points, and the observed increase relative to BDD is 12.5 points. These are benchmark effects from a single training seed, not population-level statistical estimates.

8.2 FDI-specific performance

Scenario	F1	Recall
Stealthy FDI Tier 1	0.910	0.996
Stealthy FDI Tier 2	0.902	0.974
Stealthy FDI Tier 3	0.892	1.000
Breaker manipulation	0.902	1.000
Load alteration – severe surge	0.737	0.663
Load alteration – severe drop	0.552	0.458
Load alteration – mild surge	0.464	0.370
Load alteration – mild drop	0.342	0.264

The three FDI tiers remain within a narrow F1 range of 0.892–0.910. The Tier-2 case is particularly important because it represents partial attacker knowledge: the proposed detector achieves $F1=0.902$ versus 0.628 for the vanilla LSTM, an observed 27.4-point difference. The result is substantial but should be replicated across independent training seeds before being treated as a generalized effect.

8.3 Physics-stream ablation

Ablation	Observed F1 effect	Interpretation
Remove physics stream – Tier 1	–34.8 points	Strong first-pass evidence of material physics-stream contribution
Remove physics stream – Tier 2	–38.2 points	Strong first-pass evidence of material physics-stream contribution
Remove physics stream – Tier 3	–36.0 points	Strong first-pass evidence of material physics-stream contribution
Remove GNN stream – breaker manipulation	–3.9 points	Below predefined 5-point confirmation threshold

The physics-stream ablation is the strongest component-level evidence in the study. Removing the stream causes a large performance reduction across all FDI tiers. Because the ablation changes one component while keeping the main training protocol fixed, the result supports the interpretation that explicit electrical-consistency information is materially useful. It does not establish that the residual stream is the only useful component or that the same effect will persist on unseen utility networks.

8.4 False-positive analysis

Network	Normal-operation FPR	Interpretation
IEEE 39-bus	0.0%	No false positives observed in the evaluated normal test subset
IEEE 118-bus	25.5%	Substantial topology-dependent calibration problem

The 25.5% false-positive rate on IEEE 118-bus is a central limitation. It prevents a blanket statement that the detector is uniformly calibrated. Possible mechanisms include graph-structure differences, feature scale, residual magnitude distribution, operating-state distribution or threshold calibration, but the current experiment does not establish causality. The correct conclusion is that cross-topology calibration remains unresolved.

8.5 Leakage-control result

Evaluation	Position-only F1	Position-only ROC-AUC
Original fixed-onset dataset	0.894	Not retained as final evidence
Corrected dataset	0.000	0.500

The corrected position-only result is a major validity control. It demonstrates that the final benchmark no longer permits sequence position alone to explain the attack label. The earlier 0.894 result is preserved only as evidence of the defect and the need for correction.

8.6 Time-to-detect

Attack family	Runs meeting 5-step budget	Average TTD
Stealthy FDI	94.8%	Immediate on average in primary pathway
Breaker manipulation	100%	Immediate on average in primary pathway
Load alteration	96.3%	1.49 steps

These latency measurements are simulation-step metrics. Because the nominal time step is one second, they provide an approximate simulation-time interpretation, but they do not represent end-to-end control-center, network, computation or actuation latency.

IX. DISCUSSION

9.1 What the residual evidence establishes

The study provides three mutually reinforcing forms of evidence. First, the residual is independently grounded in the defender-side electrical model rather than being an unconstrained learned feature. Second, the residual responds sharply to stealthy FDI conditions while remaining near the numerical floor

under normal/pre-attack states. Third, removing the residual stream causes a large decline in FDI detection performance. Together these observations support the conclusion that the explicit physics representation is not merely decorative in the current architecture.

9.2 Why ground truth matters

A detector can only be evaluated against a meaningful notion of physical consistency if the benchmark retains the information needed to reconstruct that consistency. The use of Ybus provides such a reference. It permits a researcher to recompute residuals independently and inspect whether an observed anomaly is associated with a physical-model discrepancy. This improves auditability relative to datasets that provide only an attack label.

9.3 Relation to existing detector families

The proposed methodology complements, rather than invalidates, classical BDD. Conventional residuals remain valuable for random or unstructured measurement errors. The contribution here is to add a separate physical-consistency channel targeted at the specific failure mode of stealthy FDI. The use of GAT and LSTM is similarly complementary: graph learning captures topology-dependent relationships, while recurrent learning captures temporal evolution. Recent graph and spatial-temporal FDI detectors demonstrate the value of these representations, but the present study isolates the contribution of an explicit residual stream [6–8].

9.4 Why the 118-bus result matters

A reviewer should not accept a pooled F1 of 0.721 as sufficient evidence of readiness. The topology-stratified false-positive result shows why. A model can perform strongly on attack cases while still generating excessive alarms under normal operation in a larger topology. This finding is scientifically useful because it identifies a concrete next experiment: topology-aware residual normalization and calibration, followed by cross-topology and cross-operating-condition evaluation.

9.5 Interpretation of the GNN ablation

The GNN removal produced a 3.9-point F1 decrease for breaker manipulation, below the predefined 5-point confirmation threshold. Therefore, this paper does not claim that the GNN branch has been

conclusively isolated as an independently necessary component. The appropriate interpretation is that a measurable difference was observed, but the evidence is insufficient for a stronger causal attribution under the single-seed experiment.

9.6 Practical meaning of the method

The method is best understood as a research-grade detection feature rather than an operational protection mechanism. It assumes a trusted electrical model, accurate topology information, appropriate measurement conventions and sufficient computational access to the required electrical variables. A real deployment would additionally require treatment of topology errors, parameter uncertainty, measurement noise, communication loss, latency, model drift and adversarial attempts to manipulate the physical-model inputs.

X. THREATS TO VALIDITY AND LIMITATIONS

- Single training seed: the reported neural-network results are one selected training run and therefore do not establish multi-seed statistical significance.
- Benchmark scope: only IEEE 39-bus and IEEE 118-bus transmission systems are evaluated; utility-specific networks and distribution systems are not covered.
- Topology-dependent calibration: the 118-bus normal-operation FPR of 25.5% demonstrates that the current threshold is not uniformly calibrated.
- Trusted defender model: the residual calculation assumes that Ybus and topology information available to the detector are correct.
- Frequency proxy: frequency is represented by a documented derived proxy rather than direct electromechanical dynamic simulation.
- Communication realism: routine network effects are represented through abstracted communication metadata; this is not equivalent to a full operational communication network.
- No systematic adversarial robustness study: attackers that explicitly optimize against the residual feature were not evaluated.
- No dedicated explainability study: attention weights and residual contributions were not subjected to a formal operator-facing evaluation.

- No live-grid validation: the study is entirely simulation/benchmark based.
- No causal explanation for topology asymmetry: the current evidence identifies the 118-bus FPR problem but does not prove its mechanism.

10.1 Reviewer-facing methodological safeguards

The study addresses several common sources of inflated performance: run-level rather than time-step splitting; randomized attack onset; bounded attack duration; explicit physical state restoration; position-only leakage control; scenario-balanced sampling; full-scale schema and label checks; sign-convention validation; and component ablation. These controls strengthen internal validity, but they do not substitute for independent external validation.

XI. REPRODUCIBILITY PROTOCOL

A reproducible implementation should execute the following sequence: (1) load the IEEE 39-bus and 118-bus benchmark cases; (2) construct the canonical topology and Ybus; (3) generate normal operating trajectories; (4) inject the three FDI tiers and other benchmark attacks; (5) randomize attack onset and duration within the corrected ranges; (6) restore the physical state after the attack interval; (7) solve and validate each step; (8) compute AC and balance residuals; (9) serialize one JSON record per run; (10) create the stratified 60/20/20 run-level split; (11) train the detector using the fixed configuration; (12) freeze the validation-selected threshold; and (13) evaluate once on the held-out test split.

For archival release, the implementation should include the exact software environment, dependency versions, random seeds, configuration files, manifest checksums, residual definitions, unit conventions, sign conventions, and a script that reproduces every table in this paper. The present manuscript does not claim a public DOI or repository unless such archival release has actually occurred.

XII. RESPONSIBLE RESEARCH AND SECURITY CONSIDERATIONS

The experiments are simulation based and use standard benchmark systems. They do not require access to operational grid infrastructure, credentials or proprietary control systems. The attack-generation

procedures are described at the level necessary to reproduce the scientific experiments without presenting operational penetration instructions for a live utility environment. The intended use is defensive research, benchmarking and validation of cyber-physical detection methods.

XIII. CONCLUSION

This paper developed and evaluated a ground-truth physics-residual methodology for detecting stealthy false data injection attacks in smart grids. The central idea is to compute an auditable electrical-consistency representation from the defender-side Ybus model and reported electrical measurements, and then combine that representation with learned spatial and temporal features. On the corrected benchmark, the proposed detector achieved pooled $F1=0.721$ and $ROC-AUC=0.815$, while FDI-tier $F1$ ranged from 0.892 to 0.910. Most importantly, removing the physics-residual stream reduced FDI $F1$ by 34.8–38.2 percentage points across the three attacker tiers, providing strong first-pass evidence that explicit physical consistency contributes materially to detection.

The study also demonstrates why rigorous benchmark validation matters. A fixed attack-onset defect was capable of producing a position-only $F1$ of 0.894; after correction, the same control achieved $F1=0.0$ and $ROC-AUC=0.5$. The final evaluation therefore rests on a leakage-controlled dataset. At the same time, the 25.5% normal-operation false-positive rate on IEEE 118-bus prevents an unconditional claim of uniform calibration or deployment readiness. The most important next steps are multi-seed replication, topology-aware calibration, adversarial robustness against residual-aware attackers, uncertainty analysis, broader network validation and operator-oriented explainability.

The principal contribution is consequently methodological: ground-truth physics residuals provide a transparent and testable information source that complements statistical, spatial and temporal learning for stealthy-FDI detection. The completed results justify further investigation, while the reported limitations define the experiments required before the method can be considered beyond benchmark-level research.

DECLARATIONS

- **Data Availability:** The benchmark is regenerable from the documented simulation pipeline. A public repository/DOI should be cited only after archival release is completed.
- **Code Availability:** The principal dataset-generation and training workflow is documented in the research records. The exact public repository revision should be cited after archival release.
- **Conflict of Interest:** The author declares no conflict of interest.
- **Funding:** No external funding statement was available in the source materials; the final submission should use the verified statement required by the target venue.
- **Ethics Statement:** The study uses simulated power-system models and public benchmark configurations and does not involve human participants or animal subjects.
- **Responsible Disclosure:** The study is simulation-based and does not involve operational grid access, credentials, proprietary control systems or live-grid penetration procedures.

REFERENCES

- [1]. Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” in *Proc. ACM CCS*, 2009, pp. 21–32, doi: 10.1145/1653662.1653666.
- [2]. Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, Art. no. 13, 2011, doi: 10.1145/1952982.1952995.
- [3]. A. Giani, E. B. Farantatos, S. Campbell, and J. D. H. G., “Stealth attacks and protection schemes for state estimators in power systems,” in *Proc. IEEE SmartGridComm*, 2010.
- [4]. A. Teixeira, G. Dán, H. Sandberg, G. Giannakis, and K. H. Johansson, “Cyber security analysis of state estimators in electric power systems,” in *Proc. IEEE CDC*, 2011.
- [5]. J. Yan, H. Guo, and C. Wen, “False data injection against state estimation in power systems with multiple cooperative attackers,” *Electric Power Systems Research*, vol. 101, pp. 225–233, 2020.

- [6]. E. Vincent, M. Korki, M. Seyedmahmoudian, A. Stojcevski, and S. Mekhilef, "Detection of false data injection attacks in cyber-physical systems using graph convolutional network," *Electric Power Systems Research*, vol. 217, Art. no. 109118, 2023, doi: 10.1016/j.epsr.2023.109118.
- [7]. X. Wang, M. Hu, X. Luo, and X. Guan, "A detection model for false data injection attacks in smart grids based on graph spatial features using temporal convolutional neural networks," *Electric Power Systems Research*, vol. 238, Art. no. 111126, 2025, doi: 10.1016/j.epsr.2024.111126.
- [8]. "False data injection attacks detection with modified temporal multi-graph convolutional network in smart grids," *Computers & Security*, vol. 124, Art. no. 103016, 2023, doi: 10.1016/j.cose.2022.103016.
- [9]. "Detection of false data injection attacks in smart grid based on joint dynamic and static state estimation," *IEEE Access*, vol. 11, pp. 45028–45038, 2023, doi: 10.1109/ACCESS.2023.3273730.
- [10]. S. Zhao, W. Luo, Q. Shu, and F. Xu, "Unifying temporal continuity and structural decomposition for stealthy FDIA detection in smart grids via implicit neural representations," *International Journal of Electrical Power & Energy Systems*, vol. 174, Art. no. 111507, 2026, doi: 10.1016/j.ijepes.2025.111507.
- [11]. S. Chaturvedula and P. K. Pareek, "AI-driven threat intelligence for smart grid resilience: A physics-informed anomaly detection and validation framework," doctoral research thesis, Nirwan University, Jaipur, 2026.
- [12]. S. Chaturvedula and P. K. Pareek, "A reproducible software-defined co-simulation benchmark dataset for cyber-physical attacks on smart grids," research manuscript and benchmark records, 2026.