A Hybrid Stacking Ensemble Model with Multidimensional Features for Electricity Theft Detection: Field Validation in West Java

Qashtalani Haramaini
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
qashtalani.haramaini@ui.ac.id

Ismi Rosyiana Fitri
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
ismi.rosyiana01@ui.ac.id

Fauzan Hanif Jufri
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
fauzanhj@ui.ac.id

Iwa Garniwa
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
iwa@eng.ui.ac.id

Hazlie Mokhlis
Department of Electrical Engineering, Universiti Malaya, Kuala Lumpur, Malaysia
hazli@um.edu.my

Budi Sudiarto
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
budi.sudiarto@ui.ac.id (corresponding author)

Received: 28 March 2026 | Revised: 27 April 2026 | Accepted: 11 May 2026

 DOI: https://doi.org/10.48084/etasr.18978

ABSTRACT

Non-Technical Losses (NTL) in electricity distribution—arising from consumer-level fraud such as meter tampering, illegal connections, and billing manipulation—constitute a persistent financial burden for utilities operating without Advanced Metering Infrastructure (AMI). Existing detection approaches rely predominantly on rule-based threshold systems that fail under high class imbalance, leaving the vast majority of fraudulent accounts undetected in manual-reading networks. This study formulates NTL detection as a binary supervised classification problem at the individual consumer level, using monthly manual-reading data representative of most Indonesian distribution networks. A hybrid stacking ensemble framework integrates 14 features across four domains: consumer behavior, technical infrastructure, socio-economic indicators, and spatial characteristics. LightGBM and CatBoost serve as base learners, and Logistic Regression acts as the meta-learner. Given a severe class imbalance ratio of 22.8:1, Precision-Recall AUC (PR-AUC) is adopted as the primary evaluation metric. The framework was developed and validated on a large-scale dataset from PT PLN (Persero) West Java that included 2.36 million consumers over a 30-month period. Under 5-fold stratified cross-validation, the ensemble achieved a test PR-AUC of 0.7764 (CV: 0.793±0.011), outperforming all single-model baselines. Field validation across 12,407 consumer inspections confirmed a detection rate of 72.0%, compared to 15.3% under the incumbent rule-based system—a 4.7-fold improvement. To the best of our knowledge, this is the largest field-validated machine learning study for electricity theft detection reported for a non-AMI distribution network.

Keywords-electricity theft detection; machine learning; non-AMI networks; stacking ensemble; imbalanced classification; precision-recall AUC; West Java; non-technical losses

I. INTRODUCTION

Electricity distribution losses represent a major financial and operational burden for power utilities in developing economies. In 2023, Indonesia reported electricity losses of 8.4%, substantially exceeding those of Singapore (2.02%), Malaysia (5.79%), and Thailand (6.11%) [1]. These losses are classified as Technical Losses (TLs), arising from energy dissipation in conductors and transformers, and Non-Technical Losses (NTLs), which encompass electricity theft, meter tampering, billing fraud, and measurement errors [2]. Electricity theft is the most economically significant component of NTLs, contributing to elevated production costs, increased subsidies, and weakened grid reliability [3, 4]. The diversity of theft techniques, ranging from meter bypassing to billing manipulation, further complicates detection efforts [5], particularly in urban settings where fraudulent consumption patterns overlap with legitimate anomalies [6].

Utilities and regulators have developed a broad portfolio of interventions against NTL, spanning regulatory, socio-technical, hardware, and data-driven domains. On the regulatory front, governments in developing economies have implemented legal penalties for tampering, mandatory self-reporting obligations, and subsidy rationalization policies designed to reduce theft incentives at the consumer level. In Indonesia, PLN's Penertiban Pemakaian Tenaga Listrik (P2TL) enforcement program provides the legal and operational framework for identifying, sanctioning, and recovering losses from fraudulent consumers [7]. Socio-technical approaches—including community-based reporting schemes, public awareness campaigns, and norm-change interventions targeting theft culture—have demonstrated measurable reductions in theft incidence in contexts where social trust in utilities is sufficiently high. Hardware countermeasures, including anti-tampering meter enclosures, current-based detection circuits, and the progressive rollout of Advanced Metering Infrastructure (AMI), address the problem at the physical and measurement layer [7, 8]. AMI in particular enables high-frequency consumption monitoring that has substantially advanced automated theft detection in well-resourced networks [9-11].

Despite this breadth, each intervention faces structural limitations in the Indonesian distribution context. Regulatory enforcement through P2TL is actionable only after fraudulent accounts have been reliably identified, as it presupposes a detection mechanism that does not yet exist on an operational scale [7]. Socio-technical approaches operate on timescales and social preconditions incompatible with the monthly billing cycle and the anonymity of dense urban low-voltage networks. Hardware countermeasures require sustained capital expenditure: full AMI deployment across PLN West Java presents a multi-year infrastructure commitment, and AMI implementation across Indonesia remains structurally incomplete [1]. Conventional field-inspection programs guided by rule-based heuristics applied to monthly billing records remain the dominant operational practice in utilities without AMI. However, PLN West Java has historically confirmed a theft-detection rate of approximately 15% under this regime, demonstrating that rule-based prioritization leaves the vast majority of fraudulent accounts undetected, and enforcement-only strategies cannot scale to millions of consumers cost-effectively. In this context, data-driven machine learning applied to existing monthly billing data is positioned not as a substitute for regulatory, socio-technical, or hardware interventions, but as a low-capex operational layer that can materially increase inspection yield in the period before AMI deployment becomes universal.

Three research streams in the machine learning literature address electricity theft detection [2, 12]. The first and most prominent is data-driven supervised learning, which gained traction with the proliferation of smart meter data. Gradient boosting-based approaches [9], support vector machines [13, 14], and deep learning architectures applied to monthly consumption time-series [15] have demonstrated strong detection performance in resource-rich settings. The second stream encompasses network-topology and power-flow analysis methods, which infer anomalies from aggregate feeder-level energy imbalances rather than individual consumption profiles. The third involves hardware-assisted approaches using physical anti-tampering devices and real-time monitoring through smart metering systems [7, 8]. Although collectively these streams have advanced the field substantially, they share a common dependency: the overwhelming majority of published methods require AMI data—high-frequency, automatically collected consumption readings—that is simply unavailable in most distribution networks across developing economies, including the majority of PLN's operational territory in Indonesia.

Peer-reviewed studies specifically targeting non-AMI, monthly manual-reading networks remain scarce relative to the expansive AMI literature. Within the Indonesian operational context, in [16], a SMOTE+KNN+LR pipeline was applied to PLN customer data, reporting 98.7% accuracy—a metric well-established in the imbalanced-classification literature as misleading under severe class imbalance [17], where the majority class dominates the metric regardless of minority-class detection performance. Institutional benchmarks conducted at Universitas Indonesia using PLN West Java data have established preliminary results: XGBoost achieved 80% overall accuracy [18] and a Logistic Regression-based pipeline reported 59% F1 [19]. Beyond Indonesia, recent studies on theft detection under manual-reading or low-AMI conditions have applied deep learning on monthly consumption data [15], ensemble machine-learning with adaptive synthetic oversampling [20], and predictive modelling with XGBoost for residential NTL detection [21]; yet these studies remain constrained by limited dataset scale, absence of field validation, and reliance on consumption-only features [13].

Together, these studies reveal a consistent and structurally determined pattern in non-AMI NTL detection research. Three limitations recur systematically across the literature, and their persistence reflects not oversight but the constraints inherent to institutional research conducted within a single utility's

operational boundaries. First, feature engineering has remained confined to billing anomalies and consumption ratios derived from meter-reading records, leaving technical infrastructure variables (transformer loading, feeder characteristics, network topology), socio-economic indicators (tariff category, consumer classification, neighborhood-level income proxies), and spatial clustering effects systematically unexploited—despite evidence from adjacent domains that these dimensions carry significant discriminative signal [22]. Second, the use of overall accuracy as the primary or sole evaluation metric—despite severe and well-documented class imbalance ratios consistently exceeding 20:1 in electricity theft datasets—has produced optimistic performance estimates that systematically overstate the operational utility of proposed models, a problem extensively documented across imbalanced classification domains [17]. Third, and most consequentially for operational adoption, none of the identified prior works in non-AMI settings have subjected their models to real-world field validation at scale; all reported performance figures reflect held-out test sets drawn from the same statistical distribution as training data, a condition that does not replicate the adversarial dynamics, concept drift, and operational heterogeneity of live deployment. These three convergent limitations—feature narrowness, metric misalignment, and absence of field validation—collectively define the design space and motivation for this study.

The research gap this study addresses can be stated as: no published study on NTL detection in non-AMI, monthly manual-reading distribution networks has simultaneously integrated multidimensional features spanning technical, socio-economic, and spatial domains beyond consumption profiles; adopted minority-class-aware evaluation as the primary metric; employed a stacking ensemble architecture with gradient-boosting base learners; and validated results through large-scale independent field inspection. Each of these elements exists in isolation within adjacent literature—multidimensional feature engineering in AMI-based detection [9, 22], PR-AUC as the primary metric in imbalanced learning [17], stacking ensembles in energy forecasting [23], and field-based outcome validation in operational inspection programs—but their integration within a non-AMI operational context at the scale of millions of consumers remains unreported. This gap is operationally consequential: without field-validated, minority-class-aware models built on multidimensional, operationally grounded features, utilities operating manual-reading networks lack the empirical foundation needed to make evidence-based decisions about transitioning from rule-based to ML-guided inspection prioritization.

This work addresses the identified gap through four specific contributions. First, a multidimensional feature-engineering framework comprising 14 variables spanning consumer behavior, technical infrastructure, socio-economic indicators, and spatial characteristics—designed specifically for non-AMI monthly billing and token-purchase data covering both postpaid and prepaid consumers. Second, a uniform 8-classifier PR-AUC benchmark establishing comparative baselines under 5-fold stratified cross-validation, with accuracy, F1-score, recall, and precision reported as supporting metrics to enable direct comparison with prior work. Third, a two-level stacking ensemble with LightGBM and CatBoost as complementary gradient-boosting base learners and Logistic Regression as meta-learner, achieving a test PR-AUC of 0.7764 and cross-validation PR-AUC of 0.793±0.011, outperforming all single-model baselines and establishing a new benchmark for this operational class. Fourth, large-scale field validation across 12,407 consumer inspections conducted independently by PLN field officers, confirmed a 72.0% theft-detection rate compared to 15.3% under the incumbent rule-based system—a 4.7-fold improvement that, to the best of our knowledge, represents the largest field-validated result reported for a non-AMI distribution network. The PLN West Java dataset serves as a supporting empirical resource rather than a standalone contribution.

II. DATASET AND MULTIDIMENSIONAL FEATURES

A. Data Source and Study Area

The dataset comprises monthly electricity consumption records from PT PLN (Persero) West Java Distribution Region, covering January 2022 to July 2024 across 93 administrative areas. After data cleaning (3.2% missing, 0.8% inconsistent, and 0.1% duplicates removed), the final dataset comprised 2,363,717 unique consumers. Ground-truth labels were derived from official PLN field inspections, yielding a class imbalance ratio of approximately 22.8:1. Figure 1 shows the spatial distribution of consumers and confirmed theft cases.

B. Multidimensional Feature Set

Fourteen features were engineered across four domains, as summarized in Table I: Consumer Profile (3 features: consumption pattern, tariff class, contracted capacity), Technical Profile (4 features: transformer load profile, phase load imbalance, feeder-level losses, transformer capacity), Social Profile (4 features: population density, bill payment regularity, GRDP, location vulnerability index), and Spatial Profile (3 features: terrain elevation, administrative unit code, feeder geographic identifier). Categorical variables were target-encoded; numerical features were standardized. Class imbalance was addressed through class-weight parameterization without oversampling, preserving the natural class distribution in the training data.

TABLE I. MULTIDIMENSIONAL FEATURE SET FOR ELECTRICITY THEFT DETECTION

Domain	Variable	Type	Coverage
Consumer	Consumption pattern (30 mo.)	Num./Cat.*	2,363,717
	Tariff class	Cat.	2,363,717
	Contracted capacity (VA)	Cat.	2,363,717
Technical	Transformer load profile	Cat.	66,777
	Phase load imbalance	Num.	66,777
	Feeder-level losses (%)	Num.	1,645
	Transformer capacity (kVA)	Num.	66,777
Social	Population density (per km²)	Num.	93 areas
	Bill payment regularity	Num.	2,363,717
	GRDP (IDR trillion)	Num.	93 areas
	Location vulnerability index	Num.	93 areas
Spatial	Terrain elevation (m)	Num.	93 areas
	Administrative unit code	Cat.	93 areas
	Feeder geographic identifier	Cat.	1,645 feeders

Type: Num. = Numerical; Cat. = Categorical

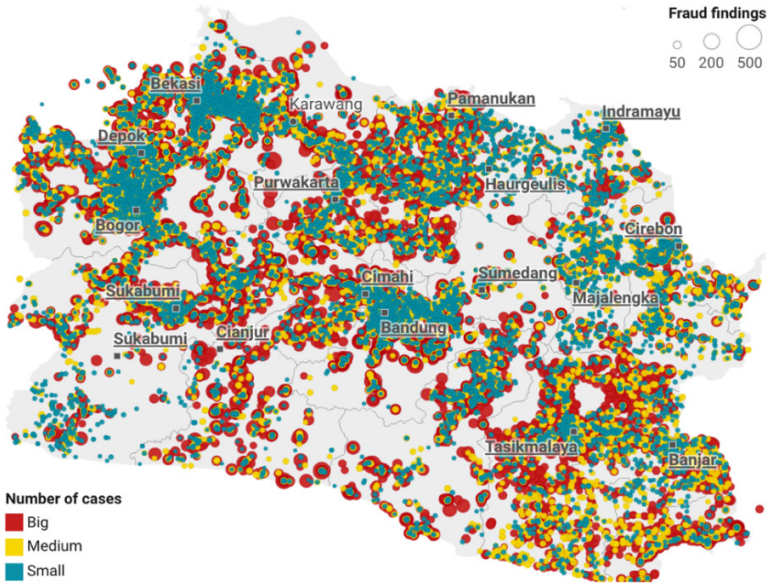

Fig. 1. Spatial distribution of consumers and confirmed theft cases across West Java.

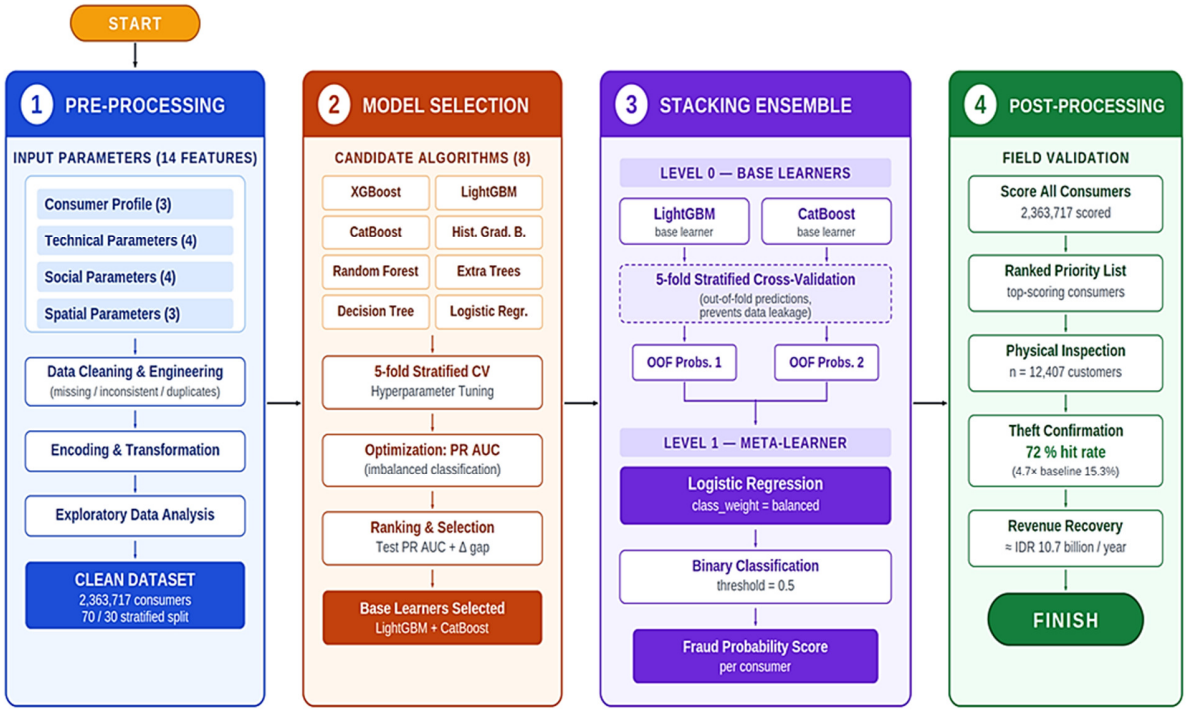

Fig. 2. Overall research methodology workflow.

III. HYBRID STACKING ENSEMBLE MODEL

Figure 2 illustrates the overall research workflow, which comprises four sequential stages: preprocessing, model selection and training, stacking ensemble construction, and post-processing. Each stage was designed to reflect the operational constraints of PLN's non-AMI distribution environment, specifically, the reliance on monthly manual-reading and prepaid token-purchase records, the absence of high-frequency AMI data, and the severe 22.8:1 class imbalance inherent to electricity theft datasets at scale.

A. Preprocessing

Raw consumption and billing records from PT PLN (Persero) West Java were consolidated into a 14-feature multidimensional dataset spanning four analytical dimensions: consumer profile (3 features), technical parameters (4 features), spatial parameters (3 features), and social parameters (4 features). This multidimensional feature design was deliberate; prior non-AMI studies have been constrained to consumption-profile features alone, neglecting the technical infrastructure, socio-economic, and spatial dimensions that carry independent

discriminative signals in utility fraud detection [22]. The dataset covers both postpay consumers monitored through monthly manual meter readings and prepay consumers monitored through token-purchase transaction patterns, reflecting the mixed-metering reality of PLN's West Java service territory.

Data engineering proceeded in three sequential steps. First, missing value imputation was applied to records with incomplete technical or spatial attributes, using median substitution for continuous variables and mode substitution for categorical fields, with imputation rates documented per feature to enable reproducibility. Second, inconsistency resolution addressed records with physically implausible consumption values—including negative readings and consumption figures exceeding feeder-level capacity constraints—which were corrected against network topology records where possible and removed otherwise. Third, duplicate consumer records arising from billing system merges were identified by consumer ID and meter serial number and deduplicated, retaining the most recent record per consumer. After these cleaning steps, the dataset underwent categorical encoding, feature transformation including log-normalization of skewed consumption distributions, and Exploratory Data Analysis (EDA) to characterize the feature distributions, inter-feature correlations, and class-conditional separation across the four domains.

The cleaned, analysis-ready dataset comprised 2,363,717 consumer records. To ensure that evaluation results reflect realistic deployment conditions rather than artificially balanced laboratory conditions, the dataset was partitioned into training (70%) and test (30%) sets using stratified random sampling. Stratification was applied jointly on the class label and consumer type (postpaid/prepaid) to preserve both the 22.8:1 overall class imbalance ratio and the metering-type distribution across both splits. No resampling, oversampling, or synthetic augmentation was applied at this stage; the imbalance was instead managed through model-level class weighting, as described in Section B.

B. Model Selection and Training

Eight heterogeneous classifiers were evaluated as candidate base learners: XGBoost, LightGBM, CatBoost, Histogram Gradient Boosting, Random Forest, Extra Trees, Decision Tree, and Logistic Regression. Each model underwent hyperparameter tuning via 5-fold stratified cross-validation, with PR-AUC as the optimization criterion. PR-AUC was selected over accuracy and ROC-AUC because, under severe class imbalance, both metrics are susceptible to inflation by the large volume of true negatives, rendering them poor indicators of minority-class detection performance [17, 24]. Following hyperparameter tuning, all models were ranked by test-set PR-AUC and generalization gap, defined as the absolute difference between cross-validation PR-AUC and test-set PR-AUC, to identify candidates that generalize reliably beyond the training distribution.

The composition of the candidate set reflects the structural characteristics of the problem. The input data are low-dimensional tabular records, 14 engineered features across 2.36 million observations, with mixed categorical and numerical variables drawn from monthly billing cycles. This data structure does not exhibit the fine-grained temporal ordering or high-dimensional sequential patterns that recurrent networks, convolutional architectures, and transformer-based models are designed to exploit. Instead, these architectures derive their representational advantage primarily from granular sub-hourly AMI interval data, which is unavailable in this operational context. Recent large-scale benchmarks have further confirmed that gradient-boosting and bagging-based tree ensembles consistently match or outperform deep neural architectures on tabular datasets of this dimensionality and structure [25, 26]. The candidate set was therefore constructed entirely from tree-based and linear methods, which are both empirically appropriate for this data regime and produce interpretable outputs compatible with PLN's operational reporting requirements.

Within this scope, the eight models were chosen to satisfy three additional properties. First, they span three structurally distinct algorithmic families. sequential boosting (XGBoost, LightGBM, CatBoost, Histogram Gradient Boosting), parallel bagging (Random Forest, Extra Trees), and linear and shallow baselines (Logistic Regression as a linear discriminant and Decision Tree as a depth-limited non-linear baseline), providing the error-profile diversity that stacking ensembles require to produce meta-level gains beyond any individual base learner [27].

Second, Logistic Regression was retained not only as a family representative but as a diagnostic baseline: a competitive linear result would indicate that the discriminative signal in the feature set is largely linearly accessible, with direct implications for the minimum model complexity warranted in operational deployment. Third, all eight candidates natively support class-weighting parameters, enabling consistent treatment of the 22.8:1 imbalance through loss-function adjustment rather than synthetic resampling. This approach preserves the original consumer distribution and avoids the boundary instability introduced by interpolation-based oversampling methods such as SMOTE under severe imbalance ratios [17].

The formulation of the problem as supervised binary classification similarly reflects a deliberate and data-driven choice. Labeled ground truth is available from 12,407 field-verified consumer inspections conducted by PLN officers under standard P2TL enforcement procedures, a resource that makes the supervised setting not only feasible but substantially more informative than unsupervised alternatives. Unsupervised anomaly detection methods such as Isolation Forest, Local Outlier Factor, and one-class SVM are most valuable when the anomalous class is poorly defined, or label acquisition is prohibitively expensive; neither condition applies here. The behavioral signatures of electricity theft in non-AMI networks—consumption drops following meter bypass, flatline reading sequences, and power factor anomalies inconsistent with declared load—are sufficiently well characterized in both the technical literature [5, 6] and PLN operational practice to support discriminative learning from confirmed examples. With 12,407 confirmed labeled cases from PLN field inspections, the labeled sample size comfortably exceeds the

threshold at which supervised classification outperforms purely unsupervised alternatives. Therefore, the supervised binary classification formulation represents the most information-efficient use of the available data, and unsupervised methods were not pursued further.

C. Stacking Ensemble Architecture

The stacking ensemble employs a two-level architecture as illustrated in Figure 3. The architectural design was motivated by two empirical observations from the model selection stage: LightGBM and CatBoost consistently achieved the highest individual PR-AUC scores while exhibiting complementary error profiles across the validation folds, suggesting that their combined predictions would yield meta-level gains beyond either model alone.

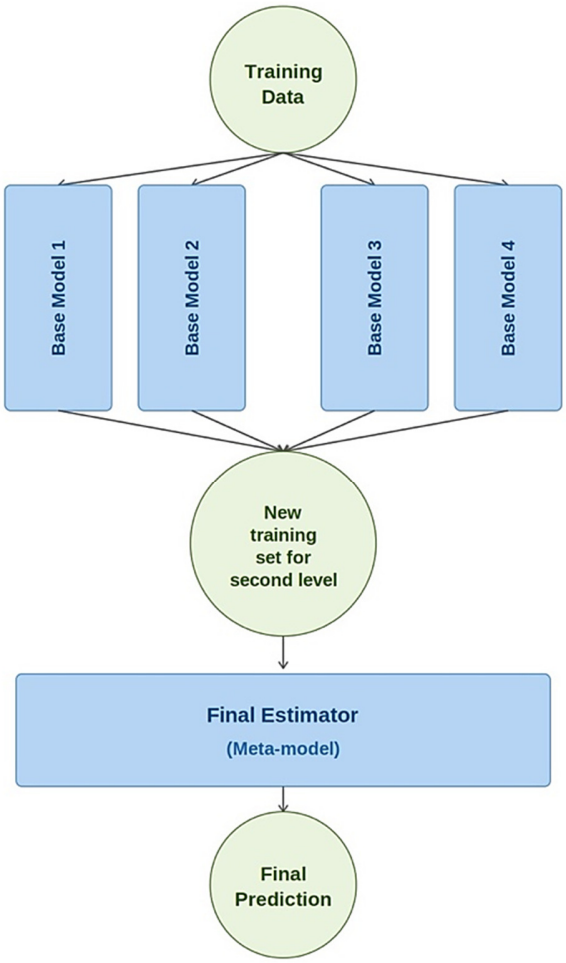

Fig. 3. Two-level stacking ensemble architecture.

At Level 0, the selected base classifiers (LightGBM and CatBoost for Stack-1, Random Forest, Decision Tree, Extra Trees, Logistic Regression, Histogram Gradient Boosting, and XGBoost for Stack-2) are trained independently on the full training partition. To prevent data leakage from the base learners into the meta-learner, all Level 0 predictions used for meta-training are generated exclusively through 5-fold stratified cross-validation applied to the training set. Under this Out-Of-Fold (OOF) prediction protocol, each training observation receives a predicted probability from a base model that was trained on the remaining four folds and never exposed to that observation during training. The complete set of OOF predictions across all five folds constitutes the meta-training feature matrix, ensuring that the meta-learner is trained only on predictions generated under conditions that approximate held-out deployment.

At Level 1, Logistic Regression was selected as the meta-learner for three interdependent reasons. First, its linear combination of base model probability outputs is interpretable and analytically well-suited to integrating the complementary probability signals produced by LightGBM and CatBoost, two models that, while sharing a gradient-boosting foundation, differ in their treatment of categorical features, regularization mechanisms, and leaf-finding algorithms, producing prediction vectors with partially decorrelated error structures. Second, a nonlinear meta-learner, such as XGBoost or a neural network, would introduce additional capacity on a meta-feature space comprising only two probability columns per Stack-1 configuration, substantially increasing the risk of meta-level overfitting without a corresponding gain in expressive power. Third, the class $weight = balanced$ parameter of Logistic Regression directly addresses the propagated class imbalance at the meta-level: even after base learner training with class weighting, the OOF probability distributions passed to the meta-learner retain the original 22.8:1 imbalance ratio, and meta-level rebalancing ensures that the final classification boundary is not biased toward the majority class.

A two-level architecture was selected over deeper stacking configurations because empirical and theoretical analyses have consistently shown that additional levels yield diminishing generalization gains while increasing the risk of overfitting on progressively smaller meta-training feature sets, and increasing computational cost without proportional performance improvement [27].

D. Post-Processing

Following ensemble construction, the finalized model was applied to all 2,363,717 consumers in the full dataset to generate a system-wide fraud probability score per consumer. These scores were rank-ordered to produce a prioritized inspection list, with the highest-scoring consumers identified as the primary candidates for physical field inspection, directly operationalizing the model output as a replacement for the rule-based heuristic prioritization currently in use by PLN field teams. A decision threshold of 0.5 was applied to the meta-learner output probability to derive binary classifications for performance metric computation, chosen to reflect the standard operating point for comparative evaluation across the literature. It is important to note that this threshold selection is a post-hoc evaluation choice and not a design constraint: in operational deployment, PLN management can adjust the threshold continuously to trade-off between detection sensitivity, the proportion of actual theft cases flagged for inspection, and inspection precision, the proportion of inspected consumers confirmed as fraudulent, depending on available field officer capacity, regulatory reporting requirements, and cost-benefit

considerations specific to each service area. Subsequently, a proof-of-concept field validation was conducted using 12,407 physical consumer inspections carried out by PLN field officers operating under standard P2TL enforcement procedures. The inspection sample was drawn from the top-ranked consumers in the model's priority list, providing a direct test of whether the model's fraud probability scores translate into confirmed theft detections under real operational conditions.

IV. MODEL DEVELOPMENT & RESULTS

A. Base Model Performance

Table II presents the test-set performance of all eight base classifiers. Rows are sorted by Test PR-AUC in descending order. Accuracy is reported for completeness and to enable comparison with prior PLN studies [16, 18] that used it as their primary metric, but it is not used for model selection here.

TABLE II. BASE MODEL PERFORMANCE BENCHMARK (TEST SET)

Algorithm	Test PR AUC	Train PR AUC	Gap (Δ)	Rec.	F1	Prec.	Acc.
LightGBM	0.7761	0.810	0.034	0.603	0.696	0.822	0.951
XGBoost	0.7722	0.900	0.128	0.575	0.684	0.844	0.951
Random Forest	0.7613	0.791	0.030	0.556	0.671	0.846	0.942
CatBoost	0.7608	0.772	0.012	0.844	0.582	0.445	0.887
Hist. Grad. Boost	0.7490	0.765	0.015	0.571	0.671	0.812	0.929
Decision Tree	0.7290	0.760	0.031	0.652	0.677	0.704	0.937
Extra Trees	0.6720	0.678	0.006	0.398	0.541	0.842	0.949
Logistic Regression	0.5937	0.600	0.006	0.394	0.507	0.713	0.948

Ranked by Test PR-AUC, LightGBM attains the best discriminative performance at 0.7761, narrowly ahead of XGBoost (0.7722), Random Forest (0.7613), and CatBoost (0.7608). The spread across the eight classifiers is substantial: the weakest model, Logistic Regression, reaches only 0.5937, confirming that a linear decision boundary is insufficient for this task. Examining Test PR-AUC jointly with the generalization gap (Δ) reveals two distinct failure modes. XGBoost has the highest Train PR-AUC (0.9001) but a Δ of 0.1279, showing that its training performance does not transfer reliably to unseen data. In contrast, CatBoost exhibits the smallest gap (Δ = 0.0117), with near-identical Train and Test PR-AUC, indicating stable generalization. LightGBM occupies a favorable middle ground: the highest Test PR-AUC with a modest gap (Δ = 0.0338), consistent with learning transferable fraud patterns rather than memorizing the training distribution.

The classification metrics expose a precision–recall trade-off that is masked when only PR-AUC is reported. A cluster of classifiers—Random Forest (Precision 0.8460), Extra Trees (0.8422), XGBoost (0.8439), and Histogram Gradient Boosting (0.8123)—achieves high precision but substantially lower recall (0.3982–0.5753), reflecting conservative decision boundaries that minimize false positives at the cost of missed fraud. In contrast, CatBoost reaches the highest Recall (0.8438) but records the lowest Precision (0.4446) and Accuracy (0.8874), implying a large false-positive load that would translate into unnecessary field-inspection workload if deployed as a standalone model. Decision Tree offers the most balanced recall-to-precision ratio (Recall 0.6517, Precision 0.7035, F1 0.6766) among tree-based models. Logistic Regression and Extra Trees record the lowest recall values (0.3937 and 0.3982, respectively), failing to detect more than 60% of actual fraud cases despite Accuracy values above 0.94—an illustration of the well-known pitfall of accuracy as a primary metric under severe class imbalance [17, 24].

Table II demonstrates that no single base classifier simultaneously optimizes discriminative power (Test PR-AUC), generalization stability (low Train-Test gap), and a balanced Precision-Recall profile. The Precision-dominant classifiers (Extra Trees, Random Forest) suppress false positives but miss most fraud cases, while the Recall-dominant classifier (CatBoost) detects most fraud but generates excessive false alarms. This performance heterogeneity—rather than representing a weakness—constitutes the empirical foundation for the stacking ensemble design. By combining classifiers with complementary error profiles through a meta-learner trained on their out-of-fold predictions, the ensemble can achieve a detection profile that neither base classifier can attain individually. The selection of LightGBM and CatBoost as the two base learners in Stack-1 is motivated primarily by PR-AUC considerations: LightGBM contributes the highest Test PR-AUC (0.7761) with a well-controlled generalization gap (Δ = 0.0338), while CatBoost contributes the most stable generalization (Δ = 0.0117) paired with the highest recall coverage. Together, they form a complementary pair whose joint probability outputs can be used by the Logistic Regression meta-learner for a more balanced final classification.

Figure 4 presents the Precision-Recall (PR) curves for all eight base classifiers, plotting Precision against Recall across all classification thresholds on both the training set (orange) and test set (blue). PR curves are particularly informative in imbalanced fraud detection settings because they directly visualize the trade-off between a classifier's ability to detect fraud (recall) and its tendency to generate false alarms (precision), without being distorted by the large number of true negatives inherent in the majority class. The degree of separation between the train and test curves serves as a visual indicator of generalization quality, complementing the quantitative generalization gap (Δ) reported in Table II.

XGBoost exhibits the most pronounced train-test divergence among all classifiers, with the train curve sitting substantially above the test curve across the entire Recall range, most notably between Recall values of 0.3 and 0.7. This visual separation corresponds to the largest generalization gap in Table II (Δ = 0.1279), representing a 14.2% drop from the Train PR-AUC (0.9001) to the Test PR-AUC (0.7722). The exceptionally high train curve indicates that XGBoost has memorized discriminative patterns specific to the training distribution, patterns that do not transfer reliably to unseen data. While its Test PR-AUC of 0.7722 remains competitive (second highest in the group), the instability implied by this gap makes XGBoost a less reliable candidate as a standalone classifier in production environments.

In contrast, LightGBM displays two nearly overlapping curves with only a narrow, consistent separation throughout the recall range, yielding the highest Test PR-AUC of 0.7761 and a moderate gap of Δ = 0.0338. The proximity of the train and test curves confirms that LightGBM has learned generalizable fraud patterns rather than overfitting to training-specific noise. The gradual, smooth descent of the test curve indicates stable Precision retention even as Recall increases—a desirable property for operational deployment where the classification threshold may need to be adjusted to balance detection rate against investigative workload.

CatBoost presents the most visually striking generalization behavior: the train and test curves are nearly indistinguishable across the entire recall spectrum, corresponding to the smallest gap of Δ = 0.0117, a mere 1.5% drop from training to test performance. This near-perfect overlap indicates that CatBoost's learned decision boundary is highly stable and not sensitive to the specific samples encountered during training. With a Test PR-AUC of 0.7608, CatBoost achieves competitive discriminative performance alongside exceptional generalization consistency. This combination makes its out-of-fold predictions particularly reliable as meta-features for the Level 1 meta-learner.

Random Forest shows a modest but visible gap between the train and test curves, particularly at higher recall values (> 0.6), where the test curve descends more steeply than the train curve. With Δ = 0.0301 and Test PR-AUC of 0.7613, Random Forest occupies a middle ground: respectable generalization with competitive discriminative performance, though slightly less stable than CatBoost or Histogram Gradient Boosting.

Extra Trees and Logistic Regression both display near-perfect train-test overlap, with gaps Δ of 0.0062 and 0.0058, respectively. However, this visual consistency must be interpreted with caution: in both cases, the overlap reflects low model capacity rather than strong generalization. Extra Trees achieves a Test PR-AUC of only 0.6720, and Logistic Regression records the lowest Test PR-AUC (0.5937) in the entire group. The Logistic Regression curve descends steeply and early, approaching near-random performance at Recall values above 0.5, confirming that a linear decision boundary is fundamentally insufficient to capture the complex, non-linear feature interactions characteristic of electricity theft fraud patterns in this dataset.

Collectively, the PR curve analysis reinforces and extends the findings from Table II. The visual evidence confirms that LightGBM and CatBoost represent the most favorable base learners for Stack-1: LightGBM delivers the highest discriminative performance on unseen data (Test PR-AUC = 0.7761) with a well-controlled generalization profile, while CatBoost provides the most stable train-test consistency (Δ = 0.0117) among classifiers with competitive Test PR-AUC (0.7608). Their complementary strengths—discriminative power versus generalization stability—provide a strong empirical basis for their selection as the core base learners in the proposed stacking ensemble architecture.

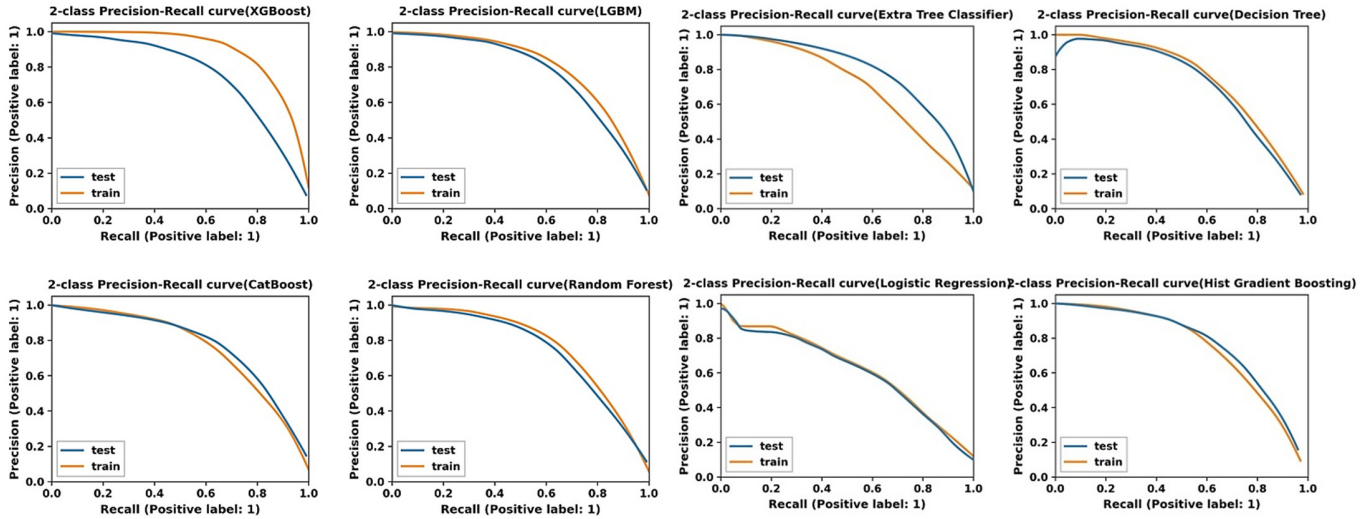

Fig. 4. Precision-Recall curves for all eight base classifiers.

B. *Stacking Ensemble Results*

Table III and Figures 5-7 present the performance of both stacking configurations on the test set. Stack-1 (LightGBM+CatBoost) emerges as the clearly superior configuration across all key metrics, while the multi-metric comparison in Figure 7 contextualizes both stacks relative to the individual base classifiers.

TABLE III. PERFORMANCE METRICS FOR STACKING CONFIGURATIONS

Configuration	PR AUC	Recall	F1	Prec.	Accuracy
Stack-1: LightGBM + CatBoost	0.7764	0.9451	0.8626	0.7933	0.9806
Stack-2: RF+DT+ET+ LR+HGB+XGB	0.7524	0.5938	0.6875	0.8164	0.9497

Stack-1 achieves a Test PR-AUC of 0.7764, the highest among all configurations evaluated, together with a Recall of 0.9451, an F1-score of 0.8626, and an Accuracy of 0.9806. Most critically, its False Negative Rate (FNR) stands at only 5.49%, meaning that it successfully identifies 94.51% of all actual fraud cases in the test set, missing only 1,614 out of 29,387 fraud instances. This represents a fundamental improvement over even the strongest individual base classifier. Compared to LightGBM (the best standalone model), Stack-1 improves Recall by +34.21 percentage points and F1-score by +16.69 points, while maintaining a comparable PR-AUC (0.7764 vs. 0.7761). The 5-fold cross-validation PR-AUC of 0.793±0.011 confirms that this performance is stable and not an artifact of a favourable test split, and the train-test generalization gap of 0.0253 indicates well-controlled overfitting. The PR curve in Figure 5 visually corroborates this, showing two closely aligned curves with only a narrow, consistent separation across the full recall range.

Stack-2 achieves a Test PR-AUC of 0.7524 with a Recall of 0.5938 and an F1-score of 0.6875—substantially below Stack-1 across every fraud-detection-relevant metric. Most notably, Stack-2's FNR is 40.61%, meaning more than 4 in 10 fraud cases are missed. Despite its marginally higher Precision (0.8164 vs. 0.7933), the practical consequence of a 40.61% FNR renders Stack-2 operationally inadequate for deployment. The performance gap between the two configurations, 35.12 percentage points in Recall and 0.0240 in PR-AUC, demonstrates that ensemble gains in this task derive from complementarity of base learners, not merely from increasing the number of classifiers. Stack-2's six diverse classifiers, despite the greater numerical variety, fail to reproduce the precision-recall synergy achieved by pairing LightGBM's discriminative power with CatBoost's recall dominance.

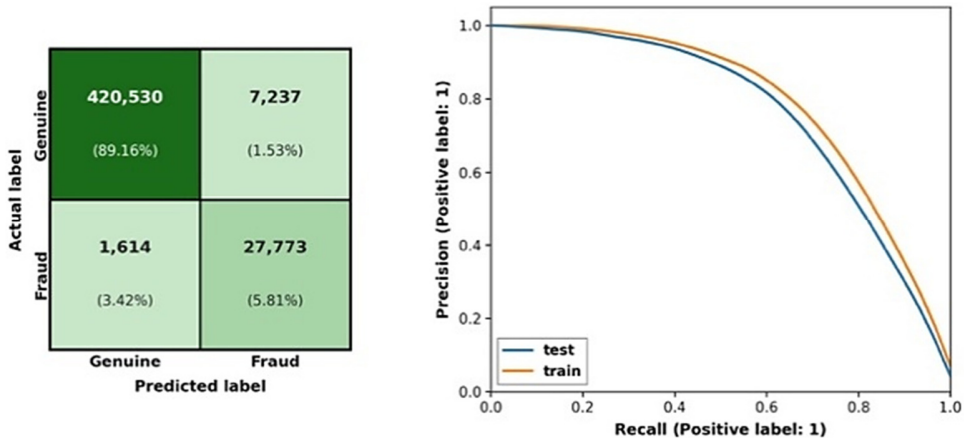

Fig. 5. Confusion matrix and Precision-Recall curve for Stack-1 (LightGBM + CatBoost).

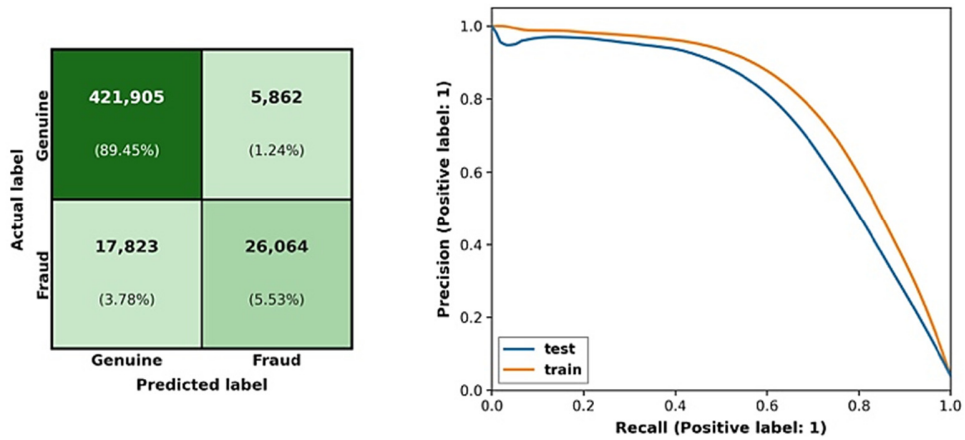

Fig. 6. Confusion matrix and Precision-Recall curve for Stack-2.

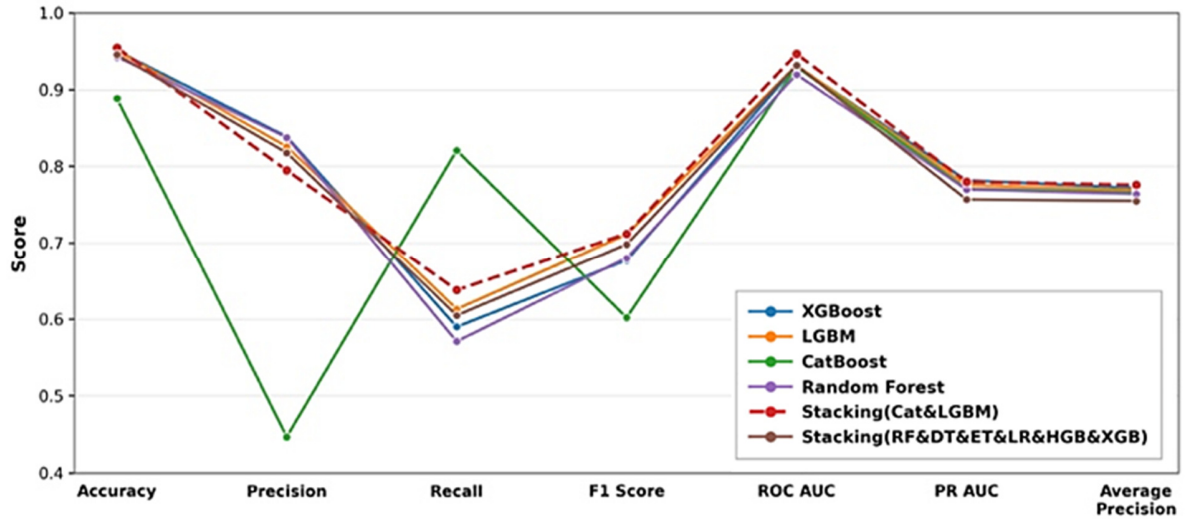

Fig. 7. Multi-metric comparison of all models, including stacking configurations.

The multi-metric comparison in Figure 7 further reinforces this finding. Stack-1 (red dashed line) consistently occupies the upper envelope across Recall, F1-score, ROC-AUC, and PR-AUC, while no individual base classifier approaches its Recall or F1 performance. CatBoost (green) achieves competitive recall in isolation but at the cost of collapsed precision. In contrast, the stacking meta-learner learns to retain CatBoost's detection coverage, while the precision discipline of LightGBM suppresses false positives—an outcome neither model achieves on its own.

Collectively, these results confirm that Stack-1 constitutes the proposed method for deployment, offering the most favourable balance between fraud detection coverage (Recall: 0.9451) and classification reliability (Precision: 0.7933, PR-AUC: 0.7764) among all configurations evaluated.

V. DISCUSSION

A. Model Interpretation

Stack-1's improvement over its base learners is most clearly captured by the joint movement of PR-AUC and the supporting metrics. Stack-1 achieves a Test PR-AUC of 0.7764, marginally exceeding LightGBM's standalone Test PR-AUC (0.7761) but significantly outperforming all other base classifiers on this primary metric—an improvement that, when read alongside the supporting metrics, reflects a substantial reorganization of the Precision-Recall trade-off rather than a marginal numerical gain. Recall rises from 0.6030 (LightGBM) to 0.9451 (Stack-1), while Precision shifts from 0.8222 to 0.7933, indicating that the meta-learner has expanded the set of detected fraud cases without proportional inflation of false positives. This pattern is the empirical signature predicted by stacking theory: meta-level gains accrue when base learners exhibit complementary error profiles across different instances [27]. LightGBM contributes the highest Test PR-AUC and a precision-disciplined probability output, while CatBoost contributes the most stable generalization ($\Delta = 0.0117$) together with the highest recall coverage (0.8438). The Logistic Regression meta-learner, with class-weighted balancing, learns to identify instances where at least one base learner signals a high fraud probability, recovering fraud cases that either individual model would suppress while preserving discriminative ranking quality in the joint probability space. Stack-2's lower Test PR-AUC (0.7524) and recall (0.5938), despite incorporating six base classifiers, demonstrates that ensemble gains in this task derive from base learner complementarity rather than from increasing the number of classifiers. Stack-2's six models share similar conservative decision boundaries and do not provide the high-recall anchor that CatBoost contributes to Stack-1. The persistence of PR-AUC as the discriminative metric of choice under the 22.8:1 class imbalance is further substantiated; both accuracy and AUC-ROC are susceptible to inflation by the majority class at this imbalance ratio and are therefore insufficient to characterize minority-class detection performance [17, 24].

B. Field Validation

Stack-1 was used to score all 2,363,717 West Java consumers, generating a ranked priority list for physical inspection. The top-scoring 12,407 consumers (0.53% of the population) were referred to PLN field inspection teams. Of these, 8,933 (72.0%) were confirmed as electricity theft cases, a precision-at-top-k performance substantially exceeding the prior rule-based system that yielded a confirmation rate of only 15.3% across comparable inspection campaigns. This is a 4.7-fold improvement in inspection efficiency, meaning PLN field officers now identify nearly five confirmed theft cases for every case the legacy system would have identified with the same inspection effort. At an estimated average annual revenue loss of IDR 1.2 million per theft customer, the 8,933 confirmed cases represent approximately IDR 10.7 billion in annual recovery potential, providing a direct economic justification for the framework's operational deployment.

C. Comparison with Related Works

Table IV benchmarks the proposed framework against seven comparable studies. Three key differentiators distinguish this work. First, it is the only study in this comparison that operates in a non-AMI setting and uses monthly manual-

reading data, directly targeting the infrastructure gap faced by utilities in developing countries. Second, it is the only study reporting real-world field validation (n=12,407); all prior studies evaluated performance exclusively on held-out test sets, which cannot account for label noise, operational drift, or the practical cost of false positives. Third, it is the only study adopting PR-AUC as the primary metric, enabling meaningful comparison of minority-class detection performance under severe class imbalance. As shown in Table IV, prior studies on PLN data [16-19] predominantly rely on accuracy as their primary metric. Similarly, recent ML-based approaches [20, 24, 28] employ accuracy-centric evaluation. However, accuracy is methodologically non-comparable under severe class imbalance [17, 24].

TABLE IV. COMPARISON WITH RELATED ELECTRICITY THEFT DETECTION STUDIES

Study	Method	Metric	Score	Field val.
[16]	SMOTE+KNN+LR	Accuracy	98.7%	None
[18]	XGBoost	Accuracy	80.0%	None
[19]	LR, NB, AdaBoost	F1-Score	59.0%	None
[28]	Rule-based ML	F1-Score	N/A	None
[24]	Hybrid DL (MLP+GRU)	Accuracy	~95%	None
[15]	Deep Learning	Accuracy	N/A	None
[20]	Ensemble ML	AUC-ROC	N/A	None
This study	Stack: LightGBM+CatBoost	PR-AUC	0.7764	72%, n=12,407

D. Hyperparameter Configurations

Table V summarizes hyperparameters for LightGBM, CatBoost, and the meta-learner, all selected via 5-fold grid search to optimize PR-AUC.

TABLE V. HYPERPARAMETER CONFIGURATIONS

Model	Key hyperparameters	Tuning
LightGBM	n_estimators=500, lr=0.05, num_leaves=63, min_child_samples=20, class_weight=balanced	Grid Search
CatBoost	iterations=500, depth=6, lr=0.05, l2_leaf_reg=3, auto_class_weights=Balanced	Grid Search
Meta-LR	C=1.0, solver=lbfgs, max_iter=1000, class_weight=balanced	Pre-specified

E. Limitations and Future Work

This study has four principal limitations. First, the ground-truth label reflects confirmed inspections rather than the true theft prevalence: the observed 4.2% theft rate is a lower bound because PLN inspection capacity limits the sample to high-priority consumers, and systematic under-labeling of uninspected areas cannot be ruled out. Second, the framework was validated within a single utility (PLN West Java); transferability to other distribution networks—with different grid topologies, consumer profiles, and loss characteristics—requires independent validation. Third, social and spatial features (population density, GRDP, vulnerability index) are aggregated at the administrative-area level (93 areas) rather than at the individual-consumer level, which may obscure within-area heterogeneity. Fourth, the fixed threshold of 0.5 was used for all reported metrics; in deployment, threshold calibration should be performed to reflect the specific cost asymmetry between missed theft (unrecovered revenue) and false positives (unnecessary inspection cost). Future directions include SHAP-based feature importance analysis to improve model interpretability and support regulatory reporting, and extension to AMI-enabled consumers as PLN expands smart meter deployment.

VI. CONCLUSION

This paper presented a hybrid stacking ensemble framework for electricity theft detection in non-AMI distribution networks. By integrating 14 multidimensional features across consumer, technical, social, and spatial domains, benchmarking eight classifiers using PR-AUC as the primary metric, and constructing a two-level stacking ensemble with LightGBM and CatBoost as complementary base learners, the proposed Stack-1 configuration achieved a test PR-AUC of 0.7764 (CV: 0.793 ± 0.011), Recall of 0.9451, and F1-score of 0.8626, outperforming all single-model baselines and the six-classifier Stack-2 alternative. The recall gain over the best standalone model (+34.21 pp) is attributed to the meta-learner's ability to exploit the complementary probability outputs of a high-precision and a high-recall base classifier in a joint probability space. Real-world field validation across 12,407 consumer inspections confirmed a 72.0% detection rate, which is a 4.7-fold improvement over the prior rule-based system, yielding an estimated IDR 10.7 billion in annual revenue recovery potential. To the best of our knowledge, this is the largest field-validated machine learning study for electricity theft detection in a non-AMI network reported in the open literature. The framework is operationally deployable, requires no AMI infrastructure, and is directly transferable to utilities in developing countries, where smart meter rollout remains years away.

CONFLICTS OF 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.

ACKNOWLEDGMENT

This research was funded by Universitas Indonesia (UI) under the Publikasi Terindeks Internasional (PUTI) 2024 scheme, grant number NKB-477/UN2.RST/HKP.05.00/2024. The authors also gratefully acknowledge PT PLN (Persero) West Java Distribution Region for providing access to data and for supporting the field validation.

DATA AVAILABILITY STATEMENT

The customer-level billing and inspection data used in this study are proprietary to PT PLN (Persero) and are not publicly available due to customer privacy obligations and company data-protection policies. The anonymized and aggregated data that support the findings of this study may be made available from the corresponding author upon reasonable request and subject to prior written approval from PT PLN (Persero). The feature-engineering framework, model configurations, and hyperparameters required to reproduce the analytical pipeline on comparable utility datasets are reported in full in this paper.

ETHICS STATEMENT

This study used operational billing and inspection records collected by PT PLN (Persero) during routine distribution-network operations. No personally identifiable consumer information is reported. Field inspections were conducted by PLN field officers under the utility's standard internal procedures for suspected non-technical loss.

REFERENCES

[1] "World Energy Outlook 2023 – Analysis," International Energy Agency, Oct. 2023. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2023.

[2] P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey," *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, pp. 760–775, Jan. 2017, https://doi.org/10.2991/ijcis.2017.10.1.51.

[3] D. Carr and M. Thomson, "Non-Technical Electricity Losses," *Energies*, vol. 15, no. 6, Mar. 2022, https://doi.org/10.3390/en15062218.

[4] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007–1015, Feb. 2011, https://doi.org/10.1016/j.enpol.2010.11.037.

[5] R. Czechowski and A. M. Kosek, "The most frequent energy theft techniques and hazards in present power energy consumption," in *2016 Joint Workshop on Cyber- Physical Security and Resilience in Smart Grids (CPSR-SG)*, Apr. 2016, pp. 1–7, https://doi.org/10.1109/CPSRSG.2016.7684098.

[6] N. Calamaro, Y. Beck, R. B. Melech, and D. Shmilovitz, "An Energy-Fraud Detection-System Capable of Distinguishing Frauds from Other Energy Flow Anomalies in an Urban Environment," *Sustainability*, vol. 13, no. 19, Sept. 2021, https://doi.org/10.3390/su131910696.

[7] M. H. Sidiq, D. Arla, S. H. Fadliqaf, and Q. Haramaini, "Anti Theft Power Smart Metering System," in *2024 International Conference on Technology and Policy in Energy and Electric Power (ICTPEP)*, Sept. 2024, pp. 287–290, https://doi.org/10.1109/ICT-PEP63827.2024.10732892.

[8] H. O. Henriques *et al.*, "Development of adapted ammeter for fraud detection in low-voltage installations," *Measurement*, vol. 56, pp. 1–7, Oct. 2014, https://doi.org/10.1016/j.measurement.2014.06.015.

[9] Z. Yan and H. Wen, "Electricity Theft Detection Base on Extreme Gradient Boosting in AMI," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021, https://doi.org/10.1109/TIM.2020.3048784.

[10] L. M. R. Raggi, F. C. L. Trindade, V. C. Cunha, and W. Freitas, "Non-Technical Loss Identification by Using Data Analytics and Customer Smart Meters," *IEEE Transactions on Power Delivery*, vol. 35, no. 6, pp. 2700–2710, Sept. 2020, https://doi.org/10.1109/TPWRD.2020.2974132.

[11] L. Wei, A. Sundararajan, A. I. Sarwat, S. Biswas, and E. Ibrahim, "A distributed intelligent framework for electricity theft detection using benford's law and stackelberg game," in *2017 Resilience Week (RWS)*, Sept. 2017, pp. 5–11, https://doi.org/10.1109/RWEEK.2017.8088640.

[12] Z. Yan and H. Wen, "Performance Analysis of Electricity Theft Detection for the Smart Grid: An Overview," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–28, 2022, https://doi.org/10.1109/TIM.2021.3127649.

[13] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010, https://doi.org/10.1109/TPWRD.2009.2030890.

[14] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016, June 2016, https://doi.org/10.1109/TII.2016.2543145.

[15] S. M. Saqib *et al.*, "Deep learning-based electricity theft prediction in non-smart grid environments," *Heliyon*, vol. 10, no. 15, Aug. 2024, https://doi.org/10.1016/j.heliyon.2024.e35167.

[16] Y. Maraden *et al.*, "Enhancing Electricity Theft Detection through K-Nearest Neighbors and Logistic Regression Algorithms with Synthetic Minority Oversampling Technique: A Case Study on State Electricity Company (PLN) Customer Data," *Energies*, vol. 16, no. 14, July 2023, https://doi.org/10.3390/en16145405.

[17] Y. Wang, M. M. Rosli, N. Musa, and F. Li, "Multi-Class Imbalanced Data Classification: A Systematic Mapping Study," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14183–14190, June 2024, https://doi.org/10.48084/etasr.7206.

[18] B. Nurhuda, "Design and development of an electricity theft detection system using XGBoost," B.S. Thesis, Universitas Indonesia, 2024.

[19] A. S. Manneken, "Electricity theft detection based on machine learning using the logistic regression method," B.S. Thesis, Universitas Indonesia, 2022.

[20] S. M. Saqib *et al.*, "Utilizing machine learning ensembles for effective electricity theft detection," *Energy Exploration & Exploitation*, vol. 44, no. 1, pp. 526–553, Jan. 2026, https://doi.org/10.1177/01445987251381989.

[21] D. Perez, M. Flores, P. Castaneda, J. Santisteban, and A. Onate-Andino, "Detection of Non-Technical Losses in Electrical Metering Systems in Northern Lima Using Predictive Modeling and Business Intelligence," *Engineering, Technology & Applied Science Research*, vol. 16, no. 1, pp. 31624–31631, Feb. 2026, https://doi.org/10.48084/etasr.14923.

[22] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, Feb. 2019, https://doi.org/10.1109/TSG.2018.2807925.

[23] Q. Wang, Z. Luo, and P. Li, "Natural Gas Consumption Forecasting Based on Homoheterogeneous Stacking Ensemble Learning," *Sustainability*, vol. 16, no. 19, Oct. 2024, https://doi.org/10.3390/su16198691.

[24] H. Iftikhar *et al.*, "Electricity theft detection in smart grid using machine learning," *Frontiers in Energy Research*, vol. 12, Mar. 2024, https://doi.org/10.3389/fenrg.2024.1383090.

[25] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Advances in Neural Information Processing Systems*, vol. 35, pp. 507–520, Dec. 2022.

[26] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022, https://doi.org/10.1016/j.inffus.2021.11.011.

[27] S. Zian, S. A. Kareem, and K. D. Varathan, "An Empirical Evaluation of Stacked Ensembles With Different Meta-Learners in Imbalanced Classification," *IEEE Access*, vol. 9, pp. 87434–87452, 2021, https://doi.org/10.1109/ACCESS.2021.3088414.

[28] S. Bahrami, E. Yumuk, A. Kerem, B. Topçu, and A. Kaya, "Electricity Theft Detection Using Rule-Based Machine Leaning (rML) Approach," *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, vol. 12, no. 2, pp. 438–456, June 2024, https://doi.org/10.29109/gujsc.1443371.