Results in Engineering 32 (2026) 113109

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Research paper

Engineering audit protocol for trustworthy electricity-theft detection on non-stationary smart-meter data

Hafiz Muhammad Azeem Akram [a,*], Siamak Talatahari [a,b], Adil Jhangeer [c,d]

[a] Department of Computer Science, Khazar University, Baku, Azerbaijan
[b] School of Computing, Macquarie University, Sydney, NSW, Australia
[c] IT4-Innovations, VSB – Technical University of Ostrava, Ostrava-Poruba, Czech Republic
[d] Department of Computer Engineering, Biruni University, Istanbul, Turkey

ARTICLE INFO

Keywords:
Advanced metering infrastructure
Data leakage
Deployment readiness
Electricity theft detection
Engineering audit protocol
Trustworthy artificial intelligence

ABSTRACT

Electricity theft is the dominant non-technical loss in advanced metering infrastructure, threatening smart-grid reliability and utility revenue worldwide. Despite a proliferation of deep-learning detectors, the data on which they are evaluated is rarely audited for leakage; undetected leakage in the evaluation pipeline can inflate the performance these detectors report. This paper reframes electricity-theft detection as a critical-infrastructure integrity problem and proposes a model-agnostic engineering audit that exposes such inflation before deployment. The audit applies four validation gates: a data-provenance audit, a temporal integrity check, a leakage detection gate, and a realistic distribution-shift test; results are reported under a three-tier disclosure standard of full population, stable regime, and audited cohort, with a 12-item checklist. On the State Grid Corporation of China dataset, whose missingness exhibits a 41.44 percentage-point structural break in a single month at January 2016, six of eight architectures lose 28.5 % to 32.6 % of their reported precision-recall area under the curve (PR-AUC; absolute drop 0.109 to 0.134) once the shortcut is removed. Customer-grouped split leakage is operationally negligible (maximum $|\Delta$ PR-AUC$| = 0.0035$), and an audit reveals detection reliability varies by a factor of seven across 22 subgroups. A permutation-importance probe identifies consumption scale and variability computed over observed days, together with a direct contribution from the missingness rate, as the signals correlated with the regime and available to every detector, indicating the likely mechanism of the inflation. The gap between reported and audited performance gives utilities a verification step that current reporting practice does not provide.

1. Introduction

Electricity theft is the unauthorized consumption of electrical energy through meter tampering, meter bypassing, the use of strong magnets to interrupt the rotation of the disc, or direct hooking from the distribution feeder. It constitutes the dominant component of non-technical losses in power distribution networks [1–4]. The phenomenon is a serious threat to three engineering objectives at once. Grid stability suffers because unmetered consumption distorts load forecasts and leads to inadequate distribution infrastructure [5,6]. Utility revenue protection is degraded because non-technical losses cost utilities approximately $96 billion per year worldwide [2,7,8]. National reports place the loss at roughly 20 % of generated electricity in India and 16 % in Brazil [3,9], while the State Grid Corporation of China region reports losses on the order of 10 to 15 % of generation [10]. Fairness to honest customers is finally compromised because they absorb the cost of non-technical losses through cross-subsidies in regulated tariffs [11,12]. A utility deploying a detection system therefore commits real inspection budget and real on-site inspection capacity at customer premises, and it incurs real financial risk from false positives. The electricity-theft-detection function within an advanced metering infrastructure is, for that reason, not a stand-alone classification task but a key component of reliable grid operation.

The deployment of advanced metering infrastructure over the past decade [11,13–15] has made data-driven detection possible, and the field has pursued it widely. Detection architectures, most of them benchmarked on the State Grid Corporation of China consumption dataset [14], fall into four method families, reviewed in Section 2. The convolutional family derives from the wide-and-deep baseline [2,10,16,17]; the attention- and graph-based family builds models on graphs constructed from consumption series [18,19]; the classical machine-learning family works on engineered features [1,9,20]; and the recurrent and autoencoder family targets temporal structure [7,21–23]. Alongside

* Corresponding author.
E-mail address: a.akram@khazar.org (H.M.A. Akram).

https://doi.org/10.1016/j.rineng.2026.113109
Received 30 May 2026; Received in revised form 22 August 2026; Accepted 19 September 2026
Available online 22 September 2026

the four families sit cloud-based detectors [8], semi-supervised detectors trained only on normal profiles and evaluated under false-data-injection attacks [24], and unsupervised observer-meter methods [25]. The reported headline values approach the ceiling of their respective metrics: F1 scores of 0.99 [2] and 0.95 [26], receiver-operating-characteristic AUC above 0.97 [26,27], and gains of 5 to 8 percentage points in F1 and AUC for a privacy-preserving variant [28]. The same dual-branch study reproduces earlier convolutional and recurrent baselines at F1 below 0.40 [26], so the range reported on a single dataset is wide, and the strongest numbers imply near-perfect detector reliability.

However, the published numbers are largely unaudited. Confidence intervals are absent from every method paper surveyed here [1,2,7–10,16,23,24]. The regime composition of the evaluation cohort is nowhere declared. The scope of preprocessing operations applied before the train-test split is rarely reported. Whether the oversampling is applied within each training fold or once before the cross-validation split is essentially never reported. For a utility considering deployment, these missing disclosures are the central problem. A detector intended for critical infrastructure cannot be accepted unless its reported performance has been verified by an audit that controls for the data-acquisition conditions of the dataset.

The practical consequence of this omission is concrete. Recent work reports F1 scores above 0.99 [2]. The F1 obtained on the full mixed-regime population, on the same dataset and the same labels, is 0.44. In operational terms, the gap between those two values is the gap between an automated deployment that processes the top 10 % of customers without manual review and a triage deployment that defers 90 % of cases to human inspection. That gap is the quantity this paper is designed to expose and to verify.

We argue that this is a reliability-engineering problem, not merely one of machine learning. Engineering systems, whether mechanical, electrical, or data-driven [29], cannot be deployed on critical infrastructure on the basis of unaudited performance claims. The analogous practices in adjacent engineering fields are structural testing in civil engineering, nondestructive testing in manufacturing, factory acceptance testing in power electronics, and certification testing in aviation. In each case, a component's reliability is established through a standardized verification protocol whose pass and fail criteria are independent of the component's design. To date, the electricity-theft-detection literature has no analogous verification protocol for detectors deployed on advanced metering infrastructure.

This paper proposes such a protocol: a four-gate verification pipeline that detects inflated performance before deployment and certifies that detectors on advanced metering infrastructure reflect audit-verified performance. The protocol is operationalized through four validation gates. Gate A is the data-provenance audit. Gate B is the temporal integrity check. Gate C is the leakage detection gate. Gate D is the realistic distribution-shift test. Results from the four gates are reported under a three-tier disclosure standard, namely the full population, the stable regime, and the audited cohort, together with a 12-item disclosure checklist. The protocol is model-agnostic. Any detector that produces a score on a held-out cohort can be evaluated under the protocol, and the validation gates apply identically to a logistic regression, an extreme gradient boosting classifier, and a convolutional autoencoder and Transformer.

The contributions of this paper are as follows:

- **An engineering audit protocol comprising four validation gates.** The protocol provides a data-provenance audit, a temporal integrity check, a leakage detection gate, and a realistic distribution-shift test, operationalized through a three-tier disclosure standard and a 12-item checklist.
- **A taxonomy of methodological vulnerabilities in electricity-theft detection.** The taxonomy defines four leakage classes: regime-correlated preprocessing leakage, customer-grouped split leakage, synthetic-oversampling artifact leakage, and temporal-autocorrelation leakage. Each class is mapped to its audit criterion and the reporting omission the protocol addresses.
- **A cross-architecture stress-test on the State Grid Corporation of China dataset.** Six of eight architectures lose 28.5–32.6 % of reported PR-AUC (absolute: 0.109–0.134) once the regime shortcut is removed; customer-grouped split leakage is operationally negligible; and an audit reveals that detection reliability varies by a factor of seven across 22 subgroups.
- **A feature-level attribution via permutation importance.** A logistic-regression probe of the shared feature space identifies the scale and variability of consumption over observed days as the dominant signals correlated with the regime, with a direct contribution from the missingness rate; the per-customer maximum contributes negligibly. The unaudited pipeline relies on this signal.

The remainder of the paper is organized as follows. Section 2 reviews the recent literature on electricity-theft detection and defines the four recurring methodological vulnerabilities. It then presents the diagnostic checklist that maps each vulnerability to its audit criterion. Section 3 formalizes the dataset and the regime structure. Section 4 introduces the engineering audit protocol and links each of its four gates to the vulnerability class that gate addresses. Section 5 describes the eight detectors, their hyperparameters, and the stress-test validation design. Section 6 reports the results across the four gates, including the mechanism attribution and the consolidated verdict, and interprets them within the trustworthy artificial intelligence frame. Section 7 concludes the paper.

2. Related work and methodological vulnerabilities

Recent work on electricity-theft detection from smart-meter data divides naturally into four method families and one parallel tradition. The four method families are convolutional models derived from the wide-and-deep baseline, attention-based extensions and hybrids, classical machine-learning approaches based on engineered features, and recurrent or autoencoder models targeting temporal structure. The parallel tradition is the reliability-engineering literature on trustworthy artificial intelligence in the energy sector, which addresses the deployment-readiness question that the four method families have not yet resolved.

2.1. Method families on the state grid corporation of China dataset

The convolutional family is based on the wide-and-deep convolutional neural network of Zheng et al. [14], which combines a wide component for global features with a deep convolutional component for the weekly periodicity of consumption. Recent extensions add attention modules and dilated convolutions [16], integrate convolutional autoencoders with Transformers [2], or replace the deep component with multi-scale differencing mechanism [17]. The attention-based family includes graph-attention models on dynamically constructed customer graphs [18,19], autocorrelation detectors deployed on a cloud server [8], and multi-objective evolutionary hybrid models [7]. The classical machine-learning family includes pattern-based dynamic time warping with k-nearest neighbors [1], decision-tree ensembles optimized by the firefly algorithm with post-hoc explanations [9,20], and focal-loss densely connected convolutional networks [10]. The recurrent and autoencoder family includes convolutional long short-term memory networks [21,22], transform-based dimensionality reduction paired with lag-embedded networks [23], and semi-supervised convolutional autoencoders trained against explicit false-data-injection attack types [24]. Alongside the four families, other approaches include model-agnostic explainable artificial intelligence [30], two-step detection strategies [31], comparative analysis of unbalanced data handling [32], statistical framework [33], anomaly detection [34,35], and clandestine-connection detection via satellite imagery [36].

2.2. Reliability engineering and trustworthy artificial intelligence in the energy sector

A second line of work frames the detection function within the reliability-engineering tradition. Trustworthy artificial intelligence in the energy sector has been the subject of recent surveys [29] and review articles [3,15,37] that identify deployment readiness as the dominant open problem. Several studies address robustness to adversarial manipulation and data-quality issues [38–43]. Generalization under distribution shift has been studied in a separate body of work [17,44,45]. Active learning has been proposed to minimize labeling cost [46]. Privacy-preserving deployment is addressed in [47,48]. Blockchain-based provenance for detection results has been proposed [49]. Large-scale hybrid detectors [50–55] share a common feature: they are novel architectures proposed without the audit protocols against which their reported performance can be verified. This paper addresses that gap directly.

2.3. From method families to methodological vulnerabilities

The literature surveyed above shares a common practice in the way results are reported. Across the four method families, reported performance is obtained under evaluation pipelines whose preprocessing scope, data splitting, oversampling scope, and temporal partitioning are rarely declared at the level of detail that would permit independent verification. The reliability-engineering tradition has identified deployment readiness as the open problem but has not produced the audit criteria that would quantify it. The result is a literature in which competitive PR-AUC values cluster near the near-ceiling regime without an external account of which fraction of that performance is verifiable.

The remainder of this section defines the four leakage classes that systematically inflate reported detection performance across this literature. Each class is a reliability-engineering vulnerability, that is, a defect in the evaluation pipeline that, if undetected, produces a detector whose verified reliability does not match its reported reliability. We survey the four classes in turn. Each entry describes the underlying mechanism, names the prior studies in which the vulnerability has gone unaudited, and specifies the audit criterion under which it is detected. The audit criterion of each class corresponds to a validation gate of the protocol defined in Section 4; and each subsection states that cross-reference explicitly.

2.4. Class 1: regime-correlated preprocessing leakage

When a dataset is collected across an evolving measurement regime, as is the case for the State Grid Corporation of China dataset with its January 2016 missingness boundary, the value that the preprocessing pipeline imputes (the mean, the median, or zero) is itself a function of the regime composition. The standard practice of fitting an imputer on the entire dataset, before the train-test split propagates the regime indicator into the imputed feature distribution. A deep learning classifier can then learn the missingness pattern as a class-discriminative shortcut. The vulnerability is present from the origin. The dataset was released together with the most widely used baseline, the wide-and-deep network of Zheng et al. [14], whose window of 2014-01-01 to 2016-10-31 already contains the January 2016 break, and neither that work nor the later models evaluated on the full window declares the regime composition. Examples include the attention-based extension with dilated convolutions of Xia et al. [16], the hybrid ConvLSTM detector of Gao et al. [55], the multi-scale model of Wang et al. [17], the graph-based detectors of Liao et al. [18,19], the convolutional autoencoder and Transformer combination of Le et al. [2], the multi-objective evolutionary hybrid model of Tursunboev et al. [7], and the ensemble optimized by the firefly algorithm of Kabir et al. [9]. The audit criterion is the data-provenance audit, addressed by Gate A in Section 4.1.

2.5. Class 2: customer-grouped split leakage

In a longitudinal smart-meter dataset a customer may contribute several rows, and a random k-fold split that distributes those rows between the training and test folds risks an inflated evaluation because the customer-identity signal crosses between the training and test folds. The SGCC matrix is not of that form. Each of its N rows is one customer carrying one label, so a row-level split is already customer-disjoint and this mechanism cannot arise. What remains is a weaker structure: exact-duplicate consumption profiles recorded under distinct customer identifiers, which a random split separates across folds in the same way. The grouped arm of the audit therefore constrains duplicate profiles rather than customer identity. The audit criterion is the temporal integrity check, addressed by Gate B in Section 4.2. Grouped splitting removes this mechanism by construction; Section 6.2 measures what it buys on this dataset.

2.6. Class 3: synthetic-oversampling artifact leakage

A common response to the class imbalance arising from the 8.53 % theft-positive class prevalence is a synthetic minority oversampling technique or one of its variants [2,9,23,32,41,50,55]; other studies use focal loss instead [10,16] or unsupervised formulations that require no resampling [51]. When the oversampling is applied before the cross-validation split, the synthetic positives generated from training-fold positives can contaminate the test fold through linear interpolation between nearest neighbors. The audit criterion is the leakage detection gate, addressed by Gate C in Section 4.3. The literature rarely reports whether oversampling is applied within each training fold or once before the cross-validation split.

2.7. Class 4: temporal-autocorrelation leakage

A daily smart-meter time series exhibits autocorrelation at lag 1 (consecutive days) and lag 7 (weekly periodicity). Where a feature extractor emits more than one sample per customer, a random k-fold split that ignores this autocorrelation can place adjacent days of the same customer in different folds, producing an inflated test-set evaluation because train and test folds share short-term temporal structure. That construction is not used here: each customer contributes exactly one row and one sample (Section 3.1), so a random or grouped fold carries a customer's whole series, and the chronological protocol's test window begins 62 days after its training window ends; no two adjacent days of a customer are ever placed in different folds, and the mechanism cannot arise. Closing it by construction is part of what permits the inflation reported under Gate D to be attributed to the regime boundary rather than to short-range temporal structure. The exposure that does remain is of a different kind: folds drawn at random share a measurement regime, a matter of distribution shift rather than of autocorrelation. The audit criterion is the realistic distribution-shift test, addressed by Gate D in Section 4.4. The literature on evaluation under distribution shift in electricity-theft [38,39,44,54–56] is sparse relative to the in-distribution literature.

2.8. Diagnostic checklist and survey of recent literature

Table 1 maps each leakage class to its audit criterion and applies those criteria to ten recent electricity-theft-detection papers as a survey of what each paper reports. The pattern is uniform. Across all four classes, only one paper [10] reports a correction for multiple comparisons. Only one [24] declares a false-data-injection threat model. The regime composition is not declared in any paper. The scope of synthetic oversampling is rarely specified. This is the common reporting practice that this paper addresses.

The four classes do not occur in isolation. Regime-correlated preprocessing leakage and customer-grouped split leakage can compound, and

Table 1
Diagnostic checklist for State Grid Corporation of China (SGCC) electricity-theft-detection papers. A dash marks an audit criterion the paper does not report; n/a and review mark criteria that do not apply to a review paper; FDI is the false-data-injection threat model.

Paper	Regime	Split type	CI on Δ	Oversampling scope	Code	Threat model
Ahir 2022 [1]	–	–	–	–	–	passive
Chen 2023 [10]	–	–	Friedman/Holm	focal loss	repository	passive
Qi 2023 [24]	–	–	–	normal-only	–	**11 FDI types**
Xia 2023 [16]	–	–	–	focal loss	–	passive
Shahzadi 2024 [23]	–	–	–	LoRAS	–	passive
Tursunboev 2024 [7]	–	–	–	–	–	passive
Kabir 2025 [9]	–	–	–	SMOTE	link	passive
Le 2025 [2]	–	28-day	–	K-SMOTE	–	passive
Gao 2022 [55]	–	64/16/20	–	borderline-SMOTE	–	passive
Iqbal 2025 [3]	–	review	n/a	review	n/a	review
Present work	**three-tier**	**customer-grouped 5-fold**	**paired bootstrap**	**training-fold**	**on accept.**	**declared**

synthetic-oversampling artifact leakage can interact with both. The proposed protocol therefore applies all four audit criteria jointly in a single evaluation pipeline. Clearing a subset of the gates is not sufficient to certify a detector.

3. Dataset and problem formulation

3.1. The state grid corporation of China dataset

The State Grid Corporation of China dataset [14] contains $N = 42{,}372$ unique customers. The raw matrix has 1,034 columns spanning 2014-01-01 to 2016-10-31, and patching a single missing calendar date yields $T = 1{,}035$ expected days. The theft-positive class prevalence is $\bar{y} = 8.53\,\%$. The consumption matrix is denoted $\mathbf{X} \in \mathbb{R}^{N\times T}$ with entries $X_{it} \in \mathbb{R} \cup \{\texttt{NaN}\}$, where i indexes customers, t indexes days, and `NaN` denotes a missing value. The binary labels are collected in $\mathbf{y} \in \{0,1\}^N$. Each row is one customer's complete series and carries a single label. No sliding windows or subsequences are extracted at any stage of the pipeline, so a customer contributes exactly one sample, and that sample falls in exactly one fold. From the perspective of the engineering audit protocol, the dataset is a measurement system, and its integrity must be verified before any detection algorithm is evaluated on it. That verification is the role of Gate A, defined in Section 4.1.

3.2. Two observation states and the January 2016 structural break

Any analysis of missing data on this dataset must distinguish between two observation states whose semantics differ. A reading can be missing, in which case the meter or communication channel failed to report. Alternatively, a reading can be observed as zero, in which case the customer used no electricity during the interval. For customer i and time index t, let $\mathcal{R}$ be the set of entries the acquisition system recorded; three binary indicators capture the distinction:

$$\begin{aligned} o_{it} &= \mathbb{1}[(i,t) \in \mathcal{R}], \\ m_{it} &= 1 - o_{it}, \\ z_{it} &= o_{it}\,\mathbb{1}[X_{it} = 0], \end{aligned} \tag{1}$$

where m_{it} encodes the state of the metering infrastructure and z_{it} the customer's behavior. The two are not interchangeable, and on this dataset they carry opposite information: over the post-2016 window the per-customer missingness fraction correlates with the theft label at $r = +0.095$ and the observed-zero fraction at $r = -0.070$, both significant at $p < 10^{-46}$, and the missingness association is unchanged at $+0.097$ once the observed consumption is controlled for, which is what makes the pattern *informative* in the sense of Ghorbani and Zou [57]. A preprocessing pipeline that conflates the two under a single zero imputation therefore averages a positive predictor of theft with a negative one, and produces a feature whose semantics depend on the dominant cause of zeros in the cohort under consideration, a hazard reported for energy systems generally [58]. That zero-filling rule was introduced together with the dataset [14] and is still applied [2,59].

The global mean daily missingness over 2014 and 2015 is 35.64 %. Over 2016 the global mean daily missingness is 1.62 %, whereas the mean of the per-customer missingness fractions is 1.94 %. The monthly trajectory descends monotonically through 2015 with a sharp discontinuity between December 2015 (43.0 %) and January 2016 (1.56 %), the largest single-month change in the dataset at 41.44 percentage points. Figs. 1–3 show this discontinuity. It is not a gradual drift but an abrupt *structural break* [60,61] in the data-acquisition process, consistent with a one-time alteration of the metering infrastructure. Because the change is one-time, with the series never switching back, we characterize the dataset as two *regimes* separated by a single break, in the sense of Bai and Perron [61], rather than as a recurring regime-switching process. Recent work in electricity-theft detection refers to such an event loosely as a regime shift.

The step is statistically anomalous relative to the background variation in the dataset. Across the 32 month-to-month transitions in the 34-month observation window excluding January 2016, the mean absolute change in monthly missingness is 1.30 percentage points with a standard deviation of 0.95 percentage points. The January 2016 step of 41.44 percentage points exceeds the next-largest single-month change in the dataset (3.21 percentage points, November 2014) by a factor of 12.9. A step of this magnitude is incompatible with the background month-to-month variation of a single-regime process. This structural break is the basis for the Class 1 vulnerability in Section 2.4.

Two temporal partitions formalize the regime structure:

$$\begin{aligned} \mathcal{T}_{\text{pre}} &= \{t : d_t < \text{2016-01-01}\}, \\ \mathcal{T}_{\text{post}} &= \{t : d_t \geq \text{2016-01-01}\}, \end{aligned} \tag{2}$$

where d_t denotes the calendar date of column t. For a window $\mathcal{W} \subseteq \{1,\ldots,T\}$ write $m_i = |\mathcal{W}|^{-1} \sum_{t\in\mathcal{W}} m_{it}$ for the fraction of days on which customer i's meter did not report, and $\zeta_i = |\mathcal{W}|^{-1} \sum_{t\in\mathcal{W}} (m_{it} + z_{it})$ for the *effective zero ratio*, the fraction carrying no energy into the model once missing entries are filled with zeros. These are the quantities the observability gate of Algorithm 1 applies, evaluated over $\mathcal{W} = \mathcal{T}_{\text{post}}$; ζ_i rather than $|\mathcal{W}|^{-1} \sum_t z_{it}$, because a day is uninformative about consumption whether the meter reported zero or reported nothing at all.

3.3. Operational objective and the inspection-budget framing

A utility operating an advanced metering infrastructure can inspect only K customers per inspection cycle. The operational value of a detector $f_\theta : \mathbb{R}^T \to [0,1]$ is the recall at top-K rather than the area under a generic receiver operating characteristic curve. The primary metric is PR-AUC,

$$\text{PR-AUC}(f_\theta) = \int_0^1 P_\theta\left(R_\theta^{-1}(r)\right) \mathrm{d}r, \tag{3}$$

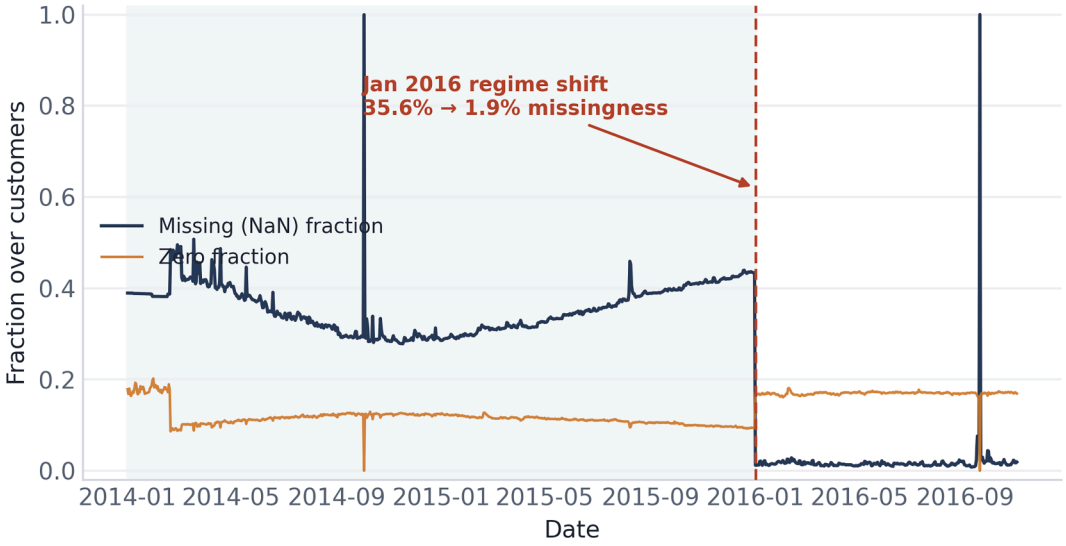

Fig. 1. Daily missingness rate (fraction of customers whose meter failed to report) and zero rate (fraction observed as zero) across the observation window. The January 2016 regime boundary marks the largest single-month change in the dataset at 41.44 percentage points.

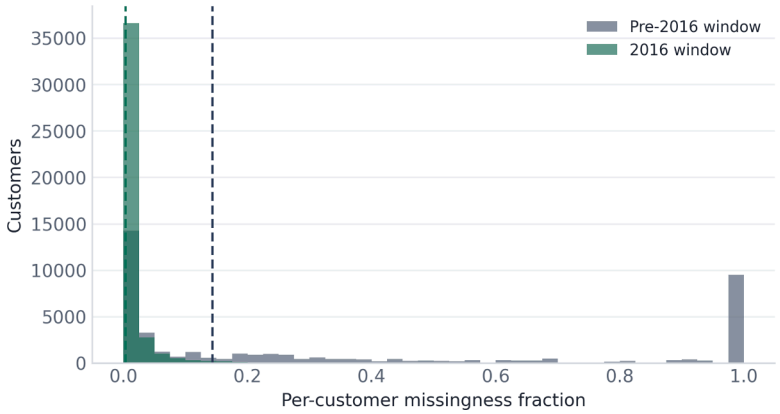

Fig. 2. Histogram of the per-customer missingness fraction, 2014–2015 versus 2016.

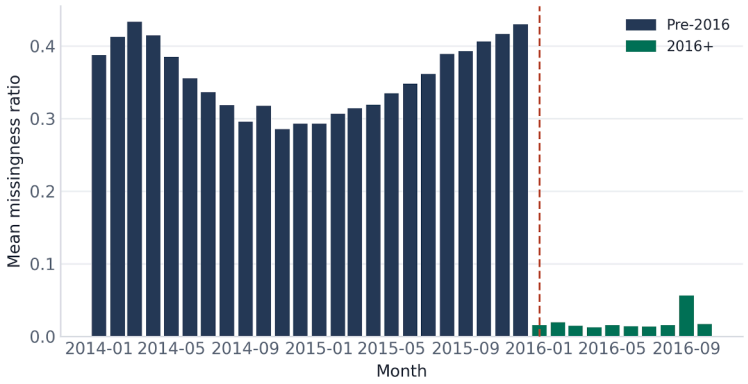

Fig. 3. Mean monthly missingness rate across customers, with the January 2016 regime boundary.

where $P_\theta(\tau)$ and $R_\theta(\tau)$ are the precision and recall of f_θ at threshold τ. The cost-optimal decision threshold under an asymmetric cost matrix with per-audit cost c_{FP} and per-event missed-theft cost c_{FN} is

$$\tau^\star = \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}. \tag{4}$$

The ratio $c_{\mathrm{FN}}/c_{\mathrm{FP}}$ is utility-specific, and to our knowledge no published estimate exists for distribution system operators; the economic strand of the non-technical-loss literature selects inspections to maximize their expected economic return rather than working through a fixed ratio [62]. The value used here is therefore a declared working assumption rather than a sourced quantity. The asymmetry itself has an operational basis: a false positive costs one site inspection, while a missed theft continues to accrue unbilled energy until the customer is next inspected. Under the working assumption $c_{\mathrm{FN}}/c_{\mathrm{FP}} = 20$, the cost-optimal threshold is $\tau^\star \approx 0.05$. Eq. (4) locates the operational regime; it does not set a threshold in the pipeline. Written as a function of the ratio, $\tau^\star(r) = 1/(1+r)$ falls from 0.167 at $r = 5$ through 0.091 at $r = 10$ and 0.048 at $r = 20$

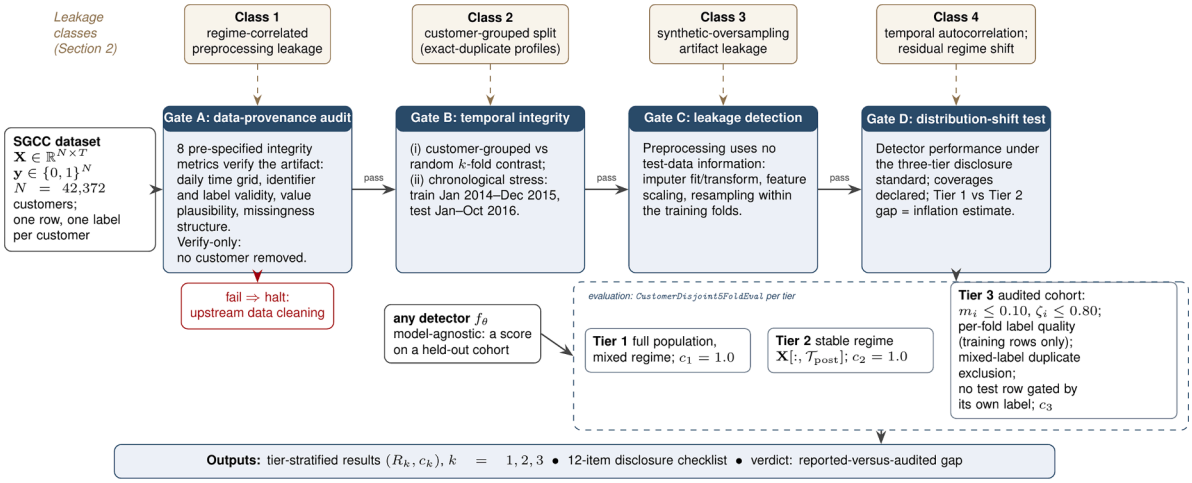

Fig. 4. Overall framework of the engineering audit protocol.

to 0.020 at $r = 50$, always far below the default 0.5. Evaluated over $\tau \in [0.02, 0.17]$ on the held-out scores, every detector's recall stays above 0.89 at every τ: the entire range keeps the detectors in the high-recall regime the ratio locates, and no reported operating point depends on the exact value of τ.

4. The engineering audit protocol

The engineering audit protocol comprises four sequential validation gates, each applying pre-specified pass and fail criteria before the process advances. Fig. 4 shows the overall framework: the information flow from the dataset artifact through the four gates to the tier-stratified outputs, and the mapping from each leakage class of Section 2 to the gate that audits it (indicated by dashed arrows). The four validation gates apply pre-specified criteria in sequence: Gate A verifies the provenance and integrity of the dataset artifact and halts the audit on failure; Gate B verifies the temporal integrity of the split construction; Gate C verifies that the preprocessing pipeline uses no information from the test data; Gate D reports detector performance under the three-tier disclosure standard with tier coverages declared. Each gate produces an integrity metric whose pass and fail criteria are declared in advance and are verifiable from the protocol outputs. The protocol is model-agnostic: any detector that produces a score on a held-out cohort enters at the evaluation stage.

4.1. Gate A: data-provenance audit

Gate A addresses the Class 1 vulnerability of Section 2.4, namely regime-correlated preprocessing leakage. The data-provenance audit verifies eight integrity metrics of the input dataset before any detector is trained on it. The metrics are calendar-grid continuity; identifier uniqueness; binary label validity; the number of customers with no reading on any date; detection of the January 2016 structural break in the missingness indicator; the 2016 missingness rate; the negative-reading rate; and the zero rate. Each metric is pre-specified and returns a pass or fail verdict. Table 2 reports the outcome of all eight metrics on the State Grid Corporation of China dataset. All eight return a pass. The data-provenance audit is the prerequisite for any subsequent gate; failure at this stage halts the audit and requires upstream data cleaning.

Table 2
Gate A (Data-Provenance Audit) on the State Grid Corporation of China dataset: all eight integrity metrics pass.

Integrity metric	Value	Pass
Dates continuous (patched)	1,035	✓
Identifier unique	42,372	✓
Labels binary $\{0, 1\}$	yes	✓
All-NaN customers rare	5 (≤ 10)	✓
Missingness drops in 2016	$0.356 \to 0.019$	✓
2016 missingness ≤ 0.05	0.0194	✓
Negative readings rare (≤ 0.001)	0.000	✓
Zero rate not extreme (≤ 0.50)	0.132	✓

4.2. Gate B: temporal integrity check

Gate B addresses the Class 2 vulnerability of Section 2.5, namely customer-grouped split leakage. The temporal integrity check verifies that the train-test split respects the engineering constraints of the deployment scenario. Three constraints apply. Under customer-disjointness, no customer appears in both the training and test folds for any combination of seed and fold. Under temporal-disjointness in the chronological-stress protocol, training is restricted to one time window and testing to a disjoint time window. Under regime-stratification, the evaluation cohort for each tier of the three-tier disclosure standard is well-defined. The customer-disjointness verification is reported in Fig. 6: the training-side and test partitions share zero customers in every (seed, fold) cell. Two deployment scenarios are distinguished, and each has a certifying protocol. A utility that scores the future behavior of customers whose history it trained on is served by the same-customer protocol: the test fold is the training customers themselves, read on the later window. A utility that screens customers the detector has never seen is served by the customer-disjoint protocol: the test fold holds out customers entirely. Both protocols train on the pre-shift window January 2014 to December 2015 and test on January to October 2016; each is a single out-of-time holdout in the sense of Maldonado et al. [63], rather than an iterated rolling-origin scheme. A third protocol, the training-window control, tests held-out customers inside the training window, so that a chronological loss is not attributable to window length or season. The contrast between the same-customer and customer-disjoint protocols is the measured cost of moving from known to unseen customers, reported in Section 6.3.

4.3. Gate C: leakage detection gate

Gate C addresses the Class 3 vulnerability of Section 2.6, namely synthetic-oversampling artifact leakage. The leakage detection gate verifies that the preprocessing pipeline and any synthetic-oversampling operation use no information from the test data. Specifically, the gate requires that the imputer is fitted on the training fold only and not on the training and test data combined, and that the feature-scaling statistics (mean, standard deviation, robust statistics) are estimated on the training fold only. The gate further requires that any synthetic

oversampling (synthetic minority oversampling technique and variants) is applied within each training fold only and not before the cross-validation split. Finally, the gate requires that the cross-validation indices, preprocessing parameters, and oversampling seeds are stored as immutable artifacts so that the audit can be reproduced exactly.

4.4. Gate D: realistic distribution-shift test

Gate D addresses the Class 4 vulnerability of Section 2.7, namely temporal-autocorrelation leakage. The realistic distribution-shift test reports detector performance under the three-tier disclosure standard. Tier 1 reports unrestricted performance on the full population, namely all customers across the full observation window, which is the conservative claim and the operating point of the unaudited literature. Tier 2 retains all customers but restricts the temporal window to $\mathcal{T}_{\text{post}}$, isolating model behavior under the stable regime; the contrast between Tier 1 and Tier 2 quantifies the regime-shortcut inflation. The contrast also shortens each customer's observation window; Section 6.5 separates the shortcut component from this shared factor. Tier 3 reports performance on an audited cohort whose customers pass pre-specified observability gates (missingness $\leq 10\,\%$, effective zero ratio $\leq 80\,\%$); within each evaluation fold the training rows additionally pass a label-quality rule computed within that fold (Section 6.9), while test rows are retained without reference to their labels; Tier 3 is the operating point of an automated triage system that defers ambiguous cases to manual review, with explicit coverage disclosure.

The procedure is given as Algorithm 1.

Algorithm 1 Gate D, the realistic distribution-shift test.

Require: Consumption matrix $\mathbf{X} \in \mathbb{R}^{N\times T}$, labels $\mathbf{y} \in \{0,1\}^N$, regime boundary $d^\star$, observability thresholds (τ_m, τ_ζ), label-quality rule on training rows (Section 6.9), classifier f_θ

Ensure: Tier-stratified performance results (R_1, R_2, R_3) with coverages (c_1, c_2, c_3)

$\mathcal{T}_{\text{post}} \leftarrow \{t : d_t \geq d^\star\}$
$\mathbf{X}^{(1)} \leftarrow \mathbf{X}$; $c_1 \leftarrow 1.0$
$\mathbf{X}^{(2)} \leftarrow \mathbf{X}[:, \mathcal{T}_{\text{post}}]$; $c_2 \leftarrow 1.0$
$\mathcal{G} \leftarrow \{i : m_i \leq \tau_m \wedge \zeta_i \leq \tau_\zeta\}$; $\mathbf{X}^{(3)} \leftarrow \mathbf{X}^{(2)}[\mathcal{G}, :]$; $c_3 \leftarrow |\mathcal{G}|/N$
for $k \in \{1,2,3\}$ **do**
    $R_k \leftarrow$ `CustomerDisjoint5FoldEval`$(f_\theta, \mathbf{X}^{(k)}, \mathbf{y})$, where for $k = 3$ each fold (i) fits the label-quality rule of Section 6.9 on its own training rows, dropping the training rows that fail, and (ii) abstains from test rows whose exact-duplicate profile group carries disagreeing training-fold labels; no test row is gated by its own label
end for
return $\{(R_k, c_k)\}$

4.5. The 12-item disclosure checklist

To operationalize the four gates without ambiguity, Table 3 lists the twelve items that a paper using the State Grid Corporation of China dataset must report in order to be reproducible and comparable across papers. The first nine items are general reproducibility disclosures. Items 10 to 12 are specific to electricity-theft detection on non-stationary smart-meter data.

4.6. Paired-bootstrap inference as the integrity metric for Gate D

All cross-protocol and cross-architecture comparisons within Gate D use the paired bootstrap with $B = 5{,}000$ replicates at confidence level $1 - \alpha = 0.95$. The procedure is given as Algorithm 2. The integrity criterion is that the 95 % confidence interval on Δ excludes zero. The gate's pass or fail verdict is therefore a verifiable, pre-specified statistical statement rather than an architectural claim.

Table 3
The 12-item disclosure checklist that operationalizes the four validation gates.

#	Disclosure item
1	Total customers, days, and readings
2	Class prevalence before and after preprocessing
3	Missingness rate (overall and per regime)
4	Preprocessing steps with code reference
5	Imputation method, or handling without imputation
6	Sentinel- or special-value encoding
7	Train, validation, and test fractions
8	Random seed(s)
9	Hardware specification
10	Regime window (start, end, breakpoints)
11	Split type (random, temporal, customer-grouped)
12	Customer-disjointness verification (overlap = 0)

Algorithm 2 Paired-bootstrap Δ PR-AUC with a 95 % confidence interval.

Require: Predictions $\hat{\mathbf{y}}^A, \hat{\mathbf{y}}^B$ of methods A and B over n paired test customers, labels $\mathbf{y}$; replicates B; confidence level $1 - \alpha$

Ensure: $\hat{\Delta}$, 95 % CI $[\Delta_{\text{lo}}, \Delta_{\text{hi}}]$

$\hat{\Delta} \leftarrow$ `PR-AUC`$(\hat{\mathbf{y}}^A, \mathbf{y})$ − `PR-AUC`$(\hat{\mathbf{y}}^B, \mathbf{y})$
for $b = 1$ to B **do**
    $\mathbf{i}_b \leftarrow$ `SampleWithReplacement`$(\{1, \ldots, n\})$
    $\hat{\Delta}^{(b)} \leftarrow$ `PR-AUC`$(\hat{\mathbf{y}}^A[\mathbf{i}_b], \mathbf{y}[\mathbf{i}_b])$ − `PR-AUC`$(\hat{\mathbf{y}}^B[\mathbf{i}_b], \mathbf{y}[\mathbf{i}_b])$
end for
$\Delta_{\text{lo}}, \Delta_{\text{hi}} \leftarrow$ percentile-bootstrap interval of $\{\hat{\Delta}^{(b)}\}$
return $\hat{\Delta}$, $[\Delta_{\text{lo}}, \Delta_{\text{hi}}]$

5. Experimental setup

The protocol is applied to eight detectors representative of the inductive-bias spectrum commonly used on smart-meter data [2,7–10,14,16]. The eight detectors are a bidirectional gated recurrent unit (BiGRU), a bidirectional long short-term memory (BiLSTM) network, and a unidirectional long short-term memory (LSTM) network. The remaining five are a temporal convolutional network (TCN), the Wide and Deep Convolutional Neural Network (CNN), a lightweight Transformer (Transformer-lite), the extreme gradient boosting (XGBoost) classifier on engineered window features [20,64], and an ℓ_2-regularized logistic regression on the same engineered features. The deep learning models share a common training procedure. The optimizer is AdamW with learning rate 10^{-3}, except for Transformer-lite at 5×10^{-4}, weight decay 10^{-4}, and gradient-norm clipping at 1.0. The batch size is 64, and training runs for at most 25 epochs and stops when the validation PR-AUC has not improved for six consecutive epochs, after which the best checkpoint is restored. The loss is focal loss with $\gamma = 2$ and $\alpha = 0.25$. Dropout is set per architecture between 0.10 and 0.20, as given in Table 4. Class imbalance is not treated uniformly across the panel: the deep learning models draw training minibatches through a with-replacement sampler weighted by inverse class frequency, in addition to the focal loss; the gradient-boosted model sets `scale_pos_weight` to the negative-to-positive ratio of the training fold; and the logistic regression uses balanced class weights. Because these three treatments differ, the panel is comparable in evaluation protocol rather than in imbalance handling, and we report the mechanism for each detector rather than a single shared setting. Hyperparameters follow each architecture's published reference and are not tuned per fold, since per-fold tuning would confound fold-level overfitting with architectural improvement.

Tables 4 and 5 detail the configurations for the deep detectors and classical baselines, respectively. Every deep detector receives the same sequence input: consumption with missing values set to zero, an observed-value mask, and a per-row z-score. Dimensions and parameter counts are quoted at the full patched input length $T = 1{,}035$, giving

Table 4
Deep detector configurations.

Detector	Structure	Parameters
BiGRU	non-overlapping 4-day patches ($T' = 259$), linear map to 64, two bidirectional GRU layers (hidden 64, dropout 0.20), mean pooling, LayerNorm, linear head	125,633
BiLSTM	as BiGRU, with LSTM cells	167,105
LSTM	as BiGRU, unidirectional, hidden 64	67,585
TCN	average pooling by 4 ($T' = 258$), six residual blocks (48 channels, kernel 5, dilation 2^i, causal, dropout 0.15), mean pooling, linear head	128,257
CNN	wide-and-deep: wide branch (FC to 90), deep branch (weekly reshape, 5 Zheng differencing convs $\gamma = 15$, global max pool, FC to 60), joint linear head	288,961
Transformer-lite	65 patches of length 16, linear map to $d_{\text{model}} = 64$, learned positional embedding, two pre-norm encoder layers (4 heads, feed-forward 128, GELU, dropout 0.10), mean pooling, LayerNorm, linear head	74,689

Table 5
Classical baseline configurations.

Baseline	Configuration
XGBoost	histogram tree method; 800 boosting rounds, maximum depth 6, learning rate 0.04, row subsample 0.85, column subsample 0.85, minimum child weight 2, ℓ_2 penalty 2; early stopping after 40 rounds without improvement in validation PR-AUC; `scale_pos_weight` $= n_-/n_+$ from the training fold
Logistic regression	features standardized with training-fold statistics; lbfgs solver, ℓ_2 penalty, $C = 1$, at most 500 iterations, balanced class weights

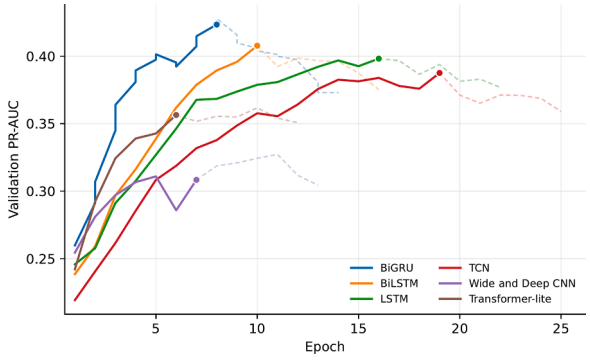

Fig. 5. Per-epoch validation PR-AUC of the six deep learning models, evaluated on fold 0 of the customer-grouped protocol (seed 42). The selected epoch for best-checkpoint restoration is marked per detector. Dashed, dimmed segments indicate the early-stopping patience window. This diagnostic changes no reported number.

$(N, 1{,}035, 3)$; window-restricted arms shorten the time axis accordingly. T' is the sequence length after downsampling. The TCN depth is derived in code from the sequence length, so that the receptive field covers the downsampled series; at $T = 1{,}035$ this yields the six blocks listed above. The classical baselines both consume the same 31 engineered features: 14 statistics of observed consumption, three missingness descriptors, the observed-zero ratio, 12 window aggregates, and one row-unobservable indicator. Fig. 5 illustrates the training convergence for all detectors.

The gradient-boosted configuration is evaluated at five additional depths, 3, 4, 8, 10, and 12, with every other setting fixed. Customer-grouped Tier-1 F1 spans 0.422 to 0.434 and PR-AUC 0.405 to 0.429 across the five depths, against 0.436 and 0.430 at the reported depth 6. Performance is flat from depth 4 upward: the depth setting neither constrains the model at the shallow end by more than 0.014 F1 nor amplifies the regime shortcut at the deep end.

5.1. Stress-test validation design

The audit applies the four validation gates jointly. For each detector the audit produces two numbers. The reported PR-AUC is the value under the unaudited evaluation protocol that the literature most commonly uses, namely Tier 1 with random k-fold. The audited PR-AUC is the value under the proposed protocol, namely Tier 2 with customer-grouped k-fold and training-fold-scoped preprocessing. The reported-versus-audited delta is a measure of the reliability of the reported number. A positive delta indicates that the reported number was inflated relative to the audit-verified number, and the magnitude of the delta is the drop in audited PR-AUC when the detector is evaluated without regime-aware preprocessing.

Three random seeds $\{7, 42, 2024\}$ combined with five folds yield $n = 15$ evaluation runs per (model, protocol, tier) combination. Figs. 6–8 report the split-integrity audit of the customer-grouped protocol, computed from the shipped run's split artifacts; the splits are identical for all eight models. Fig. 6 verifies customer-disjointness directly: in each of the 15 (seed, fold) cells, the training side, training and validation rows together, shares zero customers with the test fold, and each seed's five test folds partition all 42,372 customers, so every customer is scored exactly once per seed. This is the procedural half of Gate B(i); whether grouping changes what is measured is a separate, behavioral question, answered in Section 6.2. Fig. 7 audits stratification: every test fold's theft prevalence lies in 8.52 to 8.54 %, against a population rate of 8.53 %. The audit's ranking metric makes this a load-bearing property, because the chance level of PR-AUC equals the positive rate: a fold that drifted in prevalence would move PR-AUC with no change in detector behavior, and the flat profile removes that artifact from every cross-fold comparison in Section 6. Fig. 8 audits fold balance: test folds span 8471 to 8478 customers, so the five audited readings per seed weight the population evenly and no fold enters the $n = 15$ statistics with disproportionate mass. Table 6 summarizes the protocol parameters.

6. Results and discussion

The results that follow are organized by gate, in the order in which the gates are defined in Section 4. Gate A reports the outcome of the data-provenance audit; Gate B (i) reports customer-grouped split leakage; Gate B (ii) reports the chronological-stress audit; Gate C reports leakage detection gate compliance. Gate D is reported in four parts, namely the reported-versus-audited inflation across deep families, the

Fig. 6. Customer-disjointness verification for the customer-grouped protocol (full-population arm) across the 15 (seed, fold) cells: the training side (training ∪ validation) shares zero customers with the test fold in every cell; sub-labels give test-fold sizes. Each seed's five test folds partition all 42,372 customers. Splits are identical across models.

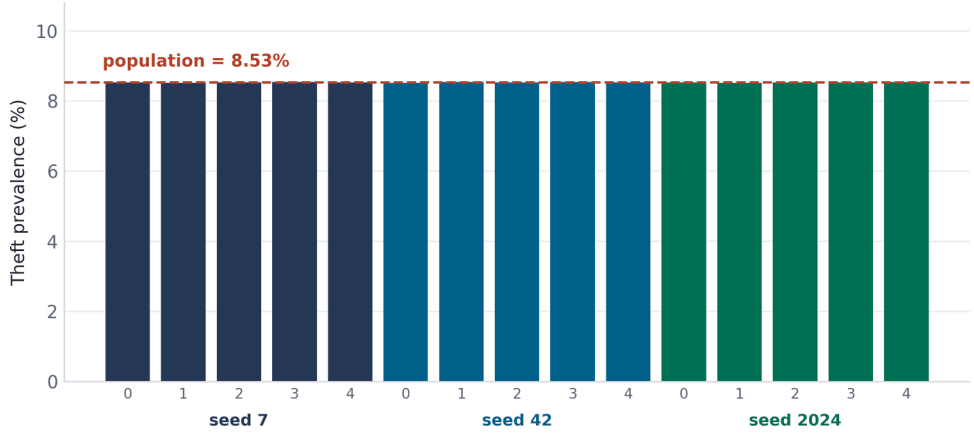

Fig. 7. Theft prevalence per test fold of the customer-grouped protocol across the three seeds. Every fold lies in 8.52–8.54 %, against a population rate of 8.53 % (dashed line).

Table 6
Protocol parameters.

Parameter	Symbol	Value
Random seed	s	7, 42, 2024
Cross-validation folds	K	5
Stratified within customer groups	–	yes
Bootstrap iterations	B	5000
Confidence level	$1-\alpha$	0.95
Decision-threshold criterion	–	validation F1, smoothed
Preprocessing scope	–	training-fold only
Feature interface	–	tabular missingness + observation block

Table 7
Random versus customer-grouped k-fold PR-AUC contrast per architecture. All eight paired tests return $p > 0.48$ (observed $\max|\Delta| = 0.0035$).

Architecture	Δ PR-AUC	p-value	95 % CI	n
BiLSTM	0.0002	0.9759	[−0.0324, 0.0291]	15
TCN	0.0003	0.9758	[−0.0680, 0.0625]	15
BiGRU	−0.0012	0.8861	[−0.0482, 0.0398]	15
Transformer-lite	−0.0019	0.7581	[−0.0345, 0.0405]	15
LSTM	−0.0024	0.7731	[−0.0667, 0.0322]	15
CNN	0.0055	0.8644	[−0.0709, 0.1078]	15
Logistic Regression	−0.0033	0.5796	[−0.0408, 0.0314]	15
XGBoost	0.0035	0.4680	[−0.0297, 0.0322]	15

three-tier reframing of the operating point, the protocol-sensitivity baseline on a fixed model, and the subgroup reliability audit. A mechanism attribution for the Gate D inflation is presented after the second part, and the consolidated verdict across the four gates is presented last. The discussion then synthesizes the findings across the four gates, interprets them within the trustworthy artificial intelligence frame, and compares the audit with the wider electricity-theft-detection literature.

6.1. Gate A: outcome of the data-provenance audit

The eight integrity metrics of Gate A, defined in Section 4.1, all return a pass on the State Grid Corporation of China dataset. The outcome of every metric is reported in Table 2 of Section 4.1. With Gate A satisfied, the audit proceeds to Gate B.

6.2. Gate B (i): no significant customer-grouped split leakage

Table 7 reports the audited contrast between random k-fold and customer-grouped k-fold under the temporal integrity check. None of the eight paired tests is statistically significant. The largest difference in magnitude is +0.0035 for XGBoost. The smallest is +0.0002 for BiLSTM. The smallest p-value across the eight architectures is 0.47. The maximum |Δ PR-AUC| observed across the eight tests is 0.0035, well below the standard deviation reported for every architecture in Table 9.

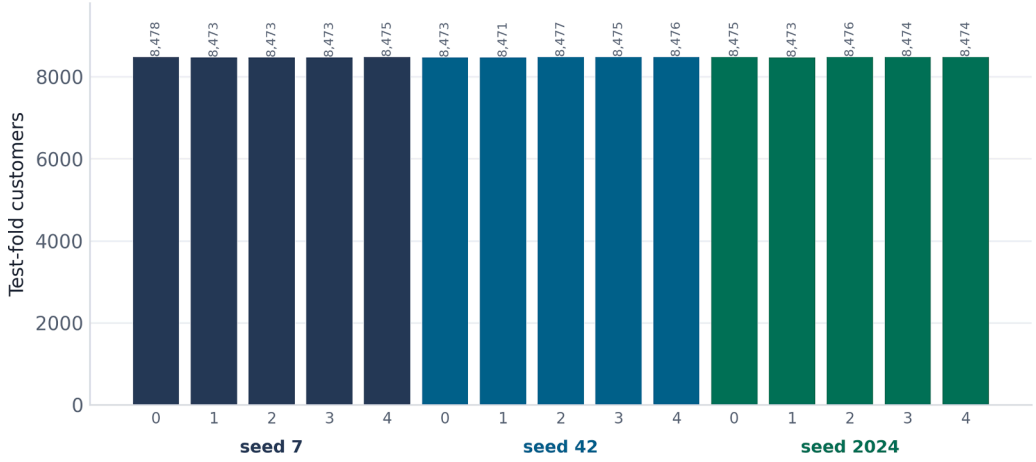

Fig. 8. Test-fold customer counts of the customer-grouped protocol across the 15 (seed, fold) cells, spanning 8,471–8,478; each customer is tested exactly once per seed.

The size of this contrast is set by how much duplicate structure the dataset actually contains. Grouping is over exact-duplicate consumption profiles, of which eleven groups with more than one member exist. The largest, comprising 2,114 identical all-zero profiles, exceeds the maximum group size of ten and is dissolved into singletons before splitting, leaving ten groups spanning twenty-two customers, or 0.05 % of the population. A constraint that binds 0.05 % of the rows cannot move a fold-level metric appreciably, and the near-zero contrast in Table 7 is what that sparsity predicts. The Class 2 vulnerability of Section 2.5 is therefore operationally negligible on this dataset, not because grouped splitting is unnecessary in general but because this benchmark holds almost no duplicate structure for it to constrain. The check retains diagnostic value: establishing that the duplicate structure is sparse is what allows the inflation reported under Gate D to be attributed to the regime boundary rather than to split construction. The leakage that matters on this dataset is structural across the regime boundary, not within the customer-by-time grid. An engineering audit that controls only for customer-grouped leakage does not certify the detector for deployment.

6.3. Gate B (ii): the out-of-time protocols measure the dominant unreported cost

Table 8 reports the three out-of-time protocols of Section 4.2 against the grouped k-fold reference under Tier 1. PR-AUC is reported per architecture as the mean±std over 15 folds. For these evaluations, the same-customer and customer-disjoint protocols train on the season-aligned window 2015-01-01 to 2015-10-31 and test on 2016-01-01 to 2016-10-31; the training-window control tests held-out customers inside the training window. The Δ represents the paired-bootstrap mean difference (customer-disjoint - grouped). Three facts emerge. First, the chronological cost is large and uniform where the regime shortcut was available: the six detectors whose Gate D inflation spans 0.109 to 0.134 lose 42.0 to 50.5 % of their grouped-protocol PR-AUC under the customer-disjoint protocol, logistic regression loses 18.9 %, and the CNN, which held no shortcut to lose, is unchanged (+0.9 %, interval spanning zero). Second, the same-customer protocol tracks the customer-disjoint protocol within 0.005 PR-AUC on every detector: scoring the customers the model trained on is no easier than scoring customers it has never seen, so the chronological cost is not a customer-identity effect. Third, the training-window control, which holds window length, season alignment, and customer-disjointness fixed while removing the regime crossing, scores below the crossing protocol for every detector (XGBoost 0.197 against 0.242); the crossing itself is therefore not the source of the loss, which traces to the reduction from the 1,035-day mixed-regime window to a single 304-day window. A grouped k-fold certificate on the full window says nothing about either chronological setting.

6.4. Gate C: leakage detection gate compliance

Gate C, defined in Section 4.3, is procedural rather than empirical. It verifies that the preprocessing pipeline and any synthetic-oversampling operation use no information from the test data. The experimental pipeline specified in Section 5 satisfies all four requirements. Re-verifying this gate for the present revision found that the implemented tabular pipeline did not meet the first requirement. The column means that fill undefined feature values were computed within each split, so validation and test features were filled with their own folds' means rather than the training fold's. The revised pipeline fits the imputer on the training fold only and reuses it unchanged for validation and test, and every reported number regenerates under the corrected pipeline in this revision. The defect touched the engineered features of the two classical baselines only; the sequence inputs of the deep learning models carry no cross-customer statistic. The feature-scaling statistics are estimated on the training fold only. Synthetic oversampling is not applied in the present audit experiments; if applied, it would be performed within each training fold rather than before the cross-validation split. The cross-validation indices, preprocessing parameters, and bootstrap seeds are stored as immutable artifacts. Gate C returns a pass for the revised pipeline, and the audit proceeds to Gate D.

6.5. Gate D (i): reported-versus-audited inflation across architectures

Table 9 reports the reported-versus-audited inflation per detector. The Tier 1 column is the PR-AUC under the unaudited Tier 1 protocol (full population, mixed regime). The Tier 2 column is the PR-AUC under the audited Tier 2 protocol (stable post-2016 regime). Six detectors show relative inflation between 28.5 and 32.6 % (absolute: 0.109 to 0.134), with 95 % paired-bootstrap intervals strictly below zero. The LSTM shows the largest inflation, $\Delta = -0.134$ with CI $[-0.145, -0.123]$, the BiGRU by -0.130 with CI $[-0.140, -0.119]$, XGBoost by -0.125 with CI $[-0.132, -0.118]$, the BiLSTM by -0.123 with CI $[-0.133, -0.114]$, the TCN by -0.117 with CI $[-0.126, -0.108]$, the Transformer-lite by -0.109 with CI $[-0.120, -0.098]$, and the Wide and Deep CNN by $+0.026$. One detector falls outside this range. Logistic regression shows 12.3 % inflation, 0.034 absolute, below the 0.053 protocol-sensitivity floor of Section 6.8, confirming that linear models lack the capacity to fully exploit the non-linear regime shortcut.

The inflation magnitudes establish that the engineering audit protocol is not a procedural refinement but a quantitatively substantial com-

Table 8
Tier 1 PR-AUC contrast between out-of-time protocols and the grouped k-fold reference.

Model	Grouped	Same-customer	Customer-disjoint	Training-window control	Δ	Rel. change
BiGRU	0.446±0.022	0.221±0.012	0.221±0.021	0.197±0.009	−0.225	−50.5 %
BiLSTM	0.421±0.017	0.215±0.010	0.215±0.017	0.190±0.013	−0.207	−49.0 %
LSTM	0.411±0.017	0.217±0.008	0.218±0.015	0.190±0.012	−0.193	−47.0 %
XGBoost	0.430±0.015	0.247±0.008	0.242±0.014	0.197±0.011	−0.188	−43.7 %
TCN	0.410±0.022	0.237±0.008	0.238±0.013	0.201±0.012	−0.172	−42.0 %
Transformer-lite	0.368±0.017	0.195±0.016	0.197±0.022	0.186±0.011	−0.171	−46.4 %
Logistic Regression	0.280±0.013	0.227±0.007	0.227±0.013	0.165±0.009	−0.053	−18.9 %
CNN	0.218±0.037	0.219±0.011	0.220±0.010	0.160±0.015	0.002	0.9 %

Table 9
Tier 1 versus Tier 2 PR-AUC per architecture (customer-disjoint 5-fold CV). Δ: paired-bootstrap mean difference (Tier 2 - Tier 1).

Model	Tier 1 (mean±std)	Tier 2 (mean±std)	Δ	Rel. change
LSTM	0.4110±0.0169	0.2771±0.0166	−0.1339	−32.58 %
BiGRU	0.4462±0.0222	0.3166±0.0150	−0.1296	−29.05 %
XGBoost	0.4299±0.0147	0.3051±0.0140	−0.1248	−29.03 %
BiLSTM	0.4214±0.0174	0.2985±0.0217	−0.1229	−29.17 %
TCN	0.4101±0.0219	0.2931±0.0145	−0.1170	−28.53 %
Transformer-lite	0.3680±0.0170	0.2592±0.0159	−0.1088	−29.56 %
Logistic Regression	0.2797±0.0135	0.2453±0.0164	−0.0344	−12.31 %
CNN	0.2181±0.0373	0.2445±0.0242	0.0264	12.10 %

ponent of the audit. A 0.43 PR-AUC detector reported under an unaudited Tier 1 evaluation reads, under the stable-regime audit, as a 0.32 PR-AUC detector. That drop of nearly one-third is the inflation that the audit quantifies on this dataset.

The Tier 1–Tier 2 contrast is not, by itself, a single-factor experiment. Tier 2 re-runs the full protocol, training, validation, and testing, inside the 2016 window, so the regime composition changes and every customer's record shortens from the full 1,035-day observation window to its 2016 portion; the customers, the training-set size, the labels, and the class balance are identical to Tier 1, because the window restricts days, not rows. The attribution of the six detectors' drop to the regime shortcut therefore rests on paired controls over these held-fixed factors. The logistic regression consumes the same 31 engineered features as XGBoost, and its 0.034 contrast, already below the protocol-sensitivity floor of Section 6.8, bounds the shared tabular-input cost; the gap to XGBoost's 0.125 is what the added capacity extracted at Tier 1 that the stable window does not support, and Section 6.7 bounds the regime-correlated signal available in the shared feature space. The contrast also compresses the panel: the eight Tier-1 readings in Table 9 span 0.228, and the same eight at Tier 2 span 0.072; the separation between architectures largely disappears inside the stable window, consistent with a shared learnable signal available at Tier 1 and absent at Tier 2. The structural break is fixed in calendar time, so no arm can hold the regime fixed while restoring the full window length; the decomposition is therefore by bounds and controls rather than by a full factorial, and the window-length effects the audit can measure directly are attributed where they dominate: the chronological losses of Section 6.3 trace to the window reduction, not the regime crossing, by the training-window control.

6.6. Gate D (ii): the three-tier cohort contrast

Table 10 contrasts the three tiers using the gradient-boosted classifier (XGBoost with the histogram tree method). In this table, Cov. refers to the retained fraction of test rows. At Tier 1 the classifier reaches F1 = 0.436 and PR-AUC = 0.430. At Tier 2, with the protocol confined to the stable 2016 window, F1 falls to 0.339 and PR-AUC to 0.305. Tier 3 retains 84.8 % of test rows after the observability gates and the mixed-label duplicate exclusion; no test row is gated by its own label, and the label-quality rule acts on training rows only, retaining on average 13,426 of 21,551 per fold. The audited cohort does not read above the stable-regime tier: F1 is 0.324 under the observability gates alone and 0.315 with the label-quality rule, and cleaning the supervision does not raise the audited reading on this dataset.

The contrast establishes that a reported number is bounded by the tier on which it was measured, and the bound is one-directional on this dataset: every restriction that removes the regime shortcut or the ambiguous supervision lowers the reading. The same dataset and labels admit an F1 of 0.44 at the unaudited full-population Tier 1 and 0.32 on the audited Tier-3 cohort at 84.8 % coverage. The two numbers answer different evaluation questions. A utility committing inspection budget to inspect every flagged customer operates at Tier 1. A utility deploying an automated triage that defers ambiguous cases to manual review, operates at Tier 3, with the coverage fraction declaring the manual-review remainder. A reported number that does not specify the operating tier is incomplete.

6.7. Feature-space attribution of the Gate D inflation

Fig. 9 reports the top fifteen of the 31 tabular features of an out-of-time logistic-regression probe, trained on the 2015 window and tested on held-out customers in the 2016 window, ranked by the mean drop in average precision when the feature is permuted on the test fold, over the 15 (seed, fold) runs. The probe reproduces the shipped experiment: its mean test average precision of 0.2268 agrees with the customer-disjoint logistic PR-AUC of Table 8 to within 4×10^{-5}. The signal concentrates on consumption scale and variability computed over observed days: the standard deviation and the mean of observed consumption carry mean drops of 0.148 and 0.136, the mean absolute day-to-day change 0.076, and the missingness ratio contributes 0.037 directly, with five of the twelve window aggregates at 0.019 to 0.033 each; the per-customer maximum, by contrast, carries 0.0004. The direct missingness contribution is consistent with the observability pattern that shifted with the regime, illustrated in Figs. 10 and 11. The probe is restricted to the feature space shared by the two classical baselines, so it bounds the regime-correlated signal available to a linear model on that space rather than measuring how each deep learning model internally weights it; establishing the latter would require feature ablation on each architecture, a direction for future work.

6.8. Gate D (iii): protocol-sensitivity baseline on a fixed model

Table 11 reports a protocol-sensitivity audit in which only the evaluation protocol varies. The model is fixed, an ℓ_2-regularized logistic regression with $C = 1$ and balanced class weight, with standardization fitted on the training fold and no consumption clipping; the features and the test population are fixed; the four rows differ solely in the split protocol: random k-fold, customer-grouped k-fold, and the same-customer and customer-disjoint out-of-time protocols of Section 4.2. Because the class weighting and the preprocessing are identical across rows, the spread is attributable to protocol choice by construction. The fixed model scores 0.276 under the random and 0.280 under the customer-grouped k-fold, and 0.227 under both out-of-time protocols (evaluated as mean±std PR-AUC over 15 folds); the spread is 0.053, and the out-of-time rows sit lowest, consistent with the window-length attribution of Section 6.3. This spread establishes the protocol-sensitivity floor,

Table 10
Tier 1, 2, and 3 performance contrast on XGBoost under the revised protocol.

Tier	Cohort	F1	Prec.	Rec.	PR-AUC	Cov.
1	Full population, full window	0.4357	0.4321	0.4454	0.4299	1.000
2	Stable regime (2016 window)	0.3385	0.3147	0.3718	0.3051	1.000
3	+ observability gates, mixed-label duplicate exclusion	0.3236	0.3026	0.3548	0.2899	0.848
3	+ label-quality rule on training rows	0.3147	0.2765	0.3715	0.2500	0.848

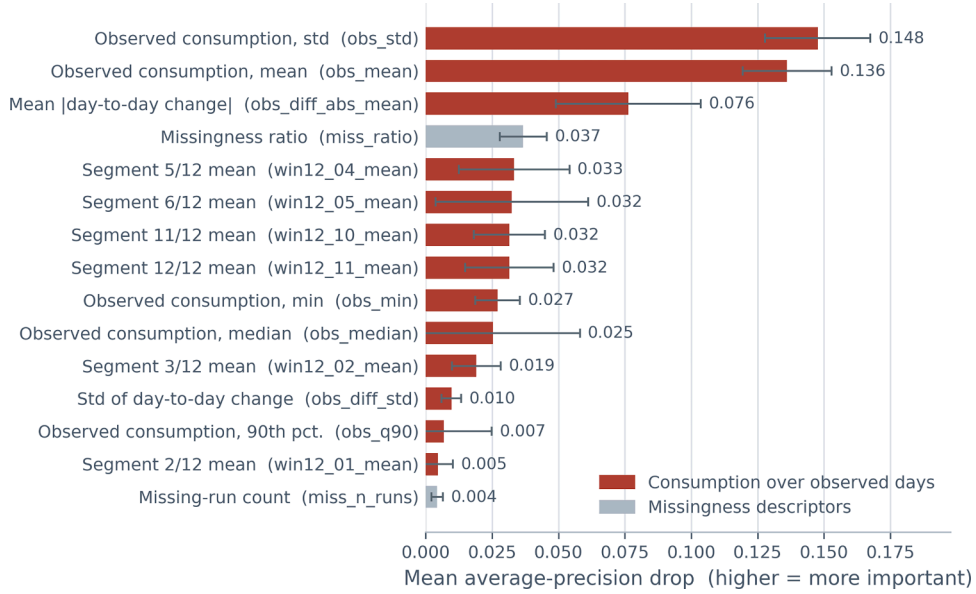

Fig. 9. Mean average-precision drop under permutation for the top fifteen of the 31 tabular features of the out-of-time logistic-regression probe; whiskers give one standard deviation over the 15 (seed, fold) runs. Red: consumption statistics computed over observed days; grey: missingness descriptors. Code names in parentheses.

Fig. 10. Wednesday-versus-Saturday valid-count gap on the 2016 cohort, by theft label.

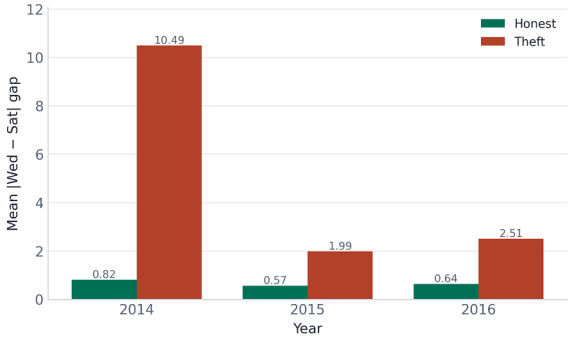

Fig. 11. Switching-coverage indicator by calendar year.

Table 11
Protocol-sensitivity floor on a fixed ℓ_2-regularized logistic regression.

Protocol	PR-AUC
Random k-fold	0.276±0.013
Customer-grouped k-fold	0.280±0.013
Same-customer out-of-time	0.227±0.007
Customer-disjoint out-of-time	0.227±0.013

namely the variation a fixed model exhibits from protocol choice alone; an inflation below this value cannot be distinguished from protocol sensitivity. The inflations of 0.109 to 0.134 absolute PR-AUC reported in Section 6.5 for the six detectors exceed the floor by 0.056 to 0.081, more than twice the floor at every point of the range; logistic regression's own contrast of 0.034 falls below it. The inflation is not an artifact of protocol choice. Two evaluation settings are excluded from the floor for cause: the Tier-2 stable-regime restriction is the quantity the inflation measures, and the training-window control of Section 6.3 is a diagnostic instrument, not a protocol under which a result would be reported.

6.9. *Gate D (iv): subgroup reliability audit and label-quality verification*

Table 12 reports the subgroup-level PR-AUC, precision at top 100, precision at top 500, and the false-positive rate at a fixed top-K operating point for the top six and bottom six of 22 subgroups, computed on the BiGRU detector under the stable-regime (Tier 2) protocol, pooled across the five held-out folds of one seed (42) so that each customer is scored exactly once. The subgroups are defined by five-way quantile bins of missingness over the 2016 window, the pre-to-post switching gap, valid count, mean consumption, missingness over the pre-2016 window, and zero rate. Bootstrap 95 % confidence intervals are computed with

Table 12
Subgroup reliability audit on the stable-regime 2016 window (BiGRU, seed 42): top and bottom six of 22 PR-AUC-ranked subgroups. FPR@K is the per-subgroup false-positive rate at a fixed prevalence-matched top-K operating point.

Subgroup	n	positives	PR-AUC [95 % CI]	p@100	p@500	FPR@K
miss_2016_bin = 2	7988	966	0.503 [0.470, 0.537]	0.86	0.658	0.0990
valid_cnt_bin = 0	10,590	1142	0.468 [0.437, 0.498]	0.88	0.646	0.0872
mean_bin = 4	8465	1509	0.426 [0.400, 0.452]	0.81	0.576	0.1984
switch_gap_2016_bin = 1	8492	1055	0.395 [0.366, 0.427]	0.76	0.530	0.0692
miss_pre_bin = 3	16,939	1879	0.367 [0.345, 0.389]	0.77	0.622	0.0672
switch_gap_2016_bin = 4	8361	720	0.352 [0.317, 0.391]	0.69	0.458	0.0691
miss_pre_bin = 1	8186	422	0.205 [0.171, 0.248]	0.43	0.238	0.0567
miss_2016_bin = 0	25,943	1882	0.200 [0.184, 0.218]	0.47	0.384	0.0512
switch_gap_2016_bin = 3	8490	422	0.195 [0.162, 0.234]	0.40	0.220	0.0487
valid_cnt_bin = 2	24,878	1785	0.194 [0.178, 0.212]	0.50	0.364	0.0504
mean_bin = 1	8465	514	0.167 [0.138, 0.199]	0.40	0.202	0.0282
mean_bin = 0	8465	273	0.071 [0.052, 0.097]	0.13	0.082	0.0065

$B = 5{,}000$ replicates. The subgroup audit shows that detection reliability varies by a factor of seven across subgroups on the same detector. The top-performing subgroup (high-missingness 2016 bin) shows PR-AUC 0.503 with CI $[0.470, 0.537]$ on 7,988 customers. The bottom subgroup (low mean-consumption bin) shows 0.071 with CI $[0.052, 0.097]$ on 8,465 customers. That spread on a fixed detector and dataset shows that full-population metrics obscure substantial heterogeneity in audited reliability. Part of this spread is composition rather than skill: the chance level of PR-AUC equals a subgroup's positive rate, and the subgroup prevalences span 3.2 to 17.8 %. Measured as lift over each subgroup's own chance level, the detector's skill still varies across subgroups by roughly a factor of two (2.2 to 4.3 times chance). Both readings matter operationally: the absolute spread is what an inspection queue experiences; the lift spread is what the detector itself contributes. The heterogeneity also propagates to the operating point: at a fixed prevalence-matched top-K threshold the per-subgroup false-positive rate spans 0.006 to 0.198, so a single global decision threshold imposes markedly unequal false-positive burdens across customer subgroups rather than the uniform burden a full-population number implies.

The label-quality audit complements the subgroup audit. Within each evaluation fold, an out-of-sample predicted probability of the observed label is computed for every training row by cross-validating a histogram gradient-boosted classifier over that fold's training rows, on the same engineered features as the classical baselines. A training row passes when this self-confidence reaches the per-class confident threshold, the mean self-confidence of its labeled class, following the confident-learning method of Northcutt et al. [65]; the rule reduces the training supervision to, on average, 13,426 of the 21,551 training rows retained by the observability gates per fold (Table 10). The score is a function of the observed label, so it exists only where a label exists: test rows are never scored or filtered by it; they are retained by the observability gates and the training-derived mixed-label duplicate exclusion of Algorithm 1, neither of which reads a test row's label. At deployment the same asymmetry holds: the rule cleans the supervision a utility already possesses from the labeled outcomes of its historical inspections, while incoming customers, which carry no label, are screened by the same gates, which read none, and ranked by the detector. The rule carries no tunable threshold: the per-class thresholds are determined by each fold's training rows, with the scorer, its inner cross-validation, and its seed fixed in the shipped code and configuration files.

6.10. Audit verdict across the four gates

The four gates return the following outcomes on the State Grid Corporation of China dataset. Gate A returns a pass on all eight data-provenance integrity metrics (Table 2). Gate B finds no significant contrast between random and customer-grouped split, on a dataset whose duplicate structure spans 0.05 % of customers, and reports out-of-time losses of 42 to 51 % of grouped-protocol PR-AUC for every competent detector. Gate C is satisfied by a preprocessing pipeline whose imputer and feature scaler are fitted on the training fold only and whose oversampling, where applied, is performed within each training fold rather than before the cross-validation split. Gate D returns a reported-versus-audited inflation of 28.5 % to 32.6 % (roughly a 30 % performance drop) across six of the eight detectors, with 12.3 % for logistic regression (below the protocol-sensitivity floor). The outlier is the Wide and Deep CNN, which exhibits a 12.1 % performance gain (+0.026). The training-window control removes the remaining distinction between the Tier-2 protocol and the non-chronological Tier-1 evaluation. In relative terms, Tier 3 retains 84.8 % of test rows with no test row gated by its own label and reads at the Tier-2 level; cleaning the training supervision does not raise the audited reading. Detection reliability varies by a factor of seven across the 22 subgroups on a single detector, in absolute PR-AUC; in lift over each subgroup's chance level the spread is roughly a factor of two. The certified audit performance on this dataset is therefore the Tier 2number, which ranges from 0.24 to 0.32 PR-AUC across the eight detectors. The Tier 1 value of 0.44 is, under this audit, leakage-inflated.

6.11. Discussion

Across the eight detectors and four leakage classes, the audit reveals a consistent pattern. The dominant inflation mechanism on the State Grid Corporation of China dataset is Class 1, regime-correlated preprocessing leakage: six of eight detectors lose 28.5 to 32.6 % of their reported PR-AUC (absolute: 0.109 to 0.134) under the leakage detection gate. Class 2 customer-grouped split leakage, by contrast, is operationally negligible, the largest paired difference reaching only 0.0035. The out-of-time protocols cost every competent detector 42 to 51 % of its grouped-protocol PR-AUC, a loss the training-window control traces to the shorter observation window rather than to the regime crossing, and the subgroup audit shows that detection reliability varies by a factor of seven on a single detector across 22 subgroups. The scale and variability of consumption over observed days, with a direct contribution from the missingness rate, emerge as the regime-correlated signal that the unaudited pipeline turns into a class-discriminative shortcut.

The protocol's value lies not in the PR-AUC gain of any specific detector but in whether the reliability claim attached to that detector is verifiable. A 0.43 PR-AUC detector verified under the four gates is a stronger candidate for deployment than a 0.99 detector that fails the temporal integrity check or the leakage detection gate. The reported-versus-audited delta quantifies this distinction: it tells the operator how much of a headline metric is audited performance and how much is leakage-induced inflation. Reporting it alongside any headline metric should be a minimum requirement for electricity-theft-detection papers that claim relevance to deployment.

Table 13
Post-hoc calibration of the eight detectors trained under the Tier-3 label-quality rule.

	Brier			ECE		
Detector	Raw	Isotonic	BSS (iso.)	Raw	Temp.	Isotonic
LSTM	0.119±0.011	0.075±0.001	0.084±0.014	0.112±0.017	0.111±0.026	0.008±0.002
BiGRU	0.127±0.016	0.075±0.001	0.087±0.013	0.128±0.028	0.135±0.042	0.008±0.002
BiLSTM	0.119±0.014	0.075±0.001	0.082±0.013	0.115±0.023	0.111±0.035	0.008±0.002
XGBoost	0.147±0.006	0.076±0.001	0.076±0.010	0.148±0.007	0.142±0.008	0.008±0.002
TCN	0.116±0.009	0.075±0.001	0.087±0.010	0.113±0.014	0.145±0.027	0.008±0.002
Transformer-lite	0.130±0.013	0.076±0.001	0.072±0.009	0.134±0.021	0.136±0.029	0.007±0.003
Logistic regression	0.168±0.006	0.076±0.001	0.070±0.008	0.173±0.007	0.301±0.012	0.007±0.002
Wide and Deep CNN	0.113±0.016	0.075±0.001	0.080±0.011	0.105±0.029	0.175±0.033	0.008±0.003

The four-gate audit is model-agnostic. The attention-based extension with dilated convolutions of Xia et al. [16], the convolutional autoencoder and Transformer combination of Le et al. [2], the multi-objective evolutionary hybrid model of Tursunboev et al. [7], the ensemble of Kabir et al. [9] optimized by the firefly algorithm, and the autocorrelation detector of Si et al. [8] can each be evaluated under it without modification. The protocol thereby separates methodological innovation from evaluation rigor: one paper can propose a new architecture, and another can validate an architecture's deployment readiness, without bundling the two. The survey of Section 2.8 shows the cost of bundling, a common reporting practice in which architectural novelty comes at the expense of audit discipline.

The present findings apply under a passive-theft threat model: the malicious customer underreports consumption through one of the canonical synthetic attack types of Punmiya and Choe [64] and Jokar et al. [13], without knowledge of the deployed detector. Three adjacent threat classes are out of scope. False-data-injection attacks against state estimation [5] target the supervisory control and data acquisition system rather than the smart meter. Clandestine connections that bypass the meter entirely belong to a complementary line of work that uses satellite imagery [36], and are undetectable by any consumption-based detector. Adversarial evasion crafted with knowledge of the decision boundary forms a third class; certifying robustness to it requires an evaluation protocol under a declared threat model, beyond the data audit developed here. The four gates do not certify robustness against any of these three classes; they certify data integrity under the passive-theft assumption.

A false-positive flag is not symmetric to a false negative: the flagged customer bears the on-site inspection, the temporary disconnection, and the reputational cost [29]. The subgroup audit of Section 6.9 measures this risk. A variation in detection reliability by a factor of seven, together with a per-subgroup false-positive rate that varies by more than an order of magnitude (0.006 to 0.198) at a fixed top-K operating point (Table 12), means a single global threshold imposes unequal precision and false-positive burden across customer subgroups. Eq. (4) quantifies the consequence: under the working assumption of a utility cost asymmetry of $c_{\mathrm{FN}}/c_{\mathrm{FP}} \approx 20$, the cost-optimal threshold falls below 0.05, and the Tier 3 cohort sits at a different operating point from the full population.

At that operating point, calibration of the underlying detector becomes essential. A PR-AUC number is unchanged by any strictly increasing transformation of the score and therefore says nothing about the calibration quality the cost-optimal threshold requires. The Brier score and the expected calibration error, rarely reported in the recent State Grid Corporation of China literature, are reported here (Table 13) and recommended as a thirteenth disclosure item. ECE uses 15 equal-width bins over $[0, 1]$. BSS is the Brier skill score against the constant base-rate forecast (no-skill Brier $= 0.082$). Computed on the Tier-3 audited cohort, the 84.8 % of test rows retained by the observability gates (the mixed-label duplicate exclusion removes no additional test rows on this dataset), over three seeds and five grouped folds, for the eight detectors trained under the Tier-3 label-quality rule, the expected calibration error on the raw detector probabilities ranges from 0.11 for the Wide and Deep CNN to 0.17 for logistic regression, with Brier scores from 0.08 to 0.17. Temperature scaling fitted on each validation fold does not repair this: it raises the expected calibration error in all 15 folds for the logistic regression and for the TCN. An isotonic map fitted on the same validation folds and applied unchanged to the test folds does: it reduces the expected calibration error to between 0.007 and 0.008 for all eight detectors. The repair is honest about what remains: against the constant base-rate forecast, whose Brier score is 0.082 at the cohort's 9.0 % prevalence, the calibrated probabilities carry a Brier skill score of 0.062 to 0.087, so the isotonic map makes the scores usable as probabilities without adding sharpness. The PR-AUC and precision values reported elsewhere in this paper are computed on the raw scores; the map is applied only where a probability enters a decision, as in the cost-optimal threshold of Eq. (4), which can then be evaluated on calibrated probabilities rather than read as approximate. Because the Tier 3 cohort definition is deterministic and auditable, any operational deployment of a Tier-3-gated detector should add a disparate-impact gate alongside the observability gates and the training-side label-quality rule of Algorithm 1.

The four gates transfer unchanged to any dataset with an identifiable structural break in its measurement process, and the 12-item checklist is dataset-agnostic; cross-utility replication is the next step. Kabir et al. [9] are the most closely related work, reporting on both the Pakistan Residential Electricity Consumption dataset and the State Grid Corporation of China dataset, though the regime structure of neither is examined. Shortcut features beyond the consumption-scale, variability, and missingness signals that the probe ranks highest cannot be ruled out without a controlled ablation study, a direction for future work. The natural extension is a regime-invariant training protocol that closes the 0.109 to 0.134 PR-AUC inflation while retaining model capacity, evaluated under the disclosure standard introduced here.

7. Conclusion

This paper reframes electricity-theft detection as a critical-infrastructure integrity problem. We have introduced a comprehensive engineering audit protocol featuring a four-class data leakage taxonomy, a three-tier evaluation standard, and a 12-item deployment checklist. This framework systematically exposes the inflation between academic reported performance and real-world audited performance.

Applying this audit to the State Grid Corporation of China dataset revealed significant vulnerabilities. Six of the eight tested detectors lost between 28.5 % and 32.6 % of their reported PR-AUC under strict auditing. A permutation-importance probe traced this performance loss directly to a regime-correlated shortcut, identifying consumption scale, variability, and missingness rates as the primary signals exploited by the models rather than genuine theft behaviors. Furthermore, testing models on future out-of-time data resulted in massive performance drops of 42 % to 51 %. Our controls demonstrated that this was caused by shorter observation windows rather than genuine concept drift. Finally, we demonstrated that proper auditing provides a realistic operational

picture: for the XGBoost model, the audited F1 score fell to 0.32 compared to its 0.44 unaudited score, while successfully retaining 84.8 % of valid test cases.

The true deployable capability of any detector is its performance under the strictest tier of this audit. The proposed protocol is model-agnostic and applies to any architecture. Future work includes standardizing this protocol for trustworthy artificial intelligence in critical infrastructure, testing cross-utility replication, and developing regime-invariant training methods to close the inflation gap. Our ultimate goal is not chasing higher detection rates on paper, but guaranteeing trustworthy and reliable deployment in the real world.

CRediT authorship contribution statement

Hafiz Muhammad Azeem Akram: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Validation, Software, Resources, Writing – original draft, Writing – review & editing; **Siamak Talatahari:** Writing – review & editing, Validation, Supervision, Methodology, Visualization; **Adil Jhangeer:** Writing – review & editing, Validation, Methodology, Funding acquisition, Visualization, Resources.

Funding

This article has been produced with the financial support of the European Union under the REFRESH – Research Excellence For Region Sustainability and High-tech Industries project, number CZ.10.03.01/00/22_003/0000048, via the Operational Programme Just Transition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The State Grid Corporation of China dataset is publicly available [14]. The source code is publicly available at https://github.com/azeem1020/electricity-theft-reproducibility.

References

[1] R.K. Ahir, B. Chakraborty, Pattern-based and context-aware electricity theft detection in smart grid, Sustain. Energy Grids Netw. 32 (2022) 100833. https://doi.org/10.1016/j.segan.2022.100833

[2] T.-D. Le, N.T.M. Duy, T.H.B. Huy, P. Van Phu, H.T. Doan, D. Kim, et al., Advanced deep learning-based electricity theft detection in smart grids using multi-dimensional analysis with convolutional autoencoder and transformer, Eng. Appl. Artif. Intell. 157 (2025) 111333. https://doi.org/10.1016/j.engappai.2025.111333

[3] M.S. Iqbal, S. Munawar, M. Adnan, A. Raza, M.A. Akbar, A. Bermak, et al., A critical review of technical case studies for electricity theft detection in smart grids: a new paradigm based transformative approach, Energy Convers. Manag.: X 26 (2025) 100965. https://doi.org/10.1016/j.ecmx.2025.100965

[4] J. Nagi, K.S. Yap, S.K. Tiong, S.K. Ahmed, F. Nagi, et al., Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system, IEEE Trans. Power Deliv. 26 (2) (2011) 1284–1285. https://doi.org/10.1109/TPWRD.2010.2055670

[5] Y. He, G.J. Mendis, J. Wei, et al., Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism, IEEE Trans. Smart Grid 8 (5) (2017) 2505–2516. https://doi.org/10.1109/TSG.2017.2703842

[6] D. Gu, Y. Gao, K. Chen, J. Shi, Y. Li, Y. Cao, et al., Electricity theft detection in AMI with low false positive rate based on deep learning and evolutionary algorithm, IEEE Trans. Power Syst. 37 (6) (2022) 4568–4578. https://doi.org/10.1109/TPWRS.2022.3150050

[7] J. Tursunboev, V. Palakonda, J.-M. Kang, et al., Multi-objective evolutionary hybrid deep learning for energy theft detection, Appl. Energy 363 (2024) 122847. https://doi.org/10.1016/j.apenergy.2024.122847

[8] Z. Si, Z. Liu, C. Mu, M. Wang, T. Gong, X. Xia, Q. Hu, Y. Xiao, et al., A new deep learning based electricity theft detection framework for smart grids in cloud computing, Comput. Stand. Interfaces 94 (2025) 104007. https://doi.org/10.1016/j.csi.2025.104007

[9] B. Kabir, U. Qasim, N. Javaid, F.A. Khan, B. Mansoor, A.K.J. Saudagar, H.S. Al-Sagri, M.N. Saqib, et al., Detecting electricity theft in smart grids using optimized machine learning ensemble techniques and eXplainable AI, Energy Rep. 14 (2025) 3142–3162. https://doi.org/10.1016/j.egyr.2025.09.019

[10] J. Chen, Y.A. Nanehkaran, W. Chen, Y. Liu, D. Zhang, et al., Data-driven intelligent method for detection of electricity theft, Int. J. Electr. Power Energy Syst. 148 (2023) 108948. https://doi.org/10.1016/j.ijepes.2023.108948

[11] C.-H. Lo, N. Ansari, CONSUMER: a novel hybrid intrusion detection system for distribution networks in smart grid, IEEE Trans. Emerg. Top. Comput. 1 (1) (2013) 33–44. https://doi.org/10.1109/TETC.2013.2274043

[12] L.M.R. Raggi, F.C.L. Trindade, V.C. Cunha, W. Freitas, et al., Non-technical loss identification by using data analytics and customer smart meters, IEEE Trans. Power Deliv. 35 (6) (2020) 2700–2710. https://doi.org/10.1109/TPWRD.2020.2974132

[13] P. Jokar, N. Arianpoo, V.C.M. Leung, et al., Electricity theft detection in AMI using customers' consumption patterns, IEEE Trans. Smart Grid 7 (1) (2016) 216–226. https://doi.org/10.1109/TSG.2015.2425222

[14] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, Y. Zhou, et al., Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids, IEEE Trans. Ind. Inform. 14 (4) (2018) 1606–1615. https://doi.org/10.1109/TII.2017.2785963

[15] T. Ahmad, R. Madonski, D. Zhang, C. Huang, A. Mujeeb, et al., Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: key developments, challenges, and future research opportunities in the context of smart grid paradigm, Renew. Sustain. Energy Rev. 160 (2022) 112128. https://doi.org/10.1016/j.rser.2022.112128

[16] R. Xia, Y. Gao, Y. Zhu, D. Gu, J. Wang, et al., An attention-based wide and deep CNN with dilated convolutions for detecting electricity theft considering imbalanced data, Electr. Power Syst. Res. 214 (2023) 108886. https://doi.org/10.1016/j.epsr.2022.108886

[17] F. Wang, S. Zhou, C. Wang, D. Meng, et al., MSDM: multi-scale differencing modeling for cross-scenario electricity theft detection, IEEE Trans. Smart Grid 16 (2) (2025) 1629–1640. https://doi.org/10.1109/TSG.2024.3455368

[18] W. Liao, R. Zhu, Z. Yang, K. Liu, B. Zhang, S. Zhu, B. Feng, et al., Electricity theft detection using dynamic graph construction and graph attention network, IEEE Trans. Ind. Inform. 20 (4) (2024) 5074–5086. https://doi.org/10.1109/TII.2023.3331131

[19] W. Liao, Z. Yang, K. Liu, B. Zhang, X. Chen, R. Song, et al., Electricity theft detection using euclidean and graph convolutional neural networks, IEEE Trans. Power Syst. 38 (4) (2023) 3514–3527. https://doi.org/10.1109/TPWRS.2022.3196403

[20] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, S. Mishra, et al., Decision tree and SVM-based data analytics for theft detection in smart grid, IEEE Trans. Ind. Inform. 12 (3) (2016) 1005–1016. https://doi.org/10.1109/TII.2016.2543145

[21] X. Xia, J. Lin, Q. Jia, X. Wang, C. Ma, J. Cui, W. Liang, et al., ETD-ConvLSTM: a deep learning approach for electricity theft detection in smart grids, IEEE Trans. Inf. Forensics Secur. 18 (2023) 2553–2568. https://doi.org/10.1109/TIFS.2023.3265884

[22] R.U. Madhure, R. Raman, S.K. Singh, et al., CNN-LSTM based electricity theft detector in advanced metering infrastructure, in: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225572

[23] N. Shahzadi, N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, S.H. Bouk, et al., A novel data driven approach for combating energy theft in urbanized smart grids using artificial intelligence, Expert Syst. Appl. 253 (2024) 124182. https://doi.org/10.1016/j.eswa.2024.124182

[24] R. Qi, Q. Li, Z. Luo, J. Zheng, S. Shao, et al., Deep semi-supervised electricity theft detection in AMI for sustainable and secure smart grids, Sustain. Energy Grids Netw. 36 (2023) 101219. https://doi.org/10.1016/j.segan.2023.101219

[25] R. Qi, J. Zheng, Z. Luo, Q. Li, et al., A novel unsupervised data-driven method for electricity theft detection in AMI using observer meters, IEEE Trans. Instrum. Meas. 71 (2022) 1–10. https://doi.org/10.1109/TIM.2022.3189748

[26] F. Tian, Y. Zhang, Q. Miao, et al., DKGT-net: dual-branch temporal network for electricity theft detection in power system, IEEE Trans. Ind. Inform. 22 (4) (2026) 2734–2745. https://doi.org/10.1109/TII.2025.3645592

[27] Y. Wang, Z. Wu, J. Ma, Q. Jin, et al., MFFTD: a multiscale feature fusion transformer detector for electricity theft based on semi-supervised learning, IEEE Trans. Instrum. Meas. 74 (2025) 1–10. https://doi.org/10.1109/TIM.2025.3552857

[28] L. Cui, F. Wu, Y. Qu, B. Gu, L. Gao, S. Yu, et al., RE4ETD: a relative entropy optimization-based method for efficient electricity theft detection with dual-privacy preservation, IEEE Trans. Smart Grid 16 (5) (2025) 4073–4086. https://doi.org/10.1109/TSG.2025.3570336

[29] S. Pelekis, E. Karakolis, G. Lampropoulos, S. Mouzakitis, O. Markaki, C. Ntanos, D. Askounis, et al., Trustworthy artificial intelligence in the energy sector: landscape analysis and evaluation framework, in: 2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC), IEEE, Funchal, Portugal, 2024, pp. 1–10. https://doi.org/10.1109/ICE/ITMC61926.2024.10794222

[30] W. Liao, S. Zhu, L. Ge, T. Ishizaki, R. Qi, Y. Yang, Z. Yang, et al., A model-agnostic approach to reveal detection basis and occurrence time of electricity theft, IEEE Trans. Instrum. Meas. 74 (2025) 1–11. https://doi.org/10.1109/TIM.2025.3595249

[31] X. Cui, S. Liu, Z. Lin, J. Ma, F. Wen, Y. Ding, L. Yang, W. Guo, X. Feng, et al., Two-step electricity theft detection strategy considering economic return based on convolutional autoencoder and improved regression algorithm, IEEE Trans. Power Syst. 37 (3) (2022) 2346–2359. https://doi.org/10.1109/TPWRS.2021.3114307

[32] J. Pereira, F. Saraiva, A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection, in: 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, pp. 1–8. https://doi.org/10.1109/CEC48606.2020.9185822

[33] J. Tao, G. Michailidis, A statistical framework for detecting electricity theft activities in smart grid distribution networks, IEEE J. Sel. Areas Commun. 38 (1) (2020) 205–216. https://doi.org/10.1109/JSAC.2019.2952181
[34] P.K. Reddy Shabad, A. Alrashide, O. Mohammed, et al., Anomaly detection in smart grids using machine learning, in: IECON 2021 - 47th Annual Conference of the IEEE Industrial Electronics Society, 2021, pp. 1–8. ISSN: 2577-1647, https://doi.org/10.1109/IECON48115.2021.9589851
[35] Z. Yan, H. Wen, Performance analysis of electricity theft detection for the smart grid: an overview, IEEE Trans. Instrum. Meas. 71 (2022) 1–28. https://doi.org/10.1109/TIM.2021.3127649
[36] A.M. Kaminski, F.G. Carloto, M.M. Jacques, V.J. Garcia, O.C. Filho, C.H. Barriquello, et al., Prospecting electricity theft through assignment of AI detected buildings to registered consumer units, Sustain. Energy Grids Netw. 41 (2025) 101609. https://doi.org/10.1016/j.segan.2024.101609
[37] X. Xia, Y. Xiao, W. Liang, J. Cui, et al., Detection methods in smart meters for electricity thefts: a survey, Proc. IEEE 110 (2) (2022) 273–319. https://doi.org/10.1109/JPROC.2021.3139754
[38] A. Takiddin, M. Ismail, U. Zafar, E. Serpedin, et al., Robust electricity theft detection against data poisoning attacks in smart grids, IEEE Trans. Smart Grid 12 (3) (2021) 2675–2684. https://doi.org/10.1109/TSG.2020.3047864
[39] A. Takiddin, M. Ismail, E. Serpedin, et al., Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids, IEEE Trans. Smart Grid 14 (1) (2023) 663–676. https://doi.org/10.1109/TSG.2022.3193989
[40] I.U. Khan, N. Javaid, C.J. Taylor, X. Ma, et al., Robust data driven analysis for electricity theft attack-resilient power grid, IEEE Trans. Power Syst. 38 (1) (2023) 537–548. https://doi.org/10.1109/TPWRS.2022.3162391
[41] I.U. Khan, A. Ali, C.J. Taylor, X. Ma, et al., Data-driven insights: boosting algorithms to uncover electricity theft patterns in AMI, IEEE Trans. Instrum. Meas. 74 (2025) 1–12. https://doi.org/10.1109/TIM.2025.3557097
[42] G.M. Messinis, A.E. Rigas, N.D. Hatziargyriou, et al., A hybrid method for non-technical loss detection in smart distribution grids, IEEE Trans. Smart Grid 10 (6) (2019) 6080–6091. https://doi.org/10.1109/TSG.2019.2896381
[43] A.N. Alkuwari, A. Albaseer, S. Al-Kuwari, M. Qaraqe, et al., Robust convLSTM model with deep reinforcement learning for stealth attack detection in smart grids, IEEE Open J. Ind. Electron. Soc. 6 (2025) 1298–1311. https://doi.org/10.1109/OJIES.2025.3594618
[44] G. Fenza, M. Gallo, V. Loia, et al., Drift-aware methodology for anomaly detection in smart grid, IEEE Access 7 (2019) 9645–9657. https://doi.org/10.1109/ACCESS.2019.2891315
[45] T. Berghout, M. Benbouzid, M.A. Ferrag, et al., Multiverse recurrent expansion with multiple repeats: a representation learning algorithm for electricity theft detection in smart grids, IEEE Trans. Smart Grid 14 (6) (2023) 4693–4703. https://doi.org/10.1109/TSG.2023.3250521
[46] L. Zhu, W. Wen, J. Li, C. Zhang, B. Zhou, Z. Shuai, et al., Deep active learning-enabled cost-effective electricity theft detection in smart grids, IEEE Trans. Ind. Inform. 20 (1) (2024) 256–268. https://doi.org/10.1109/TII.2023.3249212
[47] Z. Zhao, Y. Liu, Z. Zeng, Z. Chen, H. Zhou, et al., Privacy-preserving electricity theft detection based on blockchain, IEEE Trans. Smart Grid 14 (5) (2023) 4047–4059. https://doi.org/10.1109/TSG.2023.3246459
[48] T. Sheng, X. Gui, S. Xiong, D. Zheng, et al., Privacy-preserving electricity theft detection training model based on convolutional neural network, IEEE Trans. Smart Grid 17 (2) (2026) 1561–1573. https://doi.org/10.1109/TSG.2025.3630616
[49] F. Guo, Y. Lang, S. Li, C. Dong, Q. Wu, W.-A. Zhang, et al., A blockchain-based federated learning approach for electricity theft detection through dual-verification, IEEE Trans. Smart Grid (2026) 1–1. https://doi.org/10.1109/TSG.2026.3676843
[50] S. Ness, Hybrid KNN-LSTM framework for electricity theft detection in smart grids using SGCC smart-meter data, IEEE Access 13 (2025) 191809–191823. https://doi.org/10.1109/ACCESS.2025.3630123
[51] A.K. Mishra, B. Das, A novel density based clustering approach for electricity theft detection, IEEE Trans. Ind. Appl. 61 (4) (2025) 5537–5548. https://doi.org/10.1109/TIA.2025.3544167
[52] A.K. Mishra, B. Das, A novel anomaly detection approach for electricity theft detection, IEEE Trans. Ind. Appl. (2026) 1–14. https://doi.org/10.1109/TIA.2026.3677308
[53] W. Qin, Y. Ding, X. Luo, et al., A robust approach to electricity theft detection via tensor representation-driven contrastive distillation, IEEE Trans. Ind. Inform. 22 (5) (2026) 4561–4570. https://doi.org/10.1109/TII.2026.3659333
[54] A. Gao, F. Mei, J. Zheng, H. Sha, M. Guo, Y. Xie, et al., Electricity theft detection based on contrastive learning and non-intrusive load monitoring, IEEE Trans. Smart Grid 14 (6) (2023) 4565–4580. https://doi.org/10.1109/TSG.2023.3263219
[55] H.-X. Gao, S. Kuenzel, X.-Y. Zhang, et al., A hybrid convLSTM-based anomaly detection approach for combating energy theft, IEEE Trans. Instrum. Meas. 71 (2022) 1–10. https://doi.org/10.1109/TIM.2022.3201569
[56] A. Takiddin, M. Ismail, M. Nabil, M.M.E.A. Mahmoud, E. Serpedin, et al., Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings, IEEE Syst. J. 15 (3) (2021) 4189–4198. https://doi.org/10.1109/JSYST.2020.3030238
[57] A. Ghorbani, J.Y. Zou, Embedding for informative missingness: deep learning with incomplete data, in: 56th Annual Allerton Conference on Communication, Control, and Computing, 2018. https://doi.org/10.1109/ALLERTON.2018.8636008
[58] G. Antonesi, T. Cioara, V. Michalakopoulos, I. Papias, I. Anghel, E. Sarmas, Energy forecasting under missing data: comparative evaluation of augmented representations and decoder-only time-series imputation, Energy AI (2026) 100780. https://doi.org/10.1016/j.egyai.2026.100780
[59] R. Yao, N. Wang, W. Ke, Z. Liu, Z. Yan, X. Sheng, Electricity theft detection in incremental scenario: a novel semi-supervised approach based on hybrid replay strategy, IEEE Trans. Instrum. Meas. 72 (2023) 2530012. https://doi.org/10.1109/TIM.2023.3324674
[60] G.C. Chow, Tests of equality between sets of coefficients in two linear regressions, Econometrica 28 (3) (1960) 591–605. https://doi.org/10.2307/1910133
[61] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1) (1998) 47–78. https://doi.org/10.2307/2998540
[62] P. Massaferro, J.M.D. Martino, A. Fernández, et al., Fraud detection in electric power distribution: an approach that maximizes the economic return, IEEE Trans. Power Syst. 35 (1) (2020) 703–710. https://doi.org/10.1109/TPWRS.2019.2928276
[63] S. Maldonado, J. López, A. Iturriaga, Out-of-time cross-validation strategies for classification in the presence of dataset shift, Appl. Intell. 52 (5) (2022) 5770–5783. https://doi.org/10.1007/s10489-021-02735-2
[64] R. Punmiya, S. Choe, Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing, IEEE Trans. Smart Grid 10 (2) (2019) 2326–2329. https://doi.org/10.1109/TSG.2019.2892595
[65] C.G. Northcutt, L. Jiang, I.L. Chuang, Confident learning: estimating uncertainty in dataset labels, J. Artif. Intell. Res. 70 (2021) 1373–1411. https://doi.org/10.1613/jair.1.12125