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.

A Deep Learning-Based Adversarial Attack Detection and Defense Method for Cybersecurity Protection in Power Systems

Shuo Li

lishuo202610@yeah.net

State Grid Jibei Information & Telecommunication Company

Xiangsen Xu

State Grid Jibei Information & Telecommunication Company

Pingzhou Li

State Grid Jibei Electric Power Company Limited

Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11314898/v1

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

Additional Declarations: No competing interests reported.

A Deep Learning-Based Adversarial Attack Detection and Defense Method for Cybersecurity Protection in Power Systems

Shuo Li[1*], Xiangsen Xu[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

Shuo Li: lishuo202610@yeah.net

Xiangsen Xu: xuxeon@126.com

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

[*]corresponding：Shuo Li,lishuo202610@yeah.net

Abstract

Deep learning detectors are proposed to catch false data injection in power system state estimation, but they can themselves be attacked. We study an attacker who keeps the measurements physically consistent, so the residual test stays silent. It must also fake a minimum load change at a target bus. We propose PARRY, a two-layer detection and defence method. The first is a deep detector trained against this realisable attacker. The second is a far-bus witness that predicts each bus load from buses outside the attacker's reach. We prove that the witness bounds the load change an undetected attack can fake, however it is optimised and however slowly it is applied. State re-estimation adds a small slack to this bound, which we measure. We evaluate on IEEE 14-bus and 118-bus systems with three years of real demand. Thresholds target a nominal 1% false alarm rate, and attacks are adaptive and white-box. On the 118-bus system a standard deep detector catches 19% of attacks that fake a 30% load change. Adversarial training with PGD or TRADES catches 36% or 32%. PARRY catches 75%, and 98% at a 50% change. No method detects 10% changes, which lie within natural load variation.

Introduction

A control centre estimates the state of its grid from thousands of meter readings and acts on that estimate. Economic dispatch, contingency analysis and protection settings all rest on it. False data injection exploits this dependence. If an attacker alters readings so that they remain consistent with the network equations, the weighted least squares residual does not grow and the bad data test stays silent[1–3]. The 2015 attack on the Ukrainian grid showed that intruders can reach the operational network and stay in it for months[4].

Deep learning detectors have become the main answer to such attacks. They learn what normal measurements look like and flag the rest. They are, however, trained and tested against attacks drawn at random. An attacker who already controls meters can do better. It can choose its injection against the detector, keep the readings physically consistent so the residual test stays silent, and still move the estimated load where it wants. Robustness tools borrowed from image classification do not describe this attacker, because they perturb the measurement vector inside a norm ball and so break the network equations.

This paper asks how far a deep detector can be pushed by an attacker that is both realisable and adaptive, and what a defence can guarantee against it. Figure 1 summarises the setting and the two layers of our method. Our contributions are as follows.

- We formulate a realisable evasion attack. It passes the residual test, it must fake a minimum load change at a target bus, and it controls only the meters around that bus.
- We propose PARRY, a two-layer method. Its deep detector is trained against the realisable attacker, and its far-bus witness predicts each bus load from buses the attacker cannot touch.
- We prove that the witness caps the load change any undetected attack can fake, for any optimisation strategy and any attack duration. The cap is exact for consistent estimates and carries a small measured slack under re-estimation.
- We evaluate PARRY against seven baselines on two IEEE systems driven by three years of real demand. Attacks are random, transferred and adaptive white-box, and slow ramps are included.

The rest of the paper reviews related work, describes the methodology, reports the results, discusses them and concludes.

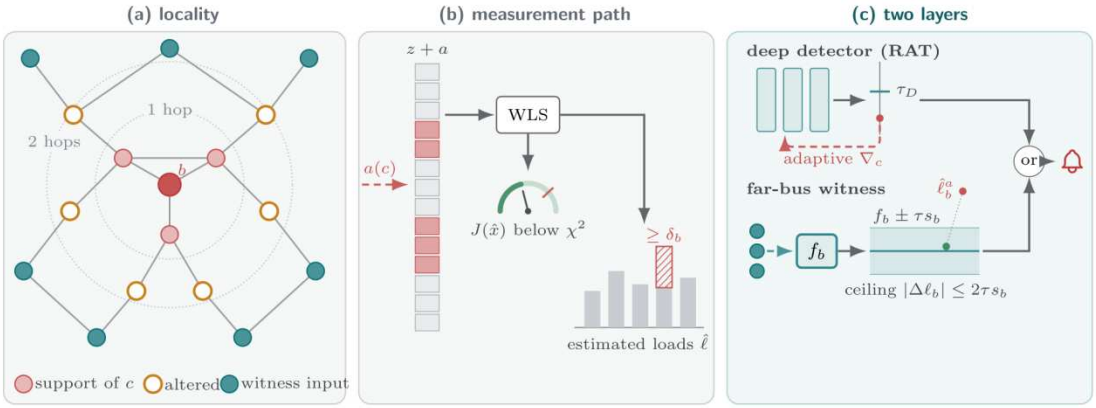


Figure 1. Attack locality, the measurement path and the two layers of PARRY.

Related work

Three lines of work meet in this paper: learned detection of false data injection, adversarial attacks on such detectors, and defences that hold under adaptive attack.

Learned detection of false data injection

Stealthy injection was first shown for DC state estimation[1,2] and later for the nonlinear AC model[3]. Surveys cover its construction, its physical and economic impact and the detectors proposed against it[5–7]. Learned detectors range from supervised and semi-supervised classifiers[8,9] to deep belief and recurrent networks that use measurement history[10,11], multilabel localisers[12] and graph neural networks that use topology[13,14]. A graph autoencoder has been hardened against poisoned training data[15], and random matrix methods separate replay attacks from injection[16]. Reviews note that robustness of these detectors to a deliberate adversary is still open[17,18]. Detection complements preventive security. Identity-based and physically unclonable function authentication[19,20], access control for grid infrastructure[21] and quantum-

resilient, learning-assisted protocols for grid devices[22] make it harder to take over meters. Our threat model starts where these layers have failed for a few meters.

Adversarial attacks on grid learning

Deep networks can be fooled by small inputs chosen against them[23–25], and optimisation attacks make this systematic[26,27]. In power systems, adversarial inputs degrade load forecasters[28] and evade learned state estimators and detectors[29,30]. Joint attacks that are stealthy both to the residual test and to a deep detector have been built[31,32], and constrained attacks respect the physics of cyber-physical systems[33]. Recent work extends evasion to detectors in protection relays[34]. Our attacker follows this line but adds an explicit impact floor, so an evasion counts only if it moves the estimated load by a set amount.

Defences under adaptive attack

Adversarial training with projected gradient descent[35] and the TRADES trade-off[36] are the standard empirical defences. Input detectors flag adversarial examples by their feature statistics[37,38] or by prediction changes under squeezing[39]. Several studies have shown that such defences fail once the attacker knows them[40–42]. Certified defences bound the effect of any perturbation inside a norm ball[43,44], but a norm ball is not the threat in a grid. For grid detectors, resilient training[45], mitigation of attacks on learned detectors[46] and robust reinforcement learning[47] have been proposed, mostly evaluated against non-adaptive or norm-bounded attacks. Related work on learned intrusion and anomaly detection defends the training stage against poisoned updates[48] or learns normal behaviour without attack labels[49]; both address threats to the detector other than evasion at test time. On the control side, moving target defences perturb line parameters to expose stealthy injection[50–52], and forecasting-aided estimation compares measurements with predictions from history[53,54]. Our witness is complementary to both. It needs no change to the physical system and it does not trust history, which matters against slow attacks.

Methodology

This section defines the grids and data, the estimator and its bad data test, the attacker, the detectors and their training, the far-bus witness with its impact ceiling, and the evaluation protocol. Figure 2 shows the inner attack, the adversarial training batch and the calibration schedule.

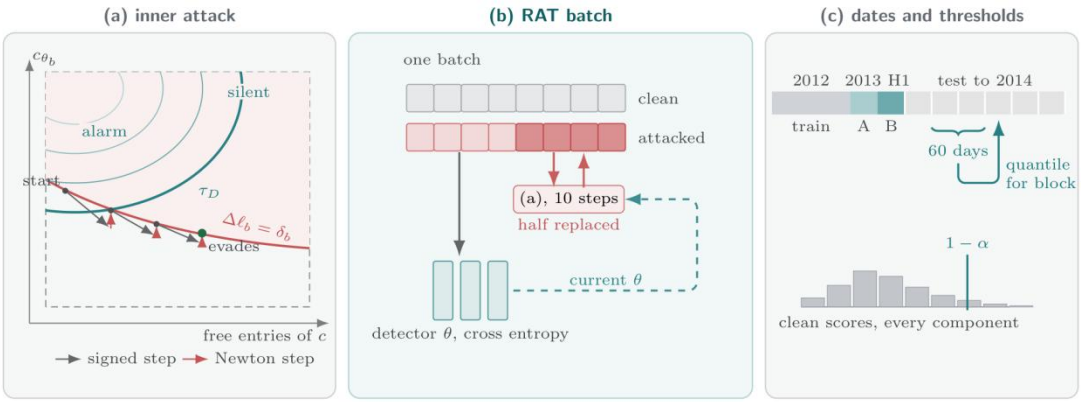


Figure 2. Inner attack, adversarial training batch and rolling thresholds.

Grids, demand and measurements

We use the IEEE 14-bus and IEEE 118-bus test systems as distributed with pandapower[55]. The demand is real. Each load of a grid is fed by the sum of 60 clients drawn at random from the UCI ElectricityLoadDiagrams set[56], which records 15-minute consumption of Portuguese clients from 2011 to 2014. We keep the 312 clients with a complete record from 2012 to 2014. The draws are independent across loads, so neighbouring loads share some clients and co-move as loads in one region do. Each aggregate is scaled so that its mean is 0.7 times the nominal load of the case. Generator set points follow the total demand in proportion and the slack bus covers losses. An AC power flow is solved for each of the 105,216 steps of the three years; every step converged. The physics is therefore simulated while the demand pattern is measured.

The meter set holds every bus voltage magnitude, every active and reactive bus injection, and the active and reactive flow at the sending end of every branch. This gives $m = 82$ readings on the 14-bus grid and $m = 726$ on the 118-bus grid. Readings carry independent Gaussian noise with standard deviation 0.002 per unit for magnitudes and 0.005 per unit (0.5 MW on the 100 MVA base) for powers. Each noise seed yields a different measurement series of the same operating points.

State estimation and bad data detection

The state $x \in \mathbb{R}^n$ holds every bus angle except the slack angle and every voltage magnitude ($n = 27$ and $n = 235$). With $h(\cdot)$ the AC measurement function and $W = \mathrm{diag}(\sigma_i^{-2})$, the control centre computes the weighted least squares estimate

$$\hat{x} = \operatorname*{argmin}_x J(x), \qquad J(x) = (z - h(x))^\top W (z - h(x)), \tag{1}$$

by Gauss–Newton iterations from a flat start[57,58]. Bad data detection raises an alarm when $J(\hat{x})$ exceeds the $1 - \alpha$ quantile of a χ^2 distribution with $m - n$ degrees of freedom. On clean test data its false alarm rate matches α (Table 1).

Threat model

The attacker wants the control centre to believe in a false load. It picks a target load bus b and a direction, and must change the estimated load at b by at least δ_b, a fraction of the mean load of b. Only buses with a mean load of at least 10 MW are targets (5 buses on the 14-bus grid, 81 on the 118-bus grid). The attacker knows the current estimate $\hat{x}$ and the network model, which is the usual assumption for stealthy injection[2,3]. It chooses a state offset c whose support is the angle and magnitude of every bus within r hops of b, and it injects

$$a = h(\hat{x} + c) - h(\hat{x}). \tag{2}$$

The tampered readings $z + a$ are those of a physically valid operating point, so the residual, and with it the bad data test, is unchanged to first order. The attacker must control every meter whose reading depends on the support. For radius 0 this is 16 meters on average on both grids and for radius 1 it is 31 to 32 meters, and for radius 2 it is 48.5 meters on the 14-bus grid and 57.9 on the 118-bus grid. Each entry of c is bounded by 0.15 rad for angles and 0.03 per unit for magnitudes.

The impact constraint is what separates this attacker from an unconstrained adversarial example. A perturbation that fools the detector but moves no load is harmless, and a perturbation that moves load but breaks consistency is caught by the residual test. The attacker therefore solves

$$\min_{c \in \mathcal{C}_b} s(z + a(c), \hat{x} + c) \quad \text{subject to} \quad \pm\left(\ell_b(\hat{x} + c) - \ell_b(\hat{x})\right) \geq \delta_b, \tag{3}$$

where ℓ_b is the estimated load at b, $\mathcal{C}_b$ is the box on the support and s is the margin of the defence being attacked (positive means an alarm). We solve it by projected signed gradient descent[35]. After every step a Newton correction on the target angle restores the impact constraint with equality, and only iterates that meet it are kept. The first restart begins at the smallest offset that meets the constraint and later restarts begin at random feasible points. The main runs use 80 steps and 2 restarts; a separate check with 400 steps and 5 restarts measures how far the attack is from convergence.

We evaluate four attacks. *Random* adds uniform jitter of up to a tenth of the box width to every free entry of the smallest feasible offset and then meets the constraint again; the same draw generates the attacked training examples. *Floor* moves only the target angle. *Transfer* solves (3) against an independently trained surrogate detector and replays the result. *Adaptive* solves (3) against the full defence, including every secondary detector it contains, with the maximum over components smoothed by a log-sum-exp for the gradient[40–42]. Every verdict is taken on a fresh WLS estimate of $z + a$, so the bad data test and the witness see what a control centre would see. Detection rates count only attack instances that meet their impact floor; the share that does not is reported.

Deep detector and training

The detector is a residual multilayer perceptron on standardised readings with three blocks of width 256 and layer normalisation. It outputs an attack logit. Training uses the 2012 data: every clean step once per epoch and an equal number of attacked steps (random attacks as defined above) with random targets, radius 0 or 1, and impact fractions drawn uniformly from 0.05 to 0.8. We compare four training modes for 20 epochs each. *Standard* training uses random attacks. *PGD-AT*[35] and *TRADES*[36] perturb every standardised input inside an ℓ_∞ ball of radius 0.25 with 7 steps; these perturbations are not physically realisable. *Realisable adversarial training* (RAT), our training mode, replaces half of the attacked examples in each batch by solutions of (3) against the current model (10 steps), so the inner adversary lives in the same set as the attacker at test time.

Two input detectors from the adversarial example literature are attached to the standard detector as secondary alarms: the Mahalanobis distance of the penultimate features to the clean class[37] and feature squeezing[39] with standardised readings rounded to a grid of 0.5, whose rounding passes the gradient straight through for the adaptive attacker.

Temporal forecast gate

As the forecasting-aided baseline[53,54] we forecast the vector of estimated bus injections at horizons of 1, 4, 16 and 96 steps. For horizon k a network predicts the value at t from the four values ending at $t - k$, the values one day and one week earlier, and the calendar, as persistence plus a learned correction. The score is the largest deviation over injections in robust units, and the gate fires when any horizon exceeds its calibrated quantile. Its history contains past estimates, so an attacker who has been active for longer than a horizon has also shaped that horizon's forecast.

Far-bus witness

The witness relies on locality. In (2) the support of c lies within r hops of b. The injection at a bus depends only on the state of that bus and of its neighbours, so the attack changes injections only within $r + 1$ hops of b. The witness of a watched bus b predicts its active and reactive injection at time t from the injections of all buses at least $\bar{r} + 2$ hops from b at the same time, plus the calendar. We set $\bar{r} = 1$. One network with masked inputs and a learned bus embedding serves every watched bus; it is trained on 2012. Its residual for bus b is $e_b = \hat{y}_b - f_b$, where $\hat{y}_b$ is the estimated active injection, and its score is $\max_b |e_b|/s_b$, with s_b a robust scale fitted on the first half of 2013.

Proposition 1 (Impact ceiling). *Let an attack of the form (2) have radius $r \le \bar{r}$ around target b. Let f_b and f_b^a be the witness predictions for b from the clean and the attacked estimate, and let $\eta \ge 0$ satisfy $|f_b^a - f_b| \le \eta s_b$. Let τ be the witness threshold and let E be the event that the clean score at time t is at most τ. On E, if the witness stays silent, the false load change at b satisfies $|\Delta\ell_b| \le (2\tau + \eta)s_b$. If the estimate under attack equals $\hat{x} + c$, then $\eta = 0$. If the clean score at t and the calibration scores are exchangeable and τ is their empirical $1 - \alpha_w$ quantile, then* $\Pr(E) \ge 1 - \alpha_w$.

Proof. Silence gives $|\hat{y}_b^a - f_b^a| \le \tau s_b$ and E gives $|\hat{y}_b - f_b| \le \tau s_b$. The triangle inequality through f_b^a and f_b bounds $|\hat{y}_b^a - \hat{y}_b|$, which equals $|\Delta\ell_b|$, by $(2\tau + \eta)s_b$. If the estimate under attack is $\hat{x} + c$, every witness input of b is an injection at a bus at least $\bar{r} + 2 > r + 1$ hops from b, which (2) leaves unchanged, so $f_b^a = f_b$ and $\eta = 0$. The probability statement is the split conformal argument for a calibrated quantile. □

The ceiling holds for any choice of c and any attack duration, because neither appears in the argument. Four caveats apply. First, the guarantee is exact only in the idealised case $\eta = 0$. A fresh WLS estimate differs slightly from $\hat{x} + c$, which moves far injections a little, so the operational bound carries the slack η. We measure η for every attack, including attacks optimised against the witness itself (Results). Second, load series are not exchangeable over months, so $\Pr(E) \ge 1 - \alpha_w$ is a calibration target rather than a guarantee, and we report the realised false alarm rate. Third, the guarantee covers radius at most $\bar{r}$; we test radius 2, whose footprint reaches the witness inputs, as a stress case. Fourth, the thresholds rely on clean reference data, discussed below.

PARRY raises an alarm when the RAT detector or the witness fires. Its nominal false alarm rate α is split as 0.8α for the detector and 0.2α for the witness. Every other two-part defence (standard with Mahalanobis, standard with squeezing, standard with witness) uses the same split.

Thresholds and protocol

Data are split by date: 2012 for training, January to June 2013 for calibration (the first half fits robust scales, the second half seeds thresholds), and July 2013 to December 2014 for testing. Every threshold is the empirical quantile of clean scores over the 60 days before the current 30-day block. The nominal false alarm rate is $\alpha = 0.01$ for every defence. It is a calibration target: the thresholds aim at it, and the realised rate on the test period is measured and reported, not assumed.

Rolling recalibration assumes that the reference window is clean. In operation this requires a vetted reference stream, such as periods confirmed by post-event audit, by an independent measurement channel, or by field inspection. Because such vetting may be unavailable or imperfect, we also evaluate two

alternatives on the same trained models. *Static* thresholds are set once on the second half of the calibration period, which precedes deployment, and are never updated. *Contaminated* rolling thresholds use reference windows in which 1% or 5% of the steps carry random realisable attacks (radius 1, false load changes of 10% to 80%). For each variant the adaptive attack is re-optimised against that variant's own thresholds. For each grid, seed, radius (0, 1, 2) and impact fraction (0.1, 0.2, 0.3, 0.5, 0.8) we draw 400 test steps with random targets and directions; all defences face the same problems. Ramps start at 150 random test steps, reach their final offset in $R \in \{1,4,16,64,256\}$ equal steps and hold it for one day. The final offset is optimised against the snapshot detector of the defence, and every tampered estimate enters the gate history. Three seeds vary the noise, the attack draws and the network initialisation. Training and attacks run on one GPU.

Results

This section first checks the threat model, then measures how deep detectors fare against adaptive attacks, how realisable adversarial training and the witness change that picture, what the witness guarantees, how slow ramps behave and what all of it costs in false alarms. Unless stated otherwise, numbers are means over three seeds, thresholds target a nominal false alarm rate of 1% with rolling recalibration, and the attack radius is 1.

The realisable attacker passes the residual test

The threat model rests on one property, and the data confirm it. Over all attack instances, the bad data test flags 0.69% of realisable attacks on the 14-bus grid and 0.92% on the 118-bus grid, which matches its 1% clean rate (Table 1). An ℓ_∞ perturbation of radius 0.25 standard deviations, computed against the standard detector, is flagged by the bad data test in 100% of cases on both grids. Every attack met its impact floor within the 2% tolerance of the optimiser, and after re-estimation every attack still produced at least 90% of its floor. A radius-0 attack controls 16 meters on average on both grids, and a radius-1 attack controls 31 to 32. The fresh WLS estimate under attack differs from $\hat{x} + c$ by at most 0.001 rad.

Table 1. Clean false alarm rate (%) on the 18-month test period, nominal rate 1%.

	IEEE 14-bus			IEEE 118-bus		
Defence	rolling	static	worst month	rolling	static	worst month
Bad data test	1.01			0.98		
Standard	1.03	0.99	2.29	1.11	1.25	2.17
PGD-AT	1.06	1.14	2.76	1.03	1.05	2.53
TRADES	1.07	1.18	3.06	1.16	1.50	4.34
Std + Mahalanobis	0.88	0.79	1.86	0.90	1.01	1.88
Std + squeezing	1.07	0.96	2.24	1.15	1.24	2.19
RAT	1.09	1.05	3.20	1.12	1.08	2.34
Temporal gate	1.63	1.26	18.87	1.52	1.53	15.16
Witness	1.51	2.17	4.56	1.71	2.16	11.24
Std + witness	1.19	1.21	2.26	1.35	1.77	3.74
PARRY	1.22	1.27	2.78	1.35	1.65	3.80

Deep detectors fail against adaptive attacks

Every detector looks strong against random attacks. On the 118-bus grid at a 30% false load change, every defence catches 98% to 100% of random attacks (Fig. 4). The picture changes when the attack is

chosen against the defence. The standard detector then catches 19%, and attacks crafted against an independently trained surrogate still evade it 65% of the time (detection 0.35). Even the floor attack, which moves only the target angle and is not optimised at all, evades it 58% of the time.

The usual defences do not close the gap (Fig. 3, Table 2). PGD adversarial training raises detection at 30% to 0.36 and TRADES to 0.32. Their ℓ_∞ perturbations do not resemble realisable attacks, and at an 80% change both fall below the detection rate of a Mahalanobis input detector attached to the standard model. That detector reaches 0.38 at 30%. Feature squeezing lowers detection to 0.15; the share of the nominal rate it takes from the detector appears to cost more than it adds.

The 14-bus grid is easier. Its few load buses make a large false change visible, and every deep detector reaches at least 0.89 at 30%. The difference lies at small changes. At 10% the standard detector catches 0.36 and PGD adversarial training 0.28.

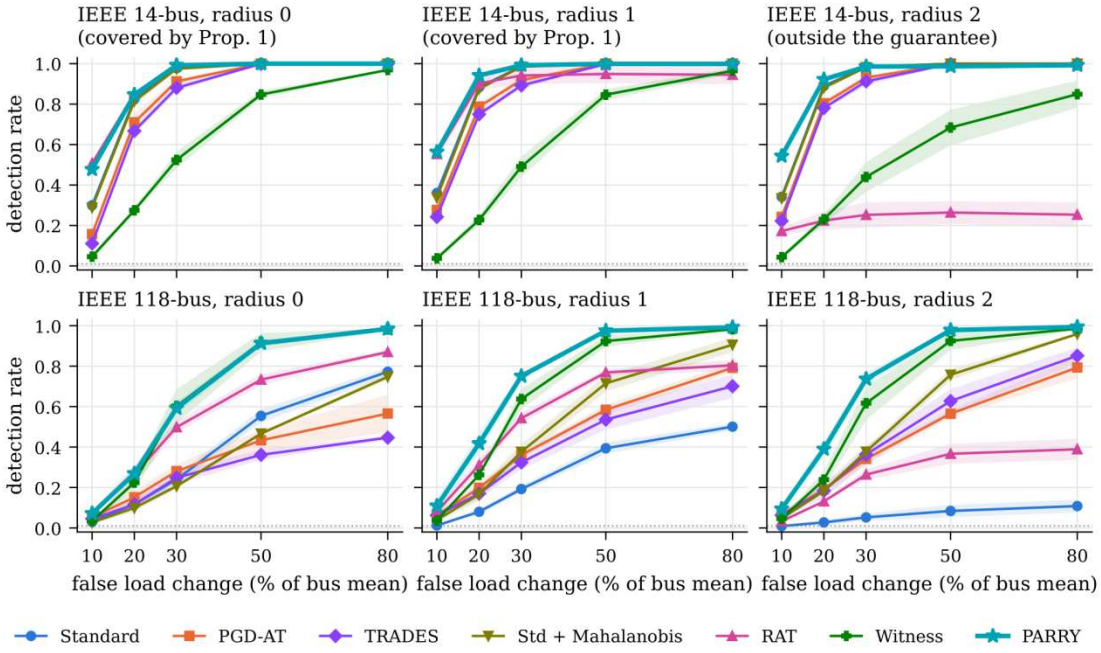


Figure 3. Detection under the adaptive attack versus false load change.

Realisable training and the witness

Training against the attacker that will be faced helps more than training against an abstract ball. On the 118-bus grid, RAT raises adaptive detection at 30% from 0.19 to 0.54, and at 50% from 0.39 to 0.77. The witness alone reaches 0.64 at 30% and 0.92 at 50%, and its rate does not depend on how the attack was optimised. PARRY combines them and catches 0.42, 0.75 and 0.98 of adaptive attacks at 20%, 30% and 50%. Against transferred attacks it reaches 0.78.

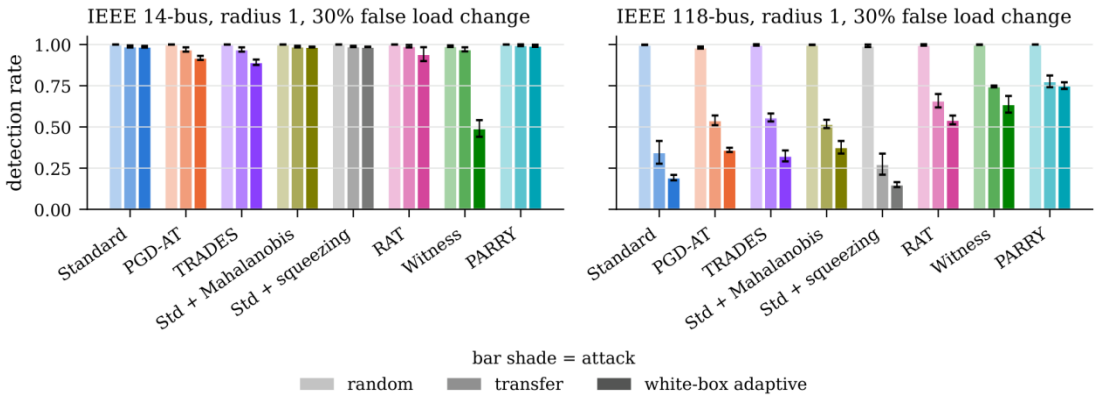


Figure 4. Random, transferred and adaptive attacks at a 30% false load change.

Table 2. Detection rate under the adaptive attack, radius 1 (mean ± s.d. over three seeds).

	IEEE 14-bus: false load change					IEEE 118-bus: false load change				
Defence	10%	20%	30%	50%	80%	10%	20%	30%	50%	80%
Standard	0.36±0.00	0.88±0.01	0.99±0.01	**1.00**±0.00	**1.00**±0.00	0.01±0.01	0.08±0.01	0.19±0.02	0.39±0.03	0.50±0.02
PGD-AT	0.28±0.02	0.79±0.02	0.92±0.01	**1.00**±0.00	**1.00**±0.00	0.06±0.02	0.20±0.03	0.36±0.01	0.58±0.03	0.79±0.02
TRADES	0.24±0.01	0.75±0.01	0.89±0.02	**1.00**±0.00	**1.00**±0.00	0.06±0.02	0.17±0.03	0.32±0.03	0.54±0.05	0.70±0.06
Std + Mahalanobis	0.34±0.01	0.88±0.01	0.99±0.00	**1.00**±0.00	**1.00**±0.00	0.04±0.01	0.17±0.03	0.38±0.04	0.71±0.04	0.91±0.04
Std + squeezing	0.34±0.01	0.88±0.01	0.99±0.00	**1.00**±0.00	**1.00**±0.00	0.01±0.00	0.05±0.01	0.15±0.02	0.33±0.02	0.41±0.03
RAT	0.55±0.03	0.90±0.03	0.94±0.04	0.95±0.05	0.95±0.05	0.08±0.02	0.31±0.04	0.54±0.03	0.77±0.02	0.80±0.03
Temporal gate	0.04±0.01	0.33±0.03	0.59±0.01	0.92±0.02	0.99±0.00	0.04±0.01	0.14±0.01	0.35±0.02	0.77±0.04	0.96±0.02
Witness	0.04±0.01	0.23±0.02	0.49±0.05	0.85±0.03	0.97±0.01	0.04±0.00	0.26±0.00	0.64±0.05	0.92±0.03	0.98±0.01
Std + witness	0.34±0.01	0.87±0.01	0.98±0.01	**1.00**±0.00	**1.00**±0.00	0.05±0.01	0.25±0.02	0.60±0.01	0.93±0.02	0.99±0.00
PARRY	**0.56**±0.00	**0.94**±0.00	**0.99**±0.01	**1.00**±0.00	**1.00**±0.00	**0.11**±0.03	**0.42**±0.05	**0.75**±0.02	**0.98**±0.01	**0.99**±0.00

On the 14-bus grid, RAT lifts detection at 10% from 0.36 to 0.55 and PARRY reaches 0.56. RAT alone, however, falls to 0.95 for changes of 50% and above, while the standard detector reaches 1.00. Training on adaptive examples has shifted its boundary away from very large changes. PARRY recovers 1.00 there because the witness catches what RAT lets through. The witness alone is weak on this grid, with 0.49 at 30%, because each target has only a handful of buses three or more hops away.

At 10% no defence exceeds 0.11 on the 118-bus grid. A false change of a tenth of the load at one bus is smaller than the variation of that load around anything the rest of the grid can predict.

The witness ceiling and its slack

Proposition 1 turns the witness threshold into a ceiling on the false load change. With the 0.2% share of the nominal rate it has inside PARRY, the median ceiling over target buses is 66% of the bus mean load on the 118-bus grid, with a tenth percentile of 49% and a ninetieth of 98% (Fig. 5a). Used alone at the full 1% budget, the median falls to 52%. On the 14-bus grid the median is 83%. These ceilings are loose. They are, however, the only quantity in this study that no attacker strategy can move.

The exact ceiling assumes that far injections are untouched, and WLS re-estimation breaks this slightly. Across all radius-1 attacks the largest change of a far injection after re-estimation was 0.21 robust units on the 118-bus grid and 0.12 on the 14-bus grid. What matters for the bound is the resulting shift of the

target's witness prediction, the slack η of Proposition 1. Over every attack type, radius 0 and 1, five impact levels and three seeds, η never exceeded 0.148 on the 14-bus grid and 0.022 on the 118-bus grid. That is at most 1% of the term 2τ, with τ near 7.6 and 7.3 robust units. Attacks optimised against the witness itself, which gain most from this slack, reached 0.019 and below 0.001. We also checked the corrected bound directly. In all 229,597 attack instances in which the witness stayed silent and the clean residual was within its threshold, the false load change stayed within $(2\tau + \eta)s_b$.

Radius-2 attacks break that assumption on purpose. They control 48.5 meters on average on the 14-bus grid and 57.9 on the 118-bus grid, and they move far injections by up to 778 and 6261 robust units, so Proposition 1 no longer applies. The witness then degrades on the small grid, from 0.85 to 0.68 at a 50% change, but on the 118-bus grid it still detects 0.62 at 30% against 0.64 at radius 1 (Fig. 5b). The deep detectors suffer more. On the 118-bus grid RAT falls from 0.54 to 0.26 at 30% and the standard detector from 0.19 to 0.05. On the 14-bus grid RAT, which was trained on radius 0 and 1 only, detects between 0.17 and 0.26 at every change size, below the standard detector. PARRY keeps 0.74 at 30% on the 118-bus grid and about 0.99 from 30% upward on the 14-bus grid.

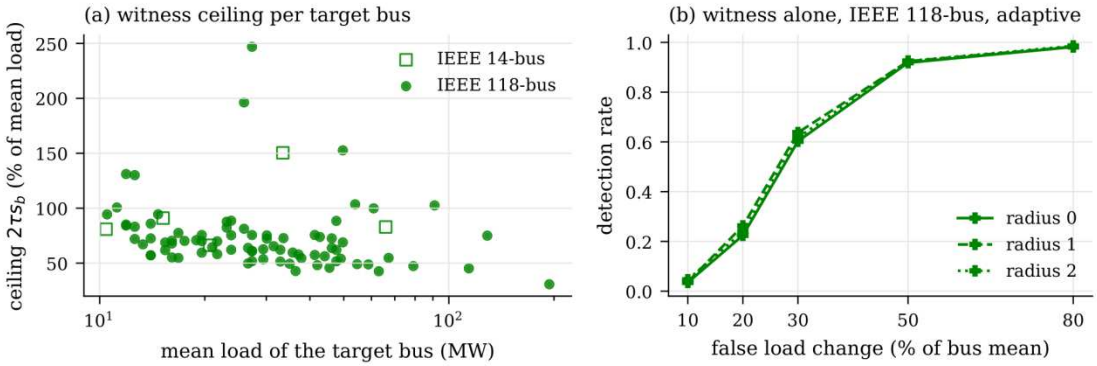


Figure 5. Witness ceiling per bus and witness detection by attack radius.

Slow ramps poison history but not the witness

A patient attacker can reach its final offset in many small steps, and each tampered estimate then enters the history of any temporal detector. We test ramps of 1 to 256 steps followed by a one-day hold (Fig. 6). On the 118-bus grid at an 80% change, the temporal gate flags 0.91 of the hold steps after a one-step jump but only 0.22 after a 256-step ramp. The witness reads only contemporaneous far injections and flags 0.99 of hold steps at every ramp length. During the ramp itself its rate falls from 0.99 to 0.66, because the offset is still small. PARRY flags at least 0.99 of steps throughout. The snapshot detectors do not depend on ramp length, as expected, so the standard detector stays at 0.57 and RAT at 0.80 to 0.81.

False alarms over eighteen months

The 1% figure is a calibration target, not a guarantee, so we report the realised rate for every defence (Table 1). With thresholds recalibrated every 30 days on the previous 60 days, the deep detectors stay between 0.88% and 1.16% on average. The witness and the temporal gate drift more, to 1.51% to 1.71%, and PARRY inherits part of this, reaching 1.22% and 1.35%. The worst single month is the clearer warning. For the temporal gate it reaches 18.87% on the 14-bus grid and 15.16% on the 118-bus grid, and

for the witness 11.24% on the 118-bus grid. For PARRY it is 2.78% and 3.80%. Thresholds fixed after the first half of 2013 would give PARRY 1.27% and 1.65%.

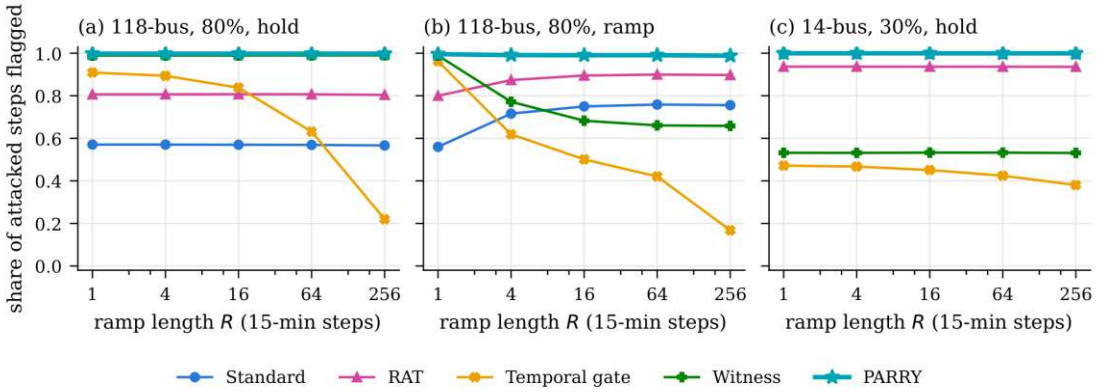


Figure 6. Detection of slow ramps during and after the ramp.

Calibration without a vetted reference stream

Rolling recalibration assumes that the reference window is clean, so we replaced it on the same trained models. Fixing the thresholds once on pre-deployment data removes the assumption at little cost. On the 118-bus grid PARRY then detects 0.76 of adaptive attacks at 30%, against 0.75 with rolling thresholds, at a realised false alarm rate of 1.6% against 1.3%. RAT alone loses more on the 14-bus grid, from 0.94 to 0.83 at 30%. Contaminated recalibration is far worse. When 1% of the reference steps carry attacks, the thresholds rise towards attack scores, the realised false alarm rate of PARRY falls to 0.2%, and its detection at 30% on the 118-bus grid falls to 0.12. With 5% contamination it is 0.04. Every defence fails in the same way, because a contamination share at the nominal rate places the threshold quantile among attack scores. Rolling recalibration is therefore safe only on vetted data.

The adaptive attack has converged

A defence can look robust only because its attack was too weak, so we repeated the 118-bus radius-1 evaluation at 30% and 50% with 400 steps and 5 restarts instead of 80 steps and 2. Across the standard detector, PGD-AT, RAT, the witness and PARRY, detection moved by at most 0.03 in either direction and by 0.001 on average. The witness, whose score no optimisation can lower, moved by the same amount, so these changes are at the level of sampling noise for 400 problems. The rankings did not change. PARRY detected 0.75 and 0.98 under the stronger attack, and the standard detector 0.19 and 0.39.

Discussion

The results separate two kinds of protection. A learned detector can be made much harder to evade by training it against the attacker it will face, and RAT roughly doubles adaptive detection at moderate changes on the larger grid. Its robustness is still empirical. A stronger optimiser, a different attack radius or a new operating regime could erode it, as the decline of RAT at very large changes on the small grid shows. The witness offers the other kind. Its ceiling follows from where the attacker can reach, not from how it searches, so an adaptive or patient attacker cannot raise it beyond the small slack that state re-estimation adds.

The ceiling is loose. A median of 66% of the bus load means that an attacker who stays inside one neighbourhood can still fake a large change at many buses without being certain of detection. The ceiling is set by how well a bus load can be predicted from distant buses, and real demand at the bus level has a large idiosyncratic part. Better predictors, more measurements or weather inputs would tighten it, and the bound would tighten with them without any change to its proof.

The study has limits that a reader should weigh. The physics is simulated while the demand is real, and the meter set is complete, so real systems with fewer meters will differ. The attacker knows the state estimate and the network model, which is the standard and conservative assumption. The witness guarantee covers attacks confined to one hop around the target; an attacker who controls a wider region can reach the witness inputs. RAT does not generalise beyond the attack radii it was trained on, so its training set should cover the widest region an operator considers plausible. False changes of a tenth of the bus load are not detected by any method we tested. The 1% false alarm rate is a calibration target. The temporal and witness components drift more than the snapshot detectors, single months can exceed the target several times over, and the witness raises the average false alarm rate of PARRY above it. Rolling recalibration also assumes a vetted clean reference stream. Where that cannot be assured, thresholds fixed on pre-deployment data are the safer choice. They cost little detection here, while contaminating 1% of the reference steps collapses detection for every defence. Finally, the deep detector is a multilayer perceptron; graph neural detectors[13] may change the numbers, though not the logic of the witness.

Conclusion

This paper studied deep detectors of false data injection under an attacker that is both realisable and adaptive. Such an attacker keeps the readings consistent with the network and must fake a real load change. On the 118-bus grid it evades a standard deep detector in most cases for false load changes up to half the bus load. Adversarial training against norm-bounded perturbations helps little. Training against the realisable attacker helps much more, though without a guarantee. The far-bus witness adds a bound that holds for every attack strategy, up to a small measured slack, and PARRY combines both layers. Small false changes remain undetectable, and the witness ceiling is still loose. Tighter bus-level prediction and sparser meter sets are the next steps.

Data availability

All data analysed in this study are publicly available. The demand data are the UCI ElectricityLoadDiagrams20112014 dataset[56], available at https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014 (DOI 10.24432/C58C86). The IEEE 14-bus and 118-bus network models are those of the Power Systems Test Case Archive (https://labs.ece.uw.edu/pstca/), used as distributed with the open-source pandapower package[55] (https://github.com/e2nIEE/pandapower). The simulated measurements and results generated in this study are available from the corresponding author on reasonable request.

References

1. Liu, Y., Ning, P. & Reiter, M. K. False data injection attacks against state estimation in electric power grids. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS)*, 21–32 (ACM, 2009).

2. Liu, Y., Ning, P. & Reiter, M. K. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security* **14**, 1–33 (2011).

3. Hug, G. & Giampapa, J. A. Vulnerability assessment of AC state estimation with respect to false data injection cyber-attacks. *IEEE Transactions on Smart Grid* **3**, 1362–1370 (2012).

4. Liang, G., Weller, S. R., Zhao, J., Luo, F. & Dong, Z. Y. The 2015 Ukraine blackout: Implications for false data injection attacks. *IEEE Transactions on Power Systems* **32**, 3317–3318 (2017).

5. Deng, R., Xiao, G., Lu, R., Liang, H. & Vasilakos, A. V. False data injection on state estimation in power systems—attacks, impacts, and defense: A survey. *IEEE Transactions on Industrial Informatics* **13**, 411–423 (2017).

6. Liang, G., Zhao, J., Luo, F., Weller, S. R. & Dong, Z. Y. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid* **8**, 1630–1638 (2017).

7. Musleh, A. S., Chen, G. & Dong, Z. Y. A survey on the detection algorithms for false data injection attacks in smart grids. *IEEE Transactions on Smart Grid* **11**, 2218–2234 (2020).

8. Ozay, M., Esnaola, I., Yarman Vural, F. T., Kulkarni, S. R. & Poor, H. V. Machine learning methods for attack detection in the smart grid. *IEEE Transactions on Neural Networks and Learning Systems* **27**, 1773–1786 (2016).

9. Zhang, Y., Wang, J. & Chen, B. Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach. *IEEE Transactions on Smart Grid* **12**, 623–634 (2021).

10. He, Y., Mendis, G. J. & Wei, J. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. *IEEE Transactions on Smart Grid* **8**, 2505–2516 (2017).

11. Ganjkhani, M., Fallah, S. N., Badakhshan, S., Shamshirband, S. & Chau, K.-w. A novel detection algorithm to identify false data injection attacks on power system state estimation. *Energies* **12**, 2209 (2019).

12. Wang, S., Bi, S. & Zhang, Y.-J. A. Locational detection of the false data injection attack in a smart grid: A multilabel classification approach. *IEEE Internet of Things Journal* **7**, 8218–8227 (2020).

13. Boyaci, O. *et al.* Graph neural networks based detection of stealth false data injection attacks in smart grids. *IEEE Systems Journal* **16**, 2946–2957 (2022).

14. Boyaci, O. *et al.* Joint detection and localization of stealth false data injection attacks in smart grids using graph neural networks. *IEEE Transactions on Smart Grid* **13**, 807–819 (2022).

15. Takiddin, A., Ismail, M., Atat, R., Davis, K. R. & Serpedin, E. Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids. *IEEE Transactions on Artificial Intelligence* **5**, 1287–1301 (2024).

16. Shen, Y. & Qin, Z. Detection, differentiation and localization of replay attack and false data injection attack based on random matrix. *Scientific Reports* **14**, 2758 (2024).

17. Sayghe, A. *et al.* Survey of machine learning methods for detecting false data injection attacks in power systems. *IET Smart Grid* **3**, 581–595 (2020).

18. Ruan, J. *et al.* Deep learning for cybersecurity in smart grids: Review and perspectives. *Energy Conversion and Economics* **4**, 233–251 (2023).

19. Badar, H. M. S. *et al.* An identity based authentication protocol for smart grid environment using physical uncloneable function. *IEEE Transactions on Smart Grid* **12**, 4426–4434 (2021).

20. Zahoor, A. *et al.* Lightweight authenticated key agreement protocol for smart power grid systems using PUF. *IEEE Open Journal of the Communications Society* **5**, 3568–3580 (2024).

21. Badar, H. M. S. *et al.* An access control protocol for IoT-based critical infrastructure in smart grid environment. *International Journal of Communication Systems* **35**, e5115 (2022).

22. Badar, H. M. S., Ahmed, S., Kajla, N. I., Fan, G. & Zhang, C. Q-BLAISE: Quantum-resilient blockchain and AI-enhanced security protocol for smart grid IoT. *IEEE Transactions on Consumer Electronics* **71**, 4959–4971 (2025).

23. Szegedy, C. *et al.* Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)* (2014).

24. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)* (2015).

25. Biggio, B. & Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* **84**, 317–331 (2018).

26. Carlini, N. & Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57 (2017).

27. Kurakin, A., Goodfellow, I. & Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)* (2017).

28. Chen, Y., Tan, Y. & Zhang, B. Exploiting vulnerabilities of load forecasting through adversarial attacks. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy)*, 1–11 (ACM, 2019).

29. Sayghe, A., Zhao, J. & Konstantinou, C. Evasion attacks with adversarial deep learning against power system state estimation. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5 (2020).

30. Tian, J., Wang, B., Li, J. & Konstantinou, C. Adversarial attack and defense methods for neural network based state estimation in smart grid. *IET Renewable Power Generation* **16**, 3507–3518 (2022).

31. Tian, J. *et al.* Joint adversarial example and false data injection attacks for state estimation in power systems. *IEEE Transactions on Cybernetics* **52**, 13699–13713 (2022).

32. Tian, J. *et al.* Exploring targeted and stealthy false data injection attacks via adversarial machine learning. *IEEE Internet of Things Journal* **9**, 14116–14125 (2022).

33. Li, J., Yang, Y., Sun, J. S., Tomsovic, K. & Qi, H. ConAML: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (ASIA CCS)*, 52–66 (ACM, 2021).

34. Saber, A. M., Maheshwari, A., Youssef, A. & Kundur, D. Adversarial attacks on deep learning-based false data injection detection in differential relays. *IEEE Transactions on Smart Grid* **16**, 4155–4166 (2025).

35. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)* (2018).

36. Zhang, H. *et al.* Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, 7472–7482 (2019). URL https://proceedings.mlr.press/v97/zhang19p.html.

37. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31 (2018). URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

38. Ma, X. *et al.* Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)* (2018).

39. Xu, W., Evans, D. & Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)* (Internet Society, 2018).

40. Carlini, N. & Wagner, D. Adversarial examples are not easily detected. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, 3–14 (ACM, 2017).

41. Athalye, A., Carlini, N. & Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80 of *Proceedings of Machine Learning Research*, 274–283 (2018). URL https://proceedings.mlr.press/v80/athalye18a.html.

42. Tramèr, F., Carlini, N., Brendel, W. & Madry, A. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33 (2020). URL https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html.

43. Wong, E. & Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80 of *Proceedings of Machine Learning Research*, 5286–5295 (2018). URL https://proceedings.mlr.press/v80/wong18a.html.

44. Cohen, J., Rosenfeld, E. & Kolter, Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, 1310–1320 (2019). URL https://proceedings.mlr.press/v97/cohen19c.html.

45. Li, J., Yang, Y., Sun, J. S., Tomsovic, K. & Qi, H. Towards adversarial-resilient deep neural networks for false data injection attack detection in power grids. In *2023 32nd International Conference on Computer Communications and Networks (ICCCN)*, 1–10 (2023).

46. Huang, R. & Li, Y. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. *IEEE Transactions on Smart Grid* **14**, 2367–2376 (2023).

47. Ran, X., Tay, W. P. & Lee, C. H. T. Robust data-driven adversarial false data injection attack detection method with deep Q-network in power systems. *IEEE Transactions on Industrial Informatics* **20**, 10405–10418 (2024).

48. Ilyas, M. *et al.* PoisonShield-FL-NIDS: A robust defense against poisoning attacks in federated learning intrusion detection. *IEEE Internet of Things Journal* **13**, 6105–6115 (2026).

49. Fan, G. *et al.* Anomal-EF: Self-supervised learning and multi-head attention based anomalous node detection in IoT networks. *Cluster Computing* **29**, 544 (2026).

50. Davis, K. R., Morrow, K. L., Bobba, R. & Heine, E. Power flow cyber attacks and perturbation-based defense. In *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, 342–347 (2012).

51. Zhang, Z., Deng, R., Yau, D. K. Y., Cheng, P. & Chen, J. Analysis of moving target defense against false data injection attacks on power grid. *IEEE Transactions on Information Forensics and Security* **15**, 2320–2335 (2020).

52. Lakshminarayana, S. & Yau, D. K. Y. Cost-benefit analysis of moving-target defense in power grids. *IEEE Transactions on Power Systems* **36**, 1152–1163 (2021).

53. Do Coutto Filho, M. B. & de Souza, J. C. S. Forecasting-aided state estimation—part I: Panorama. *IEEE Transactions on Power Systems* **24**, 1667–1677 (2009).

54. Zhao, J., Zhang, G., Dong, Z. Y. & Wong, K. P. Forecasting-aided imperfect false data injection attacks against power system nonlinear state estimation. *IEEE Transactions on Smart Grid* **7**, 6–8 (2016).

55. Thurner, L. *et al.* pandapower—an open-source Python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems* **33**, 6510–6521 (2018).

56. Trindade, A. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository (2015).

57. Abur, A. & Gómez Expósito, A. *Power System State Estimation* (CRC Press, 2004).

58. Monticelli, A. *State Estimation in Electric Power Systems* (Springer, Boston, MA, 1999).

Funding

This research received no external funding.

Author contributions

S.L. conceived and designed the study, developed the research methodology, and supervised the research. X.X. developed the deep learning models, conducted the simulations and experiments, and performed the data analysis. P.L. contributed to the validation of the proposed method, interpretation of the experimental results, and preparation of figures. S.L. and X.X. drafted the manuscript. All authors contributed to manuscript revision, reviewed the results, and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.