African Journal of Education Science and Technology (AJEST) Vol. 9 No. 1 (October, 2026)
University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone

A Systematic Review of Machine Learning Approaches for Electricity Theft Detection under Class Imbalance: Implications for Utilities in Low-Resource Settings

Daniel Wanjala Machimbo[1], Shem Mbandu Angolo[1] & Andrew Kipkebut[2]
[1]*Department of Computer Science and Information Technology,* School of Computing and Mathematics
The Co-operative University of Kenya
[2]*Department of Computer Science and Information Technology,* School of Computing and Mathematics
Kabarak University

Corresponding E-mail: danielwanjalamachimbo@gmail.com

Abstract

Electricity theft accounts for much of the non-technical loss that utilities in low- and middle-income countries absorb, and machine learning on consumption data is the most studied way to detect it. Theft cases are rare in these data, so models must handle class imbalance, and how they do so affects how far reported results can be trusted. Following PRISMA 2020, we searched the Directory of Open Access Journals for peer-reviewed studies published from 2016 to 2026 that trained a classifier for electricity theft or non-technical loss and handled class imbalance explicitly. Of 64 records, 57 studies (58 reports) met the criteria. SMOTE and related oversampling methods were the most common remedy (22 studies), followed by hybrid resampling (10) and generative models (9). Thirty-seven studies (65%) used the public State Grid Corporation of China dataset, and one used data from an African utility. Of the 55 studies that resampled or generated data, 11 kept resampling to the training data, by explicit statement or by design, and four stated that balancing came before the train and test split; the other 40 did not make the order clear. Accuracy was reported by 77% of studies, MCC by 32% and PR-AUC by 21%. No study used a time-ordered split, three tuned the decision threshold, and two checked predictions against real cases. Two studies reported values that cannot all hold for the same test set. High reported accuracies are therefore weak evidence of field performance. Utilities such as the Kenya Power and Lighting Company should judge models on inspection outcomes, time-ordered validation and precision-recall measures before deployment.

Keywords: *electricity theft detection; non-technical losses; class imbalance; data leakage; systematic review*

Introduction

Non-technical losses (NTL) are electricity losses that arise from causes other than technical faults in the network. Most come from customers who tamper with or bypass their meters, connect to the supply illegally, resell electricity to others, or benefit from incorrect meter readings (Glauner et al., 2017; Messinis & Hatziargyriou, 2018). In Kenya, the Kenya Power and Lighting Company (KPLC) reported a system loss of 23.4% in the 2022/2023 financial year, about KSh 28.7 billion (Kenya Power and Lighting Company [KPLC], 2023). The figure combines technical and non-technical losses, but it shows the revenue at stake. A utility usually confirms theft by sending a team to inspect each suspected customer's premises, so it needs a reliable way to decide which customers to visit first.

Machine learning offers such a way. A classifier takes each customer's consumption history as input and predicts whether the customer is stealing. Smart metering and public datasets have made the approach common, above all the State Grid Corporation of China (SGCC) dataset released with the study of Zheng et al. (2018), which holds daily consumption for 42,372 customers, about 8.5% of them labelled as thieves. This imbalance is the central difficulty: a classifier that calls every SGCC customer honest is 91.5% accurate and finds no thieves. Studies therefore rebalance the data by oversampling, undersampling, or synthetic generation, weight the classes in the loss function, or use ensembles that rebalance internally (Chawla et al., 2002; He & Garcia, 2009).

These remedies carry risks of their own. When oversampling runs before the data are split into training and test sets, synthetic customers built from test customers enter the training data and synthetic customers can enter the test set, so the reported score partly measures how well the model recognises its own generated samples (Kaufman et al., 2012; Santos et al., 2018). Kapoor and Narayanan (2023) traced this kind of leakage across many fields that use machine learning. The choice of metric is a second risk. The majority class dominates accuracy, and ROC curves can look favourable while most flagged customers are honest, whereas recall, precision, the precision-recall curve, and the Matthews correlation coefficient (MCC) respond to errors on the rare class (Chicco & Jurman, 2020; Saito & Rehmsmeier, 2015).

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

Earlier reviews have mapped this field. Glauner et al. (2017) surveyed artificial intelligence methods for NTL detection and named class imbalance and the choice of evaluation metric among its open challenges. Messinis and Hatziargyriou (2018) grouped detection methods into data-oriented, network-oriented, and hybrid approaches, and Savian et al. (2021) reviewed the wider NTL literature systematically. These reviews describe which methods and data sources exist, and the two earliest appeared before any of the studies included here. They do not establish, study by study, how class imbalance is handled, whether resampling is kept out of the test data, which metrics are reported, or whether models are tested on later data or in the field. It therefore remains unresolved how far the high accuracies reported in this literature can be trusted, and what they mean for a utility with few labelled cases and a limited inspection budget.

This review addresses that gap for the open-access literature, with utilities in low-resource settings in mind. It asks four questions. RQ1: Which techniques do machine learning studies of electricity theft detection use to handle class imbalance? RQ2: How do these studies evaluate their models, and do they guard against leakage from resampling? RQ3: Which datasets and settings does the evidence come from? RQ4: What do the answers imply for utilities in low-resource settings such as Kenya?

Materials and Methods

Protocol and Reporting

The review follows the PRISMA 2020 statement (Page et al., 2021). The protocol was not registered in advance.

Eligibility Criteria

Studies were eligible if they were peer-reviewed journal or conference articles in English published from 2016 to 2026; trained a machine learning or deep learning classifier to detect electricity theft or non-technical losses from consumption or billing data; handled class imbalance with an explicit technique, at the data level (resampling or synthetic generation), at the algorithm level (class weights or cost-sensitive losses) or inside an ensemble; and reported a quantitative evaluation. We excluded reviews and surveys, theses, preprints and editorials, studies of hardware or sensor-only detection, and studies whose task was not theft or non-technical loss classification.

Information Source and Search

We searched the Directory of Open Access Journals (DOAJ) through its public article search interface on 29 October 2026. DOAJ lists only open-access journals that pass its checks on peer review, editorial policy, and publishing practice, and it covers the venues in which most of this literature appears: 36 of the 57 included studies were published in IEEE Access, the MDPI journals, or Scientific Reports. We chose an open-access source deliberately, because engineers at utilities in low-resource settings can read these studies without a subscription, so this is the evidence they can act on. The search combined three groups of terms in titles and abstracts and limited publication years to 2016 to 2026 (Box 1).

To verify the search, we saved the query and every record it returned, and we ran the same query again through the DOAJ interface on 31 October 2026. It returned the same 64 records. The search therefore covers the open-access literature indexed in DOAJ; studies published only in subscription journals are outside its scope, which we treat as a limitation.

Box 1. DOAJ search string (Elasticsearch syntax), run on 29 October 2026

(bibjson.title:(THEFT) OR bibjson.abstract:(THEFT)) AND bibjson.abstract:(IMBALANCE) AND bibjson.abstract:(LEARNING) AND bibjson.year:[2016 TO 2026]

THEFT = "electricity theft" OR "energy theft" OR "power theft" OR "non-technical loss" OR "non-technical losses"

IMBALANCE = imbalance OR imbalanced OR SMOTE OR oversampling OR undersampling OR "cost-sensitive" OR "minority class" OR skewed

LEARNING = "machine learning" OR "deep learning" OR "neural network" OR XGBoost OR "gradient boosting" OR "random forest" OR ensemble

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

Selection Process

Records were exported with their abstracts and deduplicated on normalised DOI and then on normalised title. The first author screened titles and abstracts and then full texts against the criteria, assisted by an AI tool that retrieved each article and extracted its details (see the Declarations). We recorded every exclusion and its reason in a screening log, from which we computed the counts in Figure 1. When two reports described the same study, we linked them and counted the study once.

Data Collection

For each study, we recorded the dataset and its country, the share of theft cases, the imbalance technique, the model, the validation design, the point at which resampling was applied relative to the train and test split, the metrics reported, the best test results of the proposed model, the choice of decision threshold, any explanation method, and any check against field inspections. Values were taken from the full text. Resampling was coded as confined to training data only when the article said so in words, and as applied before the split only when the article described that order; otherwise it was coded as unclear or not reported. A value that could not be read reliably from the full text was flagged and left out rather than estimated.

Appraisal of Studies

We did not give studies a quality score. Each study was appraised on items adapted from the analysis domain of PROBAST (Wolff et al., 2019) and from the literature on leakage: data provenance, the order of resampling and splitting, the use of metrics suited to imbalance, temporal or external validation, threshold selection, and evidence from field use. These items appear in Supplementary Table S1 and are summarized in Figure 2.

Synthesis

The studies differ in data, models and metrics, and many omit what a pooled estimate would need, so we did not attempt a meta-analysis. We grouped the imbalance techniques into eight families (Table 1) and synthesized the findings narratively by research question. Counts are numbers of studies, with percentages of the 57 included studies unless stated otherwise.

Results and Discussion

Study Selection

The search returned 64 records. After one duplicate was removed, 63 records were screened by title and abstract, and three were excluded: two were reviews, and one forecast electricity inconsistencies in mini-grids rather than classifying theft. All 60 remaining reports were retrieved and read in full. Two were excluded, one because its full text is in Chinese and one because it used no explicit technique for class imbalance. The review therefore includes 58 reports of 57 studies (Figure 1). Two reports by the same first author describe the same AlexNet model on SGCC with identical results (Khan et al., 2023, 2024), and we counted them as one study.

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

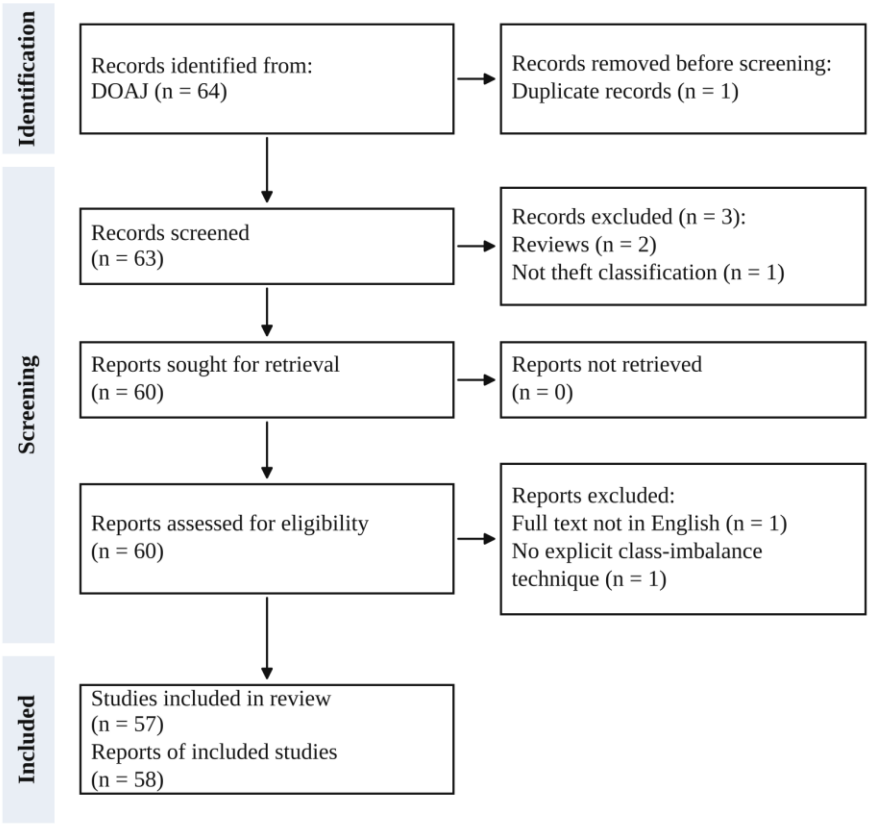


Figure 1. *PRISMA 2020 flow diagram of study selection*

Characteristics of the Included Studies

The included studies were published from 2019 to 2026, with the most in 2023 (14 studies). Supplementary Table S1 lists them. Twenty-six proposed deep learning models, 14 paired deep feature extractors with a boosting or tree classifier, 13 used tree ensembles, and four used classical single learners or broad comparisons of them. Thirty-seven studies (65%) used the SGCC dataset, 34 of them as their only data. Fourteen (25%) used utility data that had not been released publicly, from Brazil, China, Colombia, Ghana, Indonesia, Iran, Pakistan, Peru and Thailand. Two used the Irish CER smart meter data, and the remainder used other public or simulated data.

Imbalance-Handling Techniques (RQ1)

Table 1 groups the techniques into families. Synthetic oversampling with SMOTE or a variant such as ADASYN, Borderline-SMOTE, k-means SMOTE, KNNOR or LoRAS was the most common choice (22 studies). Ten studies combined oversampling with a cleaning or undersampling step, most often SMOTE with Tomek links or edited nearest neighbours. Nine generated synthetic theft profiles with generative models, among them Wasserstein GANs (Asif et al., 2022; Liu et al., 2020), TimeGAN (Y. Sun, Sun, et al., 2023), a VAE-GAN (Y. Sun, Lee, et al., 2023) and a conditional variational autoencoder (Gong et al., 2020). Four used random undersampling or oversampling on their own, four used ensembles that resample inside the learner, such as RUSBoost (Adil et al., 2020; Mujeeb et al., 2021), three created theft samples by simulating attacks on honest load profiles, and one relied on a cost-sensitive loss alone (Pamir et al., 2023). Six studies compared three or more techniques.

The comparative studies give the clearest view of what rebalancing adds. With Pakistani utility data, Ghori et al. (2021) found that untreated data already gave a recall of 0.99 for MLP, CatBoost, and XGBoost, and that logistic regression was the model most affected by sampling. Alfaverh et al. (2026), who applied SMOTE to training data only, saw accuracy rise from 94.76% to 95.03%. Two studies in which balancing preceded the split, or in which the

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

order was not stated, reported far larger gains. Kwarteng et al. (2023) report precision rising from 43.47% to 99.49% and F1 from 58.82% to 99.75% after SMOTE on Ghanaian data, but they applied SMOTE first and then divided the 375 resulting records into training, validation, and test sets, so synthetic customers were part of the test set. Qu et al. (2020) report accuracy rising from 85.53% to 94.53% with K-means SMOTE on data from 50 customers, without stating the order.

Table 1. Imbalance-handling families and evaluation practice

Family	Studies	Kept to training	Before split	Unclear or NR	None	PR-AUC or MCC	Accuracy
Synthetic oversampling (SMOTE and variants)	22	4	2	16	0	7	22
Hybrid resampling (oversampling with cleaning or undersampling)	10	2	0	8	0	6	8
Random under- or oversampling	4	0	1	3	0	3	3
Generative models (GAN, VAE)	9	0	0	8	1	7	4
Simulated theft attacks	3	0	0	3	0	0	2
Ensembles that resample inside the learner	4	4	0	0	0	0	1
Cost-sensitive learning	1	0	0	0	1	0	1
Comparison of three or more techniques	6	1	1	4	0	2	5

Note. Values are numbers of studies. Two studies used two families, so the family counts add up to 59 for 57 studies. Kept to training = resampling stated to be confined to training data, or done inside the learner. None = the study did not resample the data or add synthetic records to it. PR-AUC or MCC and Accuracy count studies reporting those metrics.

Evaluation Practice and Leakage Risk (RQ2)

Figure 2 summarises how the studies evaluated their models. Twenty-nine studies (51%) did not describe a hold-out split or cross-validation clearly enough to show how the test data were formed. Of the 55 studies that resampled or generated data, 11 (20%) kept resampling away from the test data. Seven said so in words; Abro et al. (2025), for example, write that "all oversampling methods, including Random Oversampling and SMOTE, were exclusively applied to the training set after dividing the data into training and testing partitions". The other four used ensembles that resample inside the learner, which keeps the test data out by design. Four studies (7%) stated that balancing came before the split (Ali et al., 2025; Kwarteng et al., 2023; Orozco-Cerón et al., 2026; Sánchez-Mojica & Suarez, 2026), and the remaining 40 (73%) did not report the order or described it ambiguously. In the study by Orozco-Cerón et al. (2026), the model was trained without synthetic data, but undersampling before the split left a balanced test set, so its precision of 0.695 is higher than inspectors would see at a lower field prevalence.

Accuracy was the most common headline metric (44 studies, 77%), and 50 studies (88%) reported recall or an equivalent such as the detection rate. Metrics suited to rare classes were less common: 18 studies (32%) reported MCC, 12 (21%) reported PR-AUC, and six (11%) reported both. Twenty-six studies (46%) reported accuracy with neither PR-AUC nor MCC. Some reported values are hard to reconcile. Adil et al. (2020) give an F1 score of 96.1% with a precision of 88.9% and a recall of 91.09%, although F1, the harmonic mean of the two, cannot exceed both. Pamir et al. (2023) report precision of 99.45% and F1 of 95.93% with a ROC-AUC of 71.13%, and the reported values do not show which class the precision and F1 describe.

No study used a time-ordered split, in which a model trained on earlier months is tested on later ones, although that is how a utility would use it. The only study that tested a model under conditions different from those of training used simulated data, and its accuracy fell from 99% to 94.9% when the theft power and location changed (Narayanan & Hardy, 2023).

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

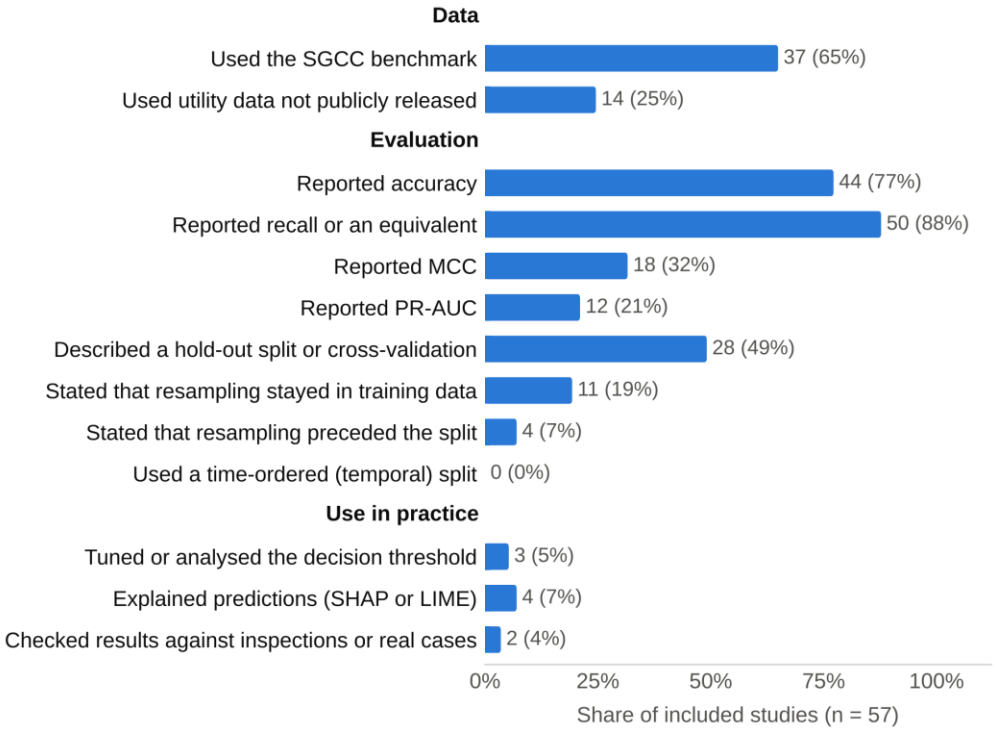


Figure 2. Data, evaluation and use-in-practice items reported by the included studies (n = 57). The two resampling items count studies that stated the order; 40 of the 55 studies that resampled or generated data did not make the order clear

Datasets and Settings (RQ3)

SGCC dominates the evidence. Of the 37 studies that used it, 12 reported a test accuracy of 95% or more, against the 91.5% that a model flagging no one would score. Real utility data came from 14 studies. They used meter readings from the Provincial Electricity Authority of Thailand (Janthong et al., 2023; Rajitroj et al., 2025), monthly billing records (Coelho et al., 2023; Orozco-Cerón et al., 2026), and readings taken every 40 to 60 days from traditional meters in Iran (Norouzi et al., 2025). In two studies, the labels came from inspections or from customers apprehended for theft (Kwarteng et al., 2023; Orozco-Cerón et al., 2026), the kind of record that any utility running inspections already keeps. The only African dataset came from the Electricity Company of Ghana and covered 294 customers, 42 of them apprehended thieves (Kwarteng et al., 2023). No study used Kenyan data.

Use in Practice

Few studies went beyond test-set scores. Three tuned or analysed the decision threshold (Coelho et al., 2023; Monteiro et al., 2025; Orozco-Cerón et al., 2026); the others used the default cut-off or did not say. Four explained their predictions with SHAP or LIME (Ali et al., 2024; Ammar et al., 2025; Monteiro et al., 2025; Rahaman & Mohamad Idris, 2026). On Peruvian utility data, for example, Monteiro et al. (2025) found extreme month-to-month variation, long periods of low consumption, and tariff category to be the strongest predictors of fraud. Two studies tested predictions against real outcomes. Coelho et al. (2023) report that field inspections guided by their model in Brazil had an accuracy of 76%, and Lin, Feng, Feng, et al. (2021) report that their model identified 44 of 47 real abnormal customers held out from training. Rahaman and Mohamad Idris (2026) built a lightweight version of their ensemble that scored a record in 3.9 ms, about 40% faster than the full model, with ROC-AUC falling only from 0.9869 to 0.9858.

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

What the Evidence Supports

The studies explore many ways of rebalancing data, but their reporting seldom lets a reader rule out leakage from resampling. Only one study in five made clear that its test data were untouched by resampling, and four said plainly that balancing came first. When synthetic minority samples reach the test set, a model can score well by recognizing interpolations of the samples it was trained on, the mechanism that Santos et al. (2018) describe for cross-validation with oversampling. The modest SMOTE gain reported by Alfaverh et al. (2026), with resampling kept to the training data, and the near-perfect scores reported by Kwarteng et al. (2023), where it was not, fit this mechanism, although two studies cannot establish it.

The choice of metrics makes the problem worse. On SGCC, a classifier that flags no one is 91.5% accurate, so accuracies in the low to mid 90s say little on their own, yet accuracy was the most reported metric and almost half the studies gave neither PR-AUC nor MCC. Saito and Rehmsmeier (2015) showed that ROC curves can look favourable on imbalanced data while precision is poor, and Chicco and Jurman (2020) recommend MCC because it uses all four cells of the confusion matrix. For utility, the more useful figure is the precision among the customers it can afford to inspect.

The evidence also rests mostly on one benchmark. SGCC is useful because it is public and lets studies be compared, but its customers, labels, and theft share are fixed, and 37 of the 57 studies used it. Small gains on such a benchmark may not carry over to a utility whose customers, tariffs, and theft methods differ, and no study tested a model on real data from a later period or another utility.

Implications for Utilities in Low-Resource Settings (RQ4)

The findings point to several steps for a utility such as KPLC. Models do not appear to require smart meters. Studies using monthly billing data (Coelho et al., 2023; Orozco-Cerón et al., 2026) and readings taken every 40 to 60 days (Norouzi et al., 2025) produced usable classifiers, and the Brazilian study converted its model into inspections with 76% accuracy. A utility can start with the billing histories and inspection results it already holds.

Inspection outcomes, including inspections that found nothing, are the labels a utility can trust. They should be stored so that each result is linked to the customer's consumption history and to the date of the visit, which is what makes time-ordered evaluation possible.

Evaluation should follow the way the model will be used. That means training on earlier months and testing on later ones, confining any resampling or class weighting to the training folds, and setting the threshold to match the number of inspections that field crews can carry out. PR-AUC, MCC, and recall and precision at that threshold describe performance far better than accuracy does.

Simple models are competitive. Random forests and extra trees were among the best performers on utility data and on public benchmarks (Abro et al., 2025; Alfaverh et al., 2026; Orozco-Cerón et al., 2026), and on 53 monthly values per customer, deep sequence models did worse than random forests (Orozco-Cerón et al., 2026). Tree ensembles run on ordinary hardware, and with SHAP, they give an inspector a reason for each flag (Monteiro et al., 2025).

Before a full rollout, a pilot in one region that compares model-guided and routine inspections over the same period would provide the field evidence that almost all of the reviewed studies lack.

Strengths and Limitations

The review coded each included study from its full text and recorded the order of resampling only from explicit statements, which keeps its estimate of leakage conservative. It has several limitations. It searched one database, DOAJ, so studies published only in subscription journals, including many IEEE Transactions articles and conference papers, are missing, and the findings describe the open-access literature. One reviewer screened the records and extracted the data with the help of an AI tool, without independent double screening, so some misclassification is possible, although the screening log and extraction form allow every decision to be checked. Only studies in English were included. Studies were appraised on selected items without a formal risk-of-bias score, and differences in data and metrics ruled out a meta-analysis. Studies with weak results may also be less likely to be published, which would raise the scores reported here.

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

Conclusion and Recommendations

Machine learning studies of electricity theft detection handle class imbalance in many ways, most often with SMOTE and its variants. Few of them show that their test data were kept free of synthetic samples or report metrics suited to rare classes. None tested a model on later data, and only two checked predictions against real outcomes. Accuracies above 95% on the SGCC benchmark should therefore be read with caution. For utilities in low-resource settings, KPLC among them, the more reliable path is to build labelled data from inspection outcomes, train simple and explainable models, evaluate them on later months with precision-recall measures at a realistic inspection budget, and confirm their value in a field pilot.

Future studies would be easier to judge and to reuse if they reported a short list of items: the data source and theft share; how the data were split, and whether by time; the stage at which any resampling or synthetic generation was applied; PR-AUC, MCC, and recall and precision at the chosen threshold, with the rule used to set it; and code, or enough detail to repeat the work. Two gaps also need direct work. Utility data from Africa are almost absent from the open literature, so partnerships that release anonymised consumption and inspection records, or at least report results on them, would help most. Temporal and field validation are missing too, and a study comparing model-guided inspections with routine inspections would be worth more to utilities than a further gain on SGCC.

Declarations

Use of artificial intelligence. An AI tool was used to collect the search records, retrieve the open-access full texts, and extract study details from them into the data extraction sheet. The authors checked the extracted data against the source articles and took full responsibility for the content.

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

Funding. This research received no specific grant from any funding agency.

Data availability. The search string, screening log, and extraction form are available from the corresponding author on request.

References

Abro, S. A., Laghari, J. A., Memon, S. A., Khan, T. A., Memon, I., Nasir, H., & Fatima, K. (2025). Non-technical loss detection in power distribution networks using machine learning. *Scientific Reports*, *15*(1), 36189. https://doi.org/10.1038/s41598-025-20048-z

Adil, M., Javaid, N., Qasim, U., Ullah, I., Shafiq, M., & Choi, J.-G. (2020). LSTM and bat-based RUSBoost approach for electricity theft detection. *Applied Sciences*, *10*(12), 4378. https://doi.org/10.3390/app10124378

Alfaverh, F., Gan, H., Miroshnyk, V., Bin Saeed, Z., Blinov, I., Shymaniuk, P., Tarassodi, P., & Mporas, I. (2026). Electricity theft detection from electricity and gas measurements using machine learning. *Energies*, *19*(9), 2045. https://doi.org/10.3390/en19092045

Ali, A., Khan, L., Javaid, N., Aslam, M., Aldegheishem, A., & Alrajeh, N. (2024). Exploiting machine learning to tackle peculiar consumption of electricity in power grids: A step towards building green smart cities. *IET Generation, Transmission & Distribution*, *18*(3), 413–445. https://doi.org/10.1049/gtd2.13056

Ali, A., Khan, L., Javaid, N., Bouk, S. H., Aldegheishem, A., & Alrajeh, N. (2025). Mitigating anomalous electricity consumption in smart cities using an AI-based stacked-generalization technique. *IET Renewable Power Generation*, *19*(1), e12785. https://doi.org/10.1049/rpg2.12785

Ammar, M., Javaid, N., & Arishi, A. (2025). An AI-explained systematic modular approach for enhanced electricity theft detection for urbanized smart grid environment. *Alexandria Engineering Journal*, *129*, 998–1023. https://doi.org/10.1016/j.aej.2025.07.040

Asif, M., Nazeer, O., Javaid, N., Alkhammash, E. H., & Hadjouni, M. (2022). Data augmentation using BiWGAN, feature extraction and classification by hybrid 2DCNN and BiLSTM to detect non-technical losses in smart grids. *IEEE Access*, *10*, 27467–27483. https://doi.org/10.1109/ACCESS.2022.3150047

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Coelho, F. J. S., Feitosa, A. R. S., Alcântara, A. L. M., Li, K., Lima, R. F., Silva, V. R., & da Silva-Filho, A. G. (2023). HyMOTree: Automatic hyperparameter tuning for non-technical loss detection based on multi-objective and tree-based algorithms. *Energies*, *16*(13), 4971. https://doi.org/10.3390/en16134971

Ghori, K. M., Awais, M., Khattak, A. S., Imran, M., Fazal-E-Amin, & Szathmary, L. (2021). Treating class imbalance in non-technical loss detection: An exploratory analysis of a real dataset. *IEEE Access*, *9*, 98928–98938. https://doi.org/10.1109/ACCESS.2021.3095145

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

Glauner, P., Meira, J. A., Valtchev, P., State, R., & Bettinger, F. (2017). The challenge of non-technical loss detection using artificial intelligence: A survey. *International Journal of Computational Intelligence Systems*, *10*(1), 760–775. https://doi.org/10.2991/ijcis.2017.10.1.51

Gong, X., Tang, B., Zhu, R., Liao, W., & Song, L. (2020). Data augmentation for electricity theft detection using conditional variational auto-encoder. *Energies*, *13*(17), 4291. https://doi.org/10.3390/en13174291

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Janthong, S., Chalermyanont, K., & Duangsoithong, R. (2023). Unbalanced data handling techniques for classifying energy theft and defective meters in the Provincial Electricity Authority of Thailand. *IEEE Access*, *11*, 46522–46540. https://doi.org/10.1109/ACCESS.2023.3274543

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 1–21. https://doi.org/10.1145/2382577.2382579

Kenya Power and Lighting Company. (2023). *Annual report and financial statements 2022/2023*.

Khan, N., Amir Raza, M., Ara, D., Mirsaeidi, S., Ali, A., Abbas, G., Shahid, M., Touti, E., Yousef, A., & Bouzguenda, M. (2023). A deep learning technique, AlexNet, to detect electricity theft in smart grids. *Frontiers in Energy Research*, *11*, 1287413. https://doi.org/10.3389/fenrg.2023.1287413

Khan, N., Shahid, Z., Alam, M. M., Sajak, A. A. B., Nazar, M., & Mazliham, M. S. (2024). A novel deep learning technique to detect electricity theft in smart grids using AlexNet. *IET Renewable Power Generation*, *18*(6), 941–958. https://doi.org/10.1049/rpg2.12846

Kwarteng, M. Y., Effah, F. B., Kwegyir, D., & Frimpong, E. A. (2023). ANN-based electricity theft classification technique for limited data distribution systems. *Jurnal Nasional Teknik Elektro*, *12*(1), 8–16. https://jnte.ft.unand.ac.id/index.php/jnte/article/view/1072

Lin, G., Feng, H., Feng, X., Wen, H., Li, Y., Hong, S., & Ni, Z. (2021). Electricity theft detection in power consumption data based on adaptive tuning recurrent neural network. *Frontiers in Energy Research*, *9*, 773805. https://doi.org/10.3389/fenrg.2021.773805

Liu, H., Li, Z., & Li, Y. (2020). Noise reduction power stealing detection model based on self-balanced data set. *Energies*, *13*(7), 1763. https://doi.org/10.3390/en13071763

Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. *Electric Power Systems Research*, *158*, 250–266. https://doi.org/10.1016/j.epsr.2018.01.005

Monteiro, T., Castor, G., Castillo Correa, C., Arias, H., Ñaupari Huatuco, D., & Molina Rodríguez, Y. (2025). A hybrid machine learning framework for electricity fraud detection: Integrating isolation forest and XGBoost for real-world utility data. *Energies*, *18*(23), 6249. https://doi.org/10.3390/en18236249

Mujeeb, S., Javaid, N., Ahmed, A., Gulfam, S. M., Qasim, U., Shafiq, M., & Choi, J.-G. (2021). Electricity theft detection with automatic labeling and enhanced RUSBoost classification using differential evolution and Jaya algorithm. *IEEE Access*, *9*, 128521–128539. https://doi.org/10.1109/ACCESS.2021.3102643

Narayanan, A., & Hardy, T. (2023). Synthetic data generation for machine learning model training for energy theft scenarios using co-simulation. *IET Generation, Transmission & Distribution*, *17*(5), 1035–1046. https://doi.org/10.1049/gtd2.12619

Norouzi, A., Ahmadi, I., Nazari, M., Pourrostami, H., & Hoseini-Kordkheili, H. (2025). A novel framework to detect electricity theft in regions with traditional metering: applying a hybrid machine learning-based approach to low-sampling-rate data. *Results in Engineering*, *25*, 104478. https://doi.org/10.1016/j.rineng.2025.104478

Orozco-Cerón, O. W., Joaqui-Barandica, O., & Manotas-Duque, D. F. (2026). Comparing machine and deep learning for electricity theft detection from monthly billing data in an emerging energy market. *Technologies*, *14*(9), 568. https://doi.org/10.3390/technologies14090568

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pamir, Javaid, N., Javed, M. U., Houran, M. A., Almasoud, A. M., & Imran, M. (2023). Electricity theft detection for energy optimization using deep learning models. *Energy Science & Engineering*, *11*(10), 3575–3596. https://doi.org/10.1002/ese3.1541

Qu, Z., Li, H., Wang, Y., Zhang, J., Abu-Siada, A., & Yao, Y. (2020). Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. *Energies*, *13*(8), 2039. https://doi.org/10.3390/en13082039

Rahaman, M. A., & Mohamad Idris, R. (2026). A stacking ensemble with Pareto optimization for scalable electricity theft detection via hybrid data repair and lightweight deployment. *Scientific Reports*, *16*(1), 14548. https://doi.org/10.1038/s41598-026-39693-z

Rajitroj, N., Seepromting, K., & Chatthaworn, R. (2025). 1D-CNN and a weight-based balancing technique for improved detection of non-technical losses in power distribution systems. *E3S Web of Conferences*, *629*, 06005. https://doi.org/10.1051/e3sconf/202562906005

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sánchez-Mojica, K. Y., & Suárez, O. J. (2026). Artificial intelligence model and balancing techniques for controlling non-technical losses in the electrical energy sector. *AiBi Revista de Investigación, Administración e Ingeniería*, *14*(1), 1–7. https://doi.org/10.15649/2346030X.5635

Santos, M. S., Soares, J. P., Abreu, P. H., Araújo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine*, *13*(4), 59–76. https://doi.org/10.1109/MCI.2018.2866730

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal

Savian, F. de S., Siluk, J. C. M., Garlet, T. B., do Nascimento, F. M., Pinheiro, J. R., & Vale, Z. (2021). Non-technical losses: A systematic contemporary article review. *Renewable and Sustainable Energy Reviews*, *147*, 111205. https://doi.org/10.1016/j.rser.2021.111205

Sun, Y., Lee, J., Kim, S., Seon, J., Lee, S., Kyeong, C., & Kim, J. (2023). Energy theft detection model based on VAE-GAN for imbalanced dataset. *Energies*, *16*(3), 1109. https://doi.org/10.3390/en16031109

Sun, Y., Sun, X., Hu, T., & Zhu, L. (2023). Smart grid theft detection based on a hybrid multi-time scale neural network. *Applied Sciences*, *13*(9), 5710. https://doi.org/10.3390/app13095710

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376

Zheng, Z., Yang, Y., Niu, X., Dai, H.-N., & Zhou, Y. (2018). Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, *14*(4), 1606–1615. https://doi.org/10.1109/TII.2017.2785963

ISSN: 2309:9240, All Rights Reserved for all authors in this Journal