Research Square

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

An AI-Driven Adversarial Attack and Defense Framework for Enhancing Cybersecurity Resilience in Smart Grid Systems

Shuo Li

lishuo202611@yeah.net

State Grid Jibei Information & Telecommunication Company

Jing Zeng

State Grid Jibei Information & Telecommunication Company

Pingzhou Li

State Grid Jibei Electric Power Company Limited

Article

Keywords: smart grid security, adversarial machine learning, intrusion detection, false data injection, adversarial training, robustness

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11518390/v1

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

Additional Declarations: No competing interests reported.

An AI-Driven Adversarial Attack and Defense Framework for Enhancing Cybersecurity Resilience in Smart Grid Systems

Shuo Li[1*], Jing Zeng[1], Pingzhou Li[2]

[1]State Grid Jibei Information & Telecommunication Company, Beijing,100032, China

[2]State Grid Jibei Electric Power Company Limited, Beijing,100032, China

Jing Zeng:760044312@qq.com

Pingzhou Li: li.pingzhou@jibei.sgcc.com.cn

[*]Corresponding author:Shuo Li, lishuo202611@yeah.net

Abstract

Smart grids rely on machine-learning detectors to flag cyber attacks in synchrophasor and control data. These detectors are accurate on clean data, yet their robustness to adversarial manipulation is rarely tested end to end. We build one framework that attacks and defends the same detectors under a single protocol. It runs on real synchrophasor measurements and on a verifiable false data injection simulation. On clean data a random forest reaches 92 percent accuracy on binary attack detection and 83 percent on 37-class detection. It reaches 100 percent on the injection task. White-box attacks drive the neural detector to chance-level accuracy or below. Black-box transfer from an independent surrogate drops the gradient-free random forest from 92 to 31 percent and from 83 to 6 percent. On the injection task it drops from 100 to 50 percent. Among four defenses, only adversarial training survives attacks adapted to it. It holds 99 percent robust accuracy on the injection task. On the binary synchrophasor task its 71 percent only matches the attack base rate. Distillation, ensembles, and feature squeezing fail under adapted attacks. A leakage-robust split lowers synchrophasor accuracy to the base rate, which shows that reported accuracy on this dataset is partly an artifact of correlated samples.

Keywords: smart grid security, adversarial machine learning, intrusion detection, false data injection, adversarial training, robustness.

1 Introduction

Modern power grids depend on networked sensing and control. Phasor measurement units stream synchronized voltage and current measurements, and supervisory control and data acquisition systems drive state estimation and dispatch. This visibility improves efficiency, and it also widens the attack surface [1–3]. A cyber attack that corrupts measurements can hide a line fault, misinform an operator, or trigger an unnecessary trip. Machine-learning detectors are now a common defense. They learn the signature of normal operation from historical measurements and flag deviations as attacks [4–6]. A separate line of work shows that false data injection attacks can be built to pass the classical residual test used in state estimation, which motivates data-driven detection of exactly these stealthy attacks [7, 8].

Machine-learning detectors inherit a weakness of machine learning itself. Small, deliberate input perturbations can flip a classifier decision, which are adversarial examples (Figure 1) [9, 10]. This threat is well studied in general, and smart-grid detectors are exposed to it, which we review in Section 2. The

attacks and defenses have not been evaluated together, under one protocol and with leakage and adaptive controls, on the widely used synchrophasor intrusion dataset. That gap is the target of this paper.

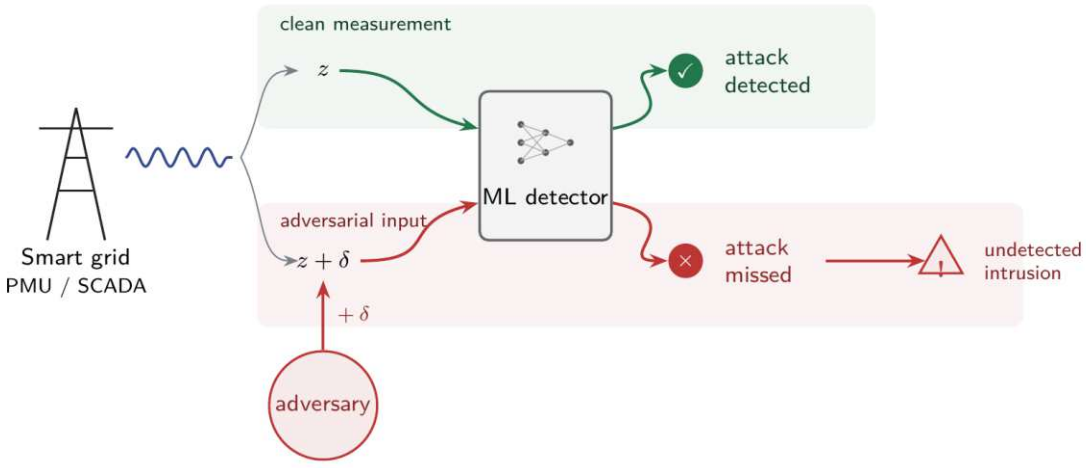


Figure 1: A tiny perturbation evades the detector.

We build and run a unified adversarial attack and defense framework for smart-grid intrusion detection, shown in Figure 2. We evaluate it on two complementary layers. The first is real synchrophasor measurements from a power-system testbed [4, 5]. The second is a false data injection simulation on the IEEE 14-bus system. The contribution is a consolidated benchmark rather than a new attack or defense algorithm.

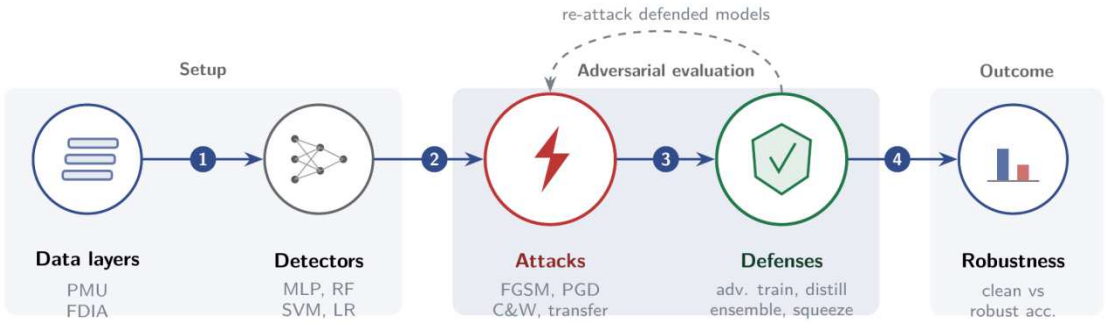


Figure 2: End-to-end attack and defense framework.

Our contributions are as follows.

- We assemble a pipeline that trains four detectors and evaluates four defenses on two smart-grid data layers, with multiple seeds.
- We apply four white-box attacks and a black-box transfer attack, and we quantify how far each degrades detection accuracy.
- We show that black-box transfer alone degrades a gradient-free random-forest detector.
- We evaluate every defense with an attack adapted to it, and report which defenses recover robustness and which only mask gradients.
- We test physically constrained attacks that respect attacker capability on both data layers.

- We add a leakage-robust evaluation of the synchrophasor dataset, and show that a large part of the usual high accuracy reflects within-recording correlation.

2 Related Work

This section reviews three strands of work and then states the gap this paper addresses. The first strand is adversarial machine learning, which supplies the attacks and defenses we study. The second is machine-learning detection in smart grids, which supplies the detectors and the data. The third is the intersection of the two, where attacks and defenses meet in the power grid and where the overlap is still thin.

2.1 Adversarial machine learning

Deep models misclassify inputs that carry small, deliberate perturbations, which are adversarial examples [9–11]. The attack toolkit is mature. The fast gradient sign method perturbs along the loss gradient [10]. The basic iterative method and projected gradient descent refine this into a strong iterative attack [12, 13]. The Carlini and Wagner attack finds small perturbations by optimization [14, 15]. Even without gradients, an adversary can train a surrogate and transfer examples to the target [16]. Defenses include adversarial training [13], defensive distillation [17], and input-level detection such as feature squeezing [18]. Poisoning at training time is a related threat, and robust aggregation can defend federated intrusion detectors against it [19]. We study evasion at inference time instead. Several defenses were later shown to give a false sense of security, because they mask gradients rather than remove the nearby wrong class, and fall to attacks that are adapted to them [20–22]. This is why a careful evaluation uses an attack designed against each defense.

2.2 Machine-learning detection in smart grids

Smart grids widen the cyber attack surface, and surveys catalogue the threats and countermeasures [1–3]. Cryptographic protocols secure authentication in grid communication networks [23, 24]. They do not protect a learned detector whose input measurements are manipulated. False data injection attacks can be built to pass the classical residual test used in state estimation, which is why data-driven detection of these stealthy attacks is needed [7, 25, 26]. Many detectors are machine-learning models trained on measurements [6, 8, 27]. The public Mississippi State University and Oak Ridge National Laboratory synchrophasor dataset is a common benchmark for this task, and it defines the attack, natural, and no-event labels we use [4, 5].

2.3 Adversarial attacks and defenses in grid machine learning

Prior work attacks grid machine learning or pairs one attack with one defense. Evasion attacks have been crafted against neural state-estimation detectors [28, 29], against a power-quality classifier with a matched defense [30], and against a power-system intrusion detector with a single mitigation [31]. A combined attack and defense benchmark exists for an industrial internet-of-things intrusion dataset [32], and data-driven attacks on grid strategies have been generated with adversarial networks [33]. Recent work studies domain-valid attacks on grid intrusion detection [34–36]. We did not find a study that evaluates the standard white-box and black-box attacks together with several defenses, under one protocol and with leakage and adaptive controls, on the widely used synchrophasor intrusion dataset. That gap is the target of this paper.

3 Methodology

This study evaluates a complete attack and defense pipeline on two complementary smart-grid data layers. The first layer is wide-area monitoring, represented by real synchrophasor measurements. The second layer is supervisory control and state estimation, represented by a false data injection simulation on a standard test system. For every detector we measure clean performance, the effect of a family of evasion attacks, and the recovery provided by a family of defenses. All experiments run on a CPU. The main runs use the seed 1337, and the headline metrics are repeated over further seeds.

3.1 Threat model

The defender trains a classifier that labels a measurement vector as normal or as an attack. The adversary observes a trained detector and perturbs the input measurements to force a misclassification, which is an evasion attack at inference time. We study three levels of adversary knowledge. A white-box adversary has the model gradients. A gray-box adversary has a deployed baseline model and attacks a separately defended model with examples crafted on that baseline. A black-box adversary has no gradients and transfers examples from an independently trained surrogate. Perturbation budgets are expressed in standardized feature units, because the synchrophasor features span very different physical scales. We return to the physical plausibility of these perturbations in Section 4.

3.2 Datasets

Power System Attack Dataset. We use the public Mississippi State University and Oak Ridge National Laboratory synchrophasor dataset [4, 5]. It contains 78,377 samples with 128 features recorded from a four-relay power-system testbed. The features are 116 physical synchrophasor and relay measurements, such as voltage and current phasor magnitudes and angles, frequency, rate of change of frequency, and apparent impedance. They also include 12 log features from control-panel, relay, and network intrusion logs. We train on the 116 physical measurement features, because these are the quantities an attacker manipulating grid telemetry would perturb. We treat the 12 log features separately and report a leakage check on them. The binary task separates attack events from natural operation. The multiclass task assigns one of 37 event scenarios. The literal string `Infinity` appears in a small number of impedance cells and becomes a missing value, which we impute with the training-split median. Tables and figures label this dataset Morris, after one of its creators.

IEEE 14-bus false data injection. We build a direct current (DC) state-estimation model for the standard IEEE 14-bus system. The branch reactances are read from the pandapower case rather than typed by hand. The measurement set contains the active power flow on every branch and the net active power injection at every non-reference bus, giving $m = 33$ measurements for $n = 13$ states. Each normal operating point comes from scaling the case loads by a random factor in $[0.7, 1.3]$ and solving a DC power flow, so every state is physically consistent. Measurements are $z = H\theta + \eta$, where H is the measurement matrix, θ is the state, and η is Gaussian meter noise with standard deviation 0.01. A stealthy false data injection adds an attack vector $a = Hc$, where c is a sparse random state shift. By construction the weighted residual of the state estimate is unchanged, so the classical chi-square bad-data test cannot flag the attack. We verify this numerically. The dataset has 12,000 balanced samples. The detector must find the attacks that the residual test misses, because the injected state lies away from the normal operating region. Tables and figures label this task FDIA IEEE-14, where FDIA stands for false data injection attack.

3.3 Detector models

We evaluate four detectors. The differentiable target is a multilayer perceptron (MLP) with two hidden layers of 128 and 64 units, rectified linear activations, and dropout. It operates on standardized features and is trained with the Adam optimizer for 40 epochs. The three baselines are a random forest (RF) with 200 trees, a support vector machine (SVM) with a radial basis kernel, and logistic regression (LR). The support vector machine is trained on a capped subset of 8,000 samples, because a radial basis kernel scales poorly with sample count. The random forest, support vector machine, and logistic regression expose no gradients, so they are attacked only by transfer.

3.4 Adversarial attacks

Figure 3 shows the geometry behind the attacks and the main defense. An evasion attack pushes an input across the decision boundary within a small budget. A transfer attack reuses a perturbation that also crosses a surrogate boundary. Adversarial training moves the boundary beyond the perturbation budget. We craft gradient attacks in the standardized feature space and project each result back into the per-feature value box observed in training.

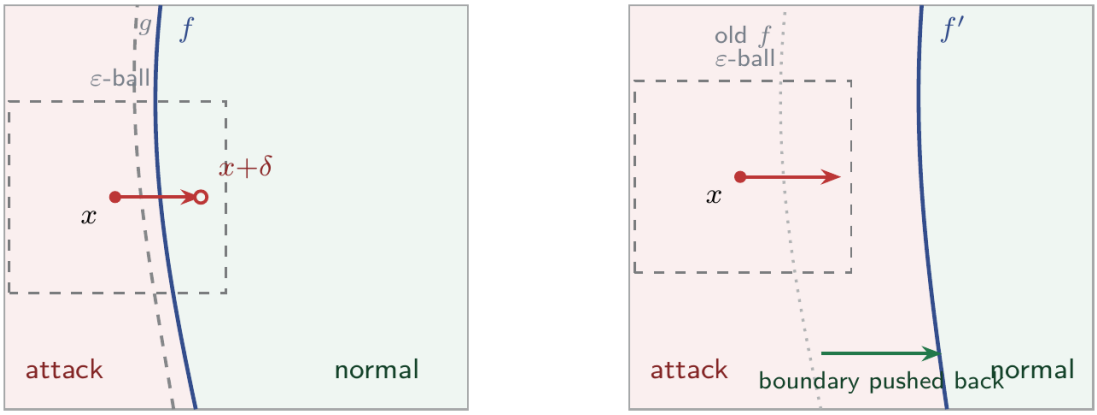


Figure 3: Evasion, transfer, and adversarial training geometry.

The fast gradient sign method (FGSM) takes a single step of size ϵ along the sign of the loss gradient [10]. The basic iterative method (BIM) applies ten smaller signed steps [12]. Projected gradient descent (PGD) applies 20 steps with a random start and a step size of $\epsilon/4$, projected to the L_∞ ball of radius ϵ [13]. We sweep $\epsilon \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$. A budget of ϵ in standardized units changes each feature by at most ϵ of its training standard deviation, so $\epsilon = 0.2$ is a one-fifth standard-deviation change per feature. The Carlini and Wagner attack minimizes an L_2 perturbation subject to misclassification, optimized for 100 steps in a hyperbolic-tangent box parameterization on a fixed subset of 1,000 test samples [14]. The black-box transfer attack crafts projected gradient descent examples on an independently trained surrogate network and evaluates them on the random forest, which shares no parameters with the surrogate [16].

We also test two physically-constrained attacks, because an unconstrained feature-space perturbation may not be realizable [35, 36]. For the synchrophasor task we restrict the perturbation support to the measurements of a single relay, which models an attacker who has compromised one phasor measurement unit. For the false data injection task we force the perturbation to lie in the column space of the

measurement matrix, written $\delta = Hc$. The weighted residual is then unchanged, so the attack stays invisible to the chi-square test while it tries to fool the detector. We optimize the state offset c and bound the per-measurement change to 0.1 per unit.

3.5 Defenses

Adversarial training (AT) augments every training batch with projected gradient descent examples at $\epsilon = 0.2$ using seven steps, so the model sees clean and adversarial inputs together [13]. Defensive distillation trains a teacher at temperature 20 and trains a student on the teacher soft labels at the same temperature [17]. The ensemble averages the output probabilities of five networks trained with different seeds, a defense whose limits against transferred attacks are known [22]. Feature squeezing quantizes each input to 3, 4, or 5 bits within the training value box, applied as a preprocessing step before the baseline network [18]. We also test two adversarial-input detectors. The first scores the L_1 distance between the softmax on an input and on its squeezed version. The second scores the pairwise disagreement among the ensemble members. We report the area under the receiver operating characteristic curve for separating clean from attacked inputs.

3.6 Evaluation protocol and leakage controls

We use two splits for the Power System Attack Dataset and report both as primary. The first is a stratified 70/30 train and test partition. The second is a leakage-robust split that trains on source files 1 to 11 and tests on files 12 to 15, so no two samples from the same recording run fall on both sides. We also train a detector on the 12 log features alone, to measure how much those features reveal the label. The false data injection task uses the stratified split, since it is balanced and fully synthetic.

Because accuracy alone is misleading under class imbalance, we report accuracy, balanced accuracy, macro-averaged F1, and per-class precision and recall [37]. Attack strength uses robust accuracy (racc), which is accuracy on adversarial inputs, and attack success rate (ASR), which is the fraction of originally correct samples that an attack flips. To separate genuine robustness from gradient masking, each defense is also evaluated with an attack adapted to it [20, 21]. These are a white-box attack on the adversarially trained and ensemble models, a temperature-corrected attack on the distilled model, and a backward-pass differentiable approximation (BPDA) attack on feature squeezing. Headline metrics are repeated over three seeds for the synchrophasor tasks and five seeds for the injection task, and reported as the mean and standard deviation. All training-set statistics used for standardization, imputation, and the value box are computed on the training split only.

3.7 Implementation

The pipeline runs on a CPU in Python 3.12 with PyTorch 2.14, scikit-learn 1.9, NumPy 2.4, SciPy, and pandapower 3.5.

4 Results and Discussion

This section reports the experiments and interprets them. We present clean detection first, then the effect of each attack, then the recovery provided by each defense, then a leakage-robust re-evaluation of the synchrophasor dataset. We then discuss why the attacks and defenses behave as they do, the threats to validity, and the limitations. Robust accuracy is accuracy on adversarial inputs. Attack success rate is the fraction of originally correct samples that an attack flips. Tables report the seed 1337 runs unless stated otherwise. Table 3 gives the headline metrics as the mean and standard deviation over several seeds.

4.1 Clean detection and the need for learning-based detection

Table 1 gives clean performance. The random forest is the strongest detector on every task. It reaches 92.1% accuracy and 89.9% macro-F1 on the binary synchrophasor task, 82.8% accuracy and 81.1% macro-F1 on the 37-class task, and 100% accuracy on the false data injection task. The neural network is weaker on the synchrophasor data, with 75.2% accuracy on the binary task and 29.0% on the multiclass task. This reflects the known advantage of tree ensembles over multilayer perceptrons on tabular features. Logistic regression reaches only 51.8% on the balanced false data injection task, which shows that those attacks are not linearly separable.

The false data injection attacks are stealthy by construction. The classical chi-square bad-data test uses a 95% threshold of 31.4 at 20 degrees of freedom. It flags only 4.8% of the stealthy attacks, while it flags 100% of random injections of the same magnitude (Figure 5, right). The mean weighted residual is 19.97 for attacked samples and 20.09 for normal samples, so the attacks are invisible to the residual test. This is the gap that a learning-based detector is meant to close, and the random forest closes it with 100% clean accuracy.

Table 1**:** Clean detection accuracy and macro-F1 (percent).

Dataset	MLP acc	MLP F1	RF acc	SVM acc	LR acc
Morris binary	75.2	61.1	92.1	71.4	71.1
Morris multiclass	29.0	21.7	82.8	11.8	23.7
FDIA IEEE-14	99.9	99.9	100.0	99.6	51.8

Accuracy alone is misleading on the binary synchrophasor task, where attacks are 71% of the samples. Table 2 reports balanced accuracy and per-class precision and recall. The random forest reaches 88.1% balanced accuracy and detects the minority natural class at 78.4% recall. The neural network reaches only 60.6% balanced accuracy and 25.9% recall on the natural class, so its 75.2% accuracy comes mostly from predicting the majority attack class. These headline numbers are stable across seeds (Table 3). The random forest clean accuracy varies by 0.04 points and the neural network projected gradient descent accuracy by 2.3 points across seeds.

The injection results are not an artifact of one setting. Table 4 sweeps meter noise and attack strength. Random forest accuracy stays between 99.8% and 100%, neural network accuracy between 99.6% and 99.9%, and the chi-square detection of the stealthy attack stays at 4.6%. The learning-based detector separates the attacks while the residual test does not, across the range tested.

Table 2**:** Imbalance-aware metrics, binary task (percent).

Model	Acc	Bal. acc	Macro-F1	Attack P/R	Natural P/R
Random forest	92.1	88.1	89.9	91.7/97.7	93.3/78.4
MLP	75.2	60.6	61.1	75.9/95.4	69.6/25.9

Table 3: Headline metrics over seeds, mean and standard deviation (percent).

Dataset	seeds	RF clean	MLP clean	MLP PGD	RF transfer	AT clean	AT PGD
Morris binary	3	92.2±0.0	75.3±0.1	10.7±2.3	30.1±3.0	72.1±0.1	70.3±0.3
Morris multiclass	3	82.9±0.1	29.0±0.3	1.0±0.1	5.6±0.8	–	–
FDIA IEEE-14	5	100.0±0.0	99.9±0.0	49.4±0.1	50.0±0.0	99.9±0.0	99.3±0.0

Table 4: Injection results over noise and attack strength (percent).

Meter noise	Attack c	RF acc	MLP acc	χ^2 stealth detect
0.01	0.1	99.9	99.9	4.6
0.01	0.25	100.0	99.9	4.6
0.01	0.5	100.0	99.9	4.6
0.02	0.1	99.9	99.8	4.6
0.02	0.25	100.0	99.9	4.6
0.02	0.5	100.0	99.9	4.6
0.05	0.1	99.8	99.6	4.6
0.05	0.25	100.0	99.9	4.6
0.05	0.5	100.0	99.9	4.6

4.2 White-box attacks collapse the neural detector

A white-box adversary with the network gradients reduces the neural detector to chance-level accuracy or below. Figure 4 shows robust accuracy against the perturbation budget ϵ for the fast gradient sign method and for projected gradient descent. On the binary synchrophasor task, projected gradient descent at $\epsilon = 0.05$ already cuts robust accuracy from 75.2% to 32.3%, and at $\epsilon = 0.2$ to 9.7% (Table 5). On the multiclass task the detector falls to 0.9% at $\epsilon = 0.2$. On the false data injection task robust accuracy reaches the 50% chance level at $\epsilon = 0.2$. The Carlini and Wagner attack reaches an attack success rate of 65.4% on the binary task, 85.1% on the multiclass task, and 57.1% on the false data injection task. Projected gradient descent is the strongest attack in every table, as expected from its iterative first-order design.

Table 5: Neural detector robust accuracy under white-box attacks, binary task.

Attack	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.5$
FGSM	45.2	29.3	14.4	9.1	6.5
BIM	38.1	23.0	12.6	9.3	6.6
PGD	32.3	18.4	9.7	7.2	5.6

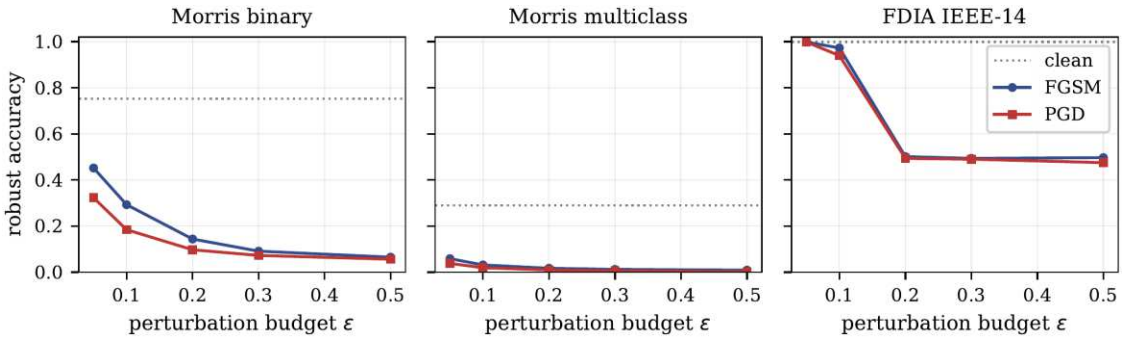


Figure 4: Robust accuracy versus perturbation budget.

4.3 Black-box transfer degrades gradient-free detectors

The random forest exposes no gradients, yet it is not safe. We craft projected gradient descent examples on an independent surrogate network at $\epsilon = 0.2$ and evaluate them on the random forest, which shares no parameters with the surrogate. Transfer alone drops the random forest from 92.1% to 30.8% on the binary task and from 82.8% to 6.1% on the multiclass task. It drops from 100% to 50.0% on the false data injection task (Figure 5, left, and Table 6). The corresponding attack success rates are 67.5%, 93.0%, and 50.0%. A detector that is accurate and gradient-free is therefore still exposed to an adversary who only trains a substitute model.

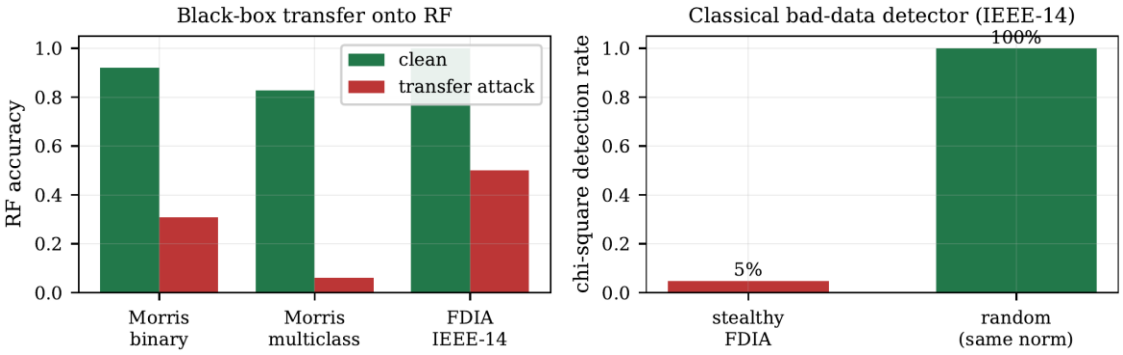


Figure 5: Transfer vulnerability and false data injection stealth.

Table 6: Black-box transfer onto the random forest and detection AUC.

Dataset	RF clean	RF racc (transfer)	RF ASR	Squeeze AUC	Ensemble AUC
Morris binary	92.1	30.8	67.5	0.17	0.51
Morris multiclass	82.8	6.1	93.0	0.39	0.96
FDIA IEEE-14	100.0	50.0	50.0	0.40	0.65

4.4 Physically constrained attacks still degrade the detector

An unconstrained feature-space perturbation may not be realizable, so we test two attacks that respect attacker capability (Table 7). On the synchrophasor task we allow the attacker to perturb only the 29 measurements of a single relay. This single-relay attack drops the neural detector from 75.2% to 34.8%, against 9.7% for the unconstrained attack. On the injection task we force the perturbation into the stealth subspace, so the residual is unchanged. This stealth-constrained attack drops the detector from 99.9% to

55.5%, and the chi-square test still flags only 4.8% of these attacks. A realizable, stealthy perturbation therefore still removes most of the detection.

Table 7**:** Robust accuracy under physically constrained attacks (percent).

Setting	Constraint	Robust acc
Morris binary	clean	75.2
unconstrained PGD	all 116 feat.	9.7
single-PMU PGD	29 feat.	34.8
FDIA IEEE-14	clean	99.9
unconstrained PGD	free	49.3
stealth-constrained	$\delta = Hc$	55.5

4.5 Only adversarial training recovers robustness

Table 8 and Figure 6 compare the defenses at $\epsilon = 0.2$. Adversarial training is the one defense whose robust accuracy stays high. On the false data injection task it raises robust accuracy from 49.3% to 99.5%, with a clean accuracy of 99.8% and 99.3% under white-box re-attack. This is genuine robustness on a balanced task. On the binary task it raises projected gradient descent robust accuracy from 9.7% to 71.5%, at a clean-accuracy cost of 3.2 percentage points, and it holds 70.2% under a white-box re-attack on the hardened model. These values sit at the 71% attack base rate, which a detector that always predicts attack would also reach. The other defenses do not hold. Defensive distillation gives 17.5% robust accuracy on the binary task and 54.9% on the false data injection task, which matches the known result that its gradient masking is broken by an iterative attack. The five-model ensemble gives 12.5% and 49.7%. Feature squeezing at three bits reaches 25.4% on the binary task and lower at higher bit depths.

Two caveats are important. First, on the binary task the adversarially trained model has a macro-F1 of 46.6%, below the undefended 61.1%. A constant attack prediction would give a macro-F1 of about 41.5%. Most of its binary robustness therefore comes from predicting the majority attack class rather than from better discrimination. Second, adversarial training does not help the weak multiclass detector. Its clean accuracy there falls from 29.0% to 18.1% and its robust accuracy is 12.4%, so a detector that is already poor cannot be made robust by this procedure.

The two adversarial-input detectors are weak in most settings. The feature-squeezing detector never exceeds an area under the curve of 0.40. The ensemble-disagreement detector reaches 0.959 on the multiclass task, where attacks induce large disagreement among members, but only 0.512 on the binary task and 0.652 on the false data injection task (Table 6). Detection is therefore task-dependent and not a general safeguard.

Table 8: Defense clean accuracy and robust accuracy at $\epsilon = 0.2$ (percent).

Defense	Clean acc	FGSM racc	PGD racc
Morris binary ($\epsilon = 0.2$)			
Undefended MLP	75.2	14.4	9.7
Adversarial training	72.0	71.3	71.5
Defensive distillation	73.4	20.9	17.5
Ensemble (5)	75.3	17.2	12.5
Morris multiclass ($\epsilon = 0.2$)			
Undefended MLP	29.0	1.7	0.9
Adversarial training	18.1	15.1	12.4
Defensive distillation	22.9	3.7	2.0
Ensemble (5)	30.3	2.2	1.5
FDIA IEEE-14 ($\epsilon = 0.2$)			
Undefended MLP	99.9	50.1	49.3
Adversarial training	99.8	99.6	99.5
Defensive distillation	99.9	71.0	54.9
Ensemble (5)	99.9	52.9	49.7

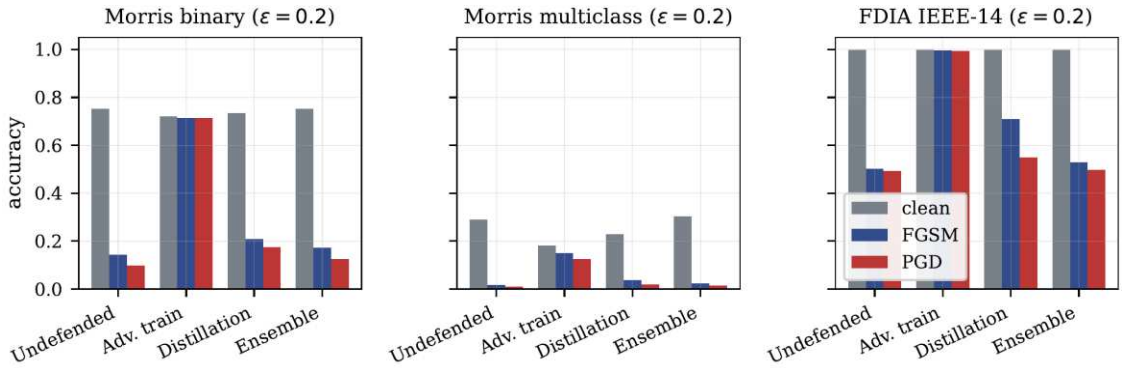


Figure 6: Clean and robust accuracy by defense.

4.6 Adaptive attacks confirm the defense ranking

The numbers above use a gray-box attack crafted on the baseline. A defense can look robust against that attack while failing against an attack designed for it, so we re-evaluate each defense adaptively (Table 9 and Figure 7). Adversarial training holds: its white-box robust accuracy is 99.3% on the injection task and 70.2% on the binary task, the latter at the base rate. The other defenses fall. On the injection task a naive white-box attack on the distilled model reports 89.9% robust accuracy, which is a false reading caused by gradient masking. A temperature-corrected attack recovers the gradient and drops it to 49.4%, the chance level. The adaptive ensemble attack and the backward-pass attack on feature squeezing reach 49.6% and 49.8% on the injection task, and 5.0% and 6.0% on the binary task. Only adversarial training survives an attack built against it.

Table 9: Clean and adaptive robust accuracy by defense (percent).

Defense	Clean acc	Adaptive robust acc
Morris binary		
Adversarial training	72.0	70.2
Defensive distillation (naive)	73.4	12.7
Defensive distillation (adaptive)	73.4	6.3
Ensemble (5)	75.3	5.0
Feature squeezing (BPDA)	–	6.0
FDIA IEEE-14		
Adversarial training	99.8	99.3
Defensive distillation (naive)	99.9	89.9
Defensive distillation (adaptive)	99.9	49.4
Ensemble (5)	99.9	49.6
Feature squeezing (BPDA)	–	49.8

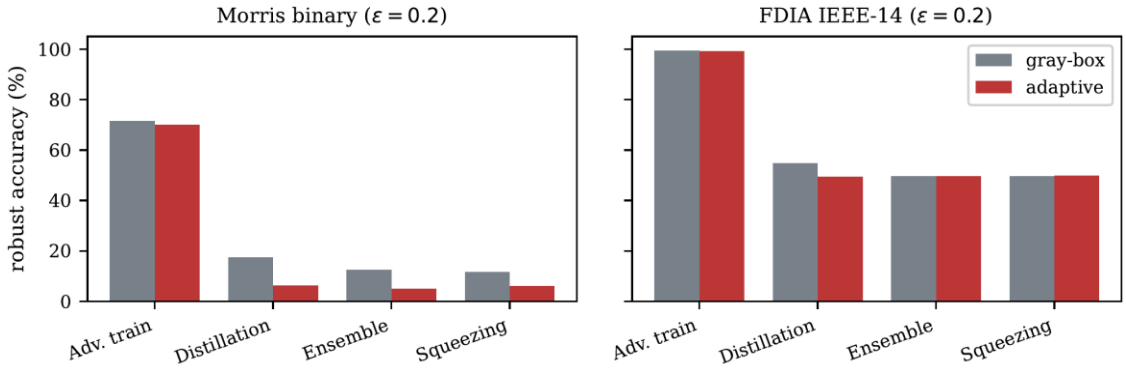


Figure 7: Gray-box versus adaptive robust accuracy.

4.7 Leakage-robust evaluation tempers the reported accuracy

The synchrophasor samples form short recorded runs, so a random split can place correlated neighbours on both sides. Table 10 reports a leakage-robust split that trains on files 1 to 11 and tests on files 12 to 15. Binary accuracy then falls to 71.4% for the neural network and 70.7% for the random forest, close to the 71% attack base rate, against 92.1% for the random forest on the random split. Much of the usual high accuracy on this dataset therefore reflects within-recording correlation rather than transferable detection skill. A separate check trains on the 12 log features alone. This reaches only 71.5% on the binary task, the base rate, and 7.8% on the multiclass task, which confirms that the log features carry little label information and justifies training on the physical measurements.

The attack and defense findings hold on this leakage-robust split (Table 11). Projected gradient descent still drops the neural detector to 17.2%, and transfer still drops the random forest to 31.4%. The vulnerability is therefore not an artifact of the random split. Adversarial training gives 70.0% robust accuracy at a clean accuracy of 71.3%. Both values sit at the base rate, so on this split it prevents the collapse but adds no detection skill.

Table 10**:** Leakage checks on the synchrophasor dataset (percent).

Check	Accuracy
Log-features-only RF (binary)	71.5
Log-features-only RF (multiclass)	7.8
By-file split MLP (binary)	71.4
By-file split RF (binary)	70.7

Table 11**:** Attacks and defense on the leakage-robust split (percent).

Metric (by-file split)	Value
RF clean accuracy	70.7
MLP clean accuracy	71.4
MLP PGD robust accuracy	17.2
RF transfer robust accuracy	31.4
Adv. training clean / PGD	71.3 / 70.0

4.8 Why the attacks and defenses behave as they do

The results give a consistent picture, and the reasons are structural. Learning-based detectors are accurate on clean smart-grid data and fragile under adversarial perturbation. The neural detector collapses under white-box attacks on every task, and the gradient-free random forest collapses under black-box transfer. A trained detector places its decision boundary close to the data, so a small move along the loss gradient crosses it. Projected gradient descent follows this gradient over several steps, which is why it is the strongest attack in every table. Transfer succeeds because independent models trained on the same measurements learn similar boundaries, so a perturbation that crosses the surrogate boundary tends to cross the target boundary as well. This is why a random forest with no gradients still loses most of its accuracy. The effect is largest on the multiclass task, where many classes create many nearby boundaries for a perturbation to cross. A detector being a tree ensemble, or being accurate on clean data, is not evidence that it is safe against an adversary.

Adversarial training works because it reshapes the loss so that the gradient near each training point no longer points at a nearby wrong class. The cost appears in two places. On the binary synchrophasor task the hardened model gains robust accuracy but loses macro-F1, which means it leans toward the majority attack class rather than discriminating better. On the multiclass task the model is too weak to begin with, so adversarial training lowers both its clean and its robust accuracy. Robustness is therefore not free, and it cannot repair a detector that does not fit the clean task. The other defenses fail for known reasons. Defensive distillation hides the gradient but does not remove the nearby wrong class, so an adaptive attack recovers the direction and the robustness disappears. The ensemble averages models that share the same vulnerable directions, so averaging does not remove them. Feature squeezing quantizes continuous phasor measurements, which changes the input too little to undo a perturbation that already crossed the boundary.

4.9 Threats to validity

Two measurement choices shape how the numbers should be read, and we address both directly. First, perturbation budgets are set in standardized feature units. When mapped back to raw measurements, a

budget that is small in standardized units can be large on high-magnitude channels such as bus voltages. A real attacker may not control every synchrophasor channel independently, so the standardized-budget attacks are an upper bound on what an unconstrained adversary achieves. We therefore test constraint-aware attacks that respect attacker capability, and find that a single-relay perturbation and a stealth-preserving injection both still degrade the detector. Second, the false data injection data is a simulation on the IEEE 14-bus system. It is valuable because the stealthiness of each attack is exact and verifiable against the chi-square test, but it is not a field deployment, and its clean accuracies near 100% reflect that controlled setting. The sensitivity sweep shows how accuracy changes as meter noise and attack strength vary.

The leakage-robust evaluation carries its own message. On a split that keeps each recorded run on one side, binary accuracy on the synchrophasor data falls to the base rate. Reported accuracies on this dataset, including some in the literature, are likely inflated by within-run correlation. A security claim built on a random split can overstate how well a detector will generalize to a new event, and this holds before any adversary is considered. Benchmarks on this dataset should report a grouped split alongside the random split, which matches wider guidance on avoiding data leakage in security machine learning [38].

4.10 Limitations

This study has clear limits. It runs on a CPU, which bounds model size, the number of Carlini and Wagner iterations, and the size of the attacked subset. It uses one real dataset and one simulation, and a single neural architecture as the differentiable target. It does not evaluate certified defenses. On the binary synchrophasor task no tested defense yields robust detection above the base rate.

5 Conclusion

This work places the standard white-box and black-box attacks and four defenses on the same smart-grid detectors under one protocol. It adds leakage controls, adaptive-attack evaluation, physically constrained attacks, and a verifiable stealth model. Accurate detectors on both data layers collapse under attack. White-box attacks drive the neural detector to chance or below. Black-box transfer drives the gradient-free random forest from high accuracy to well below the base rate. Among the defenses, only adversarial training survives adapted attacks. It gives genuine robustness on the injection task. On the binary synchrophasor task its robust accuracy only matches the base rate, and it does not help the under-fit multiclass detector. Distillation, ensembles, and feature squeezing fall under attacks adapted to them. A leakage-robust split lowers synchrophasor accuracy to the base rate, so part of the usual reported accuracy is an artifact of correlated samples. Resilient smart-grid detection will need defense in depth, physically constrained tests, and evaluation that does not rely on correlated splits.

Data availability

The Power System Attack Dataset analysed here is the public Mississippi State University and Oak Ridge National Laboratory synchrophasor dataset [4, 5]. It is publicly available from the dataset authors at https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets (Dataset 1, Power System Datasets). The IEEE 14-bus false data injection dataset is generated by the simulation described in Section 3, using the standard IEEE 14-bus case distributed with the open-source pandapower library.

Author contributions

To be completed by the authors.

Competing interests

The authors declare no competing interests.

Declarations

Funding:not applicable.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable. This study did not involve human participants, human data, human tissue or animals.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

Clinical trial number: not applicable.

References

[1] Wenye Wang and Zhuo Lu. Cyber security in the smart grid: Survey and challenges. *Computer Networks*, 57 (5): 1344–1371, 2013. doi: 10.1016/j.comnet.2012.12.017.

[2] Zakaria El Mrabet, Naima Kaabouch, Hassan El Ghazi, and Hamid El Ghazi. Cyber-security in smart grid: Survey and challenges. *Computers and Electrical Engineering*, 67: 469–482, 2018. doi: 10.1016/j.compeleceng.2018.01.015.

[3] Chih-Che Sun, Adam Hahn, and Chen-Ching Liu. Cyber security of a power grid: State-of-the-art. *International Journal of Electrical Power and Energy Systems*, 99: 45–56, 2018. doi: 10.1016/j.ijepes.2017.12.020.

[4] Raymond C. Borges Hink, Justin M. Beaver, Mark A. Buckner, Tommy Morris, Uttam Adhikari, and Shengyi Pan. Machine learning for power system disturbance and cyber-attack discrimination. In *2014 7th International Symposium on Resilient Control Systems (ISRCS)*, pages 1–8. IEEE, 2014. doi: 10.1109/ISRCS.2014.6900095.

[5] Shengyi Pan, Thomas Morris, and Uttam Adhikari. Developing a hybrid intrusion detection system using data mining for power systems. *IEEE Transactions on Smart Grid*, 6 (6): 3104–3113, 2015. doi: 10.1109/TSG.2015.2409775.

[6] Mete Ozay, Inaki Esnaola, Fatos Tunay Yarman Vural, Sanjeev R. Kulkarni, and H. Vincent Poor. Machine learning methods for attack detection in the smart grid. *IEEE Transactions on Neural Networks and Learning Systems*, 27 (8): 1773–1786, 2016. doi: 10.1109/TNNLS.2015.2404803.

[7] Yao Liu, Peng Ning, and Michael K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14 (1): 1–33, 2011. doi: 10.1145/1952982.1952995.

[8] Youbiao He, Gihan J. Mendis, and Jin Wei. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. *IEEE Transactions on Smart Grid*, 8 (5): 2505–2516, 2017. doi: 10.1109/TSG.2017.2703842.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[11] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84: 317–331, 2018. doi: 10.1016/j.patcog.2018.07.023.

[12] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In Roman V. Yampolskiy, editor, *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. doi: 10.1201/9781351251389-8.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[14] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. doi: 10.1109/SP.2017.49.

[15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582. IEEE, 2016. doi: 10.1109/CVPR.2016.282.

[16] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS)*, pages 506–519. ACM, 2017. doi: 10.1145/3052973.3053009.

[17] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. doi: 10.1109/SP.2016.41.

[18] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the 25th Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2018.

[19] Muhammad Ilyas, Nadeem Iqbal Kajla, Michael G. Madden, Gaojuan Fan, Chongsheng Zhang, and Hafiz Muhammad Sanaullah Badar. PoisonShield-FL-NIDS: A robust defense against poisoning attacks in federated learning intrusion detection. *IEEE Internet of Things Journal*, 13 (4): 6105–6115, 2026. doi: 10.1109/JIOT.2025.3634357.

[20] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283. PMLR, 2018.

[21] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 1633–1645. Curran Associates, 2020. doi: 10.5555/3495724.3495862.

[22] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations (ICLR)*, 2018.

[23] Hafiz Muhammad Sanaullah Badar, Khalid Mahmood, Waseem Akram, Zahid Ghaffar, Muhammad Umar, and Ashok Kumar Das. Secure authentication protocol for home area network in smart grid-based smart cities. *Computers and Electrical Engineering*, 108: 108721, 2023. doi: 10.1016/j.compeleceng.2023.108721.

[24] Hafiz Muhammad Sanaullah Badar, Shafiq Ahmed, Nadeem Iqbal Kajla, Gaojuan Fan, and Chongsheng Zhang. Q-BLAISE: Quantum-resilient blockchain and AI-enhanced security protocol for smart grid IoT. *IEEE Transactions on Consumer Electronics*, 71 (2): 4959–4971, 2025. doi: 10.1109/TCE.2025.3570969.

[25] Ahmed S. Musleh, Guo Chen, and Zhao Yang Dong. A survey on the detection algorithms for false data injection attacks in smart grids. *IEEE Transactions on Smart Grid*, 11 (3): 2218–2234, 2020. doi: 10.1109/TSG.2019.2949998.

[26] Ali Sayghe, Yaodan Hu, Ioannis Zografopoulos, XiaoRui Liu, Raj Gautam Dutta, Yier Jin, and Charalambos Konstantinou. Survey of machine learning methods for detecting false data injection attacks in power systems. *IET Smart Grid*, 3 (5): 581–595, 2020a. doi: 10.1049/iet-stg.2020.0015.

[27] Mohammad Esmalifalak, Lanchao Liu, Nam Nguyen, Rong Zheng, and Zhu Han. Detecting stealthy false data injection using machine learning in smart grid. *IEEE Systems Journal*, 11 (3): 1644–1652, 2017. doi: 10.1109/JSYST.2014.2341597.

[28] Ali Sayghe, Junbo Zhao, and Charalambos Konstantinou. Evasion attacks with adversarial deep learning against power system state estimation. In *2020 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2020b. doi: 10.1109/PESGM41954.2020.9281719.

[29] Jiwei Tian, Buhong Wang, Jing Li, Zhen Wang, Bowen Ma, and Mete Ozay. Exploring targeted and stealthy false data injection attacks via adversarial machine learning. *IEEE Internet of Things Journal*, 9 (15): 14116–14125, 2022a. doi: 10.1109/JIOT.2022.3147040.

[30] Jiwei Tian, Buhong Wang, Jing Li, and Zhen Wang. Adversarial attacks and defense for CNN based power quality recognition in smart grid. *IEEE Transactions on Network Science and Engineering*, 9 (2): 807–819, 2022b. doi: 10.1109/TNSE.2021.3135565.

[31] Rong Huang and Yuancheng Li. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. *IEEE Transactions on Smart Grid*, 14 (3): 2367–2376, 2023. doi: 10.1109/TSG.2022.3217060.

[32] Shiming Li, Jingxuan Wang, Yuhe Wang, Guohui Zhou, and Yan Zhao. EIFDAA: Evaluation of an IDS with function-discarding adversarial attacks in the IIoT. *Heliyon*, 9 (2): e13520, 2023. doi: 10.1016/j.heliyon.2023.e13520.

[33] Zengji Liu, Qi Wang, Yujian Ye, and Yi Tang. A GAN-based data injection attack method on data-driven strategies in power systems. *IEEE Transactions on Smart Grid*, 13 (4): 3203–3213, 2022. doi: 10.1109/TSG.2022.3159842.

[34] Nelson Makau Mutua, Simin Nadjm-Tehrani, and Petr Matoušek. Penetrating the power grid: Realistic adversarial attacks on smart grid intrusion detection systems. In *Critical Information Infrastructures Security (CRITIS 2024)*, Lecture Notes in Computer Science, pages 249–268. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-84260-3_15.

[35] Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni. Modeling realistic adversarial attacks against network intrusion detection systems. *Digital Threats: Research and Practice*, 3 (3): 1–19, 2022. doi: 10.1145/3469659.

[36] Nour Alhussien, Ahmed Aleroud, Abdullah Melhem, and Samer Y. Khamaiseh. Constraining adversarial attacks on network intrusion detection systems: Transferability and defense analysis. *IEEE Transactions on Network and Service Management*, 21 (3): 2751–2772, 2024. doi: 10.1109/TNSM.2024.3357316.

[37] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124. IEEE, 2010. doi: 10.1109/ICPR.2010.764.

[38] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3971–3988. USENIX Association, 2022.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- smartgridadvoverleaf.zip