Journal of Electrical and Electronic Engineering
2026, Vol. 14, No. 1, pp. 46–53
https://doi.org/10.11648/j.jeee.20261401.15

Research Article

Robust Energy Theft Detection in Smart Distribution Method Using a Data-driven Method

Olushola Akintola*, Babatunde Adetokun, Oghenewvogaga Oghorada

Electrical and Electronics Engineering Department, Nile University of Nigeria, Abuja, Nigeria

Abstract

Energy theft poses a significant challenge to modern power systems, leading to economic losses, reduced efficiency, and compromised reliability in smart grids. Detecting such anomalies requires robust, scalable analytical frameworks that can accurately distinguish normal consumption, marginally increased usage, and patterns indicative of electricity theft across diverse operating conditions. This study investigates the application of machine learning techniques for energy theft detection using a dataset of recorded consumption values. Two numerical features, energy used by theft in per unit and normal energy, were employed as predictors. At the same time, the target variable comprises three categorical conditions: Theft detected, Normal, and Energy slightly higher. Four classifiers were implemented and compared: Decision Tree, Support Vector Machine (SVM) with Error-Correcting Output Codes (ECOC), Random Forest, and k-Nearest Neighbors (kNN). The models were trained and evaluated using MATLAB with an 80/20 hold-out validation approach. Performance was assessed using accuracy metrics and confusion matrices. Results demonstrated that SVM achieved the highest accuracy (86.67%), followed closely by Random Forest (83.33%) and kNN (82.33%), while Decision Tree yielded the lowest accuracy (73.33%). Confusion matrix analysis showed that all classifiers detected theft-based cases with high accuracy, whereas most classification errors arose from overlap and ambiguity between normal consumption and elevated energy usage conditions. The study adds to the expanding literature on data-driven energy management by providing practical evidence of how machine-learning techniques can strengthen grid security, minimize financial losses, and enhance overall operational efficiency.

Keywords

Distribution Network, Energy Theft, Energy Theft Detection, Machine Learning

1. Introduction

Malfunctions in metering equipment or electricity leakage can lead to significant revenue losses for power utilities, thereby undermining their financial interests. To address this, power supply centers typically deploy inspectors or metrology personnel to conduct routine inspections of metering devices to detect leakage or electricity theft, while also collecting consumption data and abnormal alarm signals for further analysis [1]. However, the presence of false alarms and redundant information at the terminal level makes it difficult to rapidly identify faulty meters and suspicious users, despite the large volume of abnormal electricity consumption data generated through these measures [2].

*Correspondence: Olushola Akintola (sholakintola@gmail.com)

Received: 21 January 2026; **Accepted:** 31 January 2026; **Published:** 11 February 2026

Electricity theft has long been a widespread and persistent problem. In Nigeria, numerous households engage in various forms of electricity theft and meter tampering, while in India, more than one-fifth of total electricity generation is lost due to such activities [3]. In Ghana, illegal practices, including electricity theft, result in losses of nearly 30% of the power supplied by utility companies [4]. Non-technical losses (NTLs) constitute approximately 10–40% of total system losses, with the majority attributed to fraud and energy theft. India loses about 25% of its electricity generation, whereas China loses roughly 16%, leading to a combined annual economic loss estimated at $96 billion [5].

Developed countries are also not immune to this issue. For instance, British Columbia Hydro reports annual losses of around $100 million in Canada. In the United States, electricity utilities lose approximately $6 billion each year, with nearly 80% of these losses linked to electricity theft, metering faults, and related issues.

Electricity theft techniques often involve interrupting the current loop or inserting series resistance to lower the voltage, thereby altering meter operation and causing meters to run slowly, stop, or even record reverse energy flow. Beyond financial implications, electricity theft also complicates load estimation for utilities, potentially leading to generation unit overloading and, in severe cases, widespread blackouts [2].

To mitigate losses caused by energy theft in smart grids, a wide range of Energy Theft Detection (ETD) methodologies has been investigated. The ETD approaches are generally categorized into hardware-based and data-driven methods [6]. Hardware-based ETD techniques were primarily developed to identify theft resulting from physical tampering with mechanical meters in conventional power systems. In contrast, data-driven ETD methods aim to detect suspected energy theft by analyzing electricity consumption profiles and power flow patterns.

2. Literature Review

With the evolution of smart grids, research attention has increasingly shifted toward data-driven ETD, as modern energy theft is more commonly associated with data manipulation rather than direct physical interference with metering devices [7]. Most existing studies have employed machine learning and deep learning techniques to ensure the effective performance of data-driven ETD systems. While deep learning–based ETD models can achieve high accuracy when trained on balanced datasets, their effectiveness often deteriorates in the presence of imbalanced data distributions [8]. Under such conditions, ETD models may become biased, perform well only in limited scenarios, and face significant challenges in terms of generalization and scalability.

Figure 1 illustrates the key technical components involved in designing a data-driven ETD framework. These typically include data acquisition, data preprocessing, malicious behaviour modelling, and the development of intelligent algorithms capable of identifying theft activities.

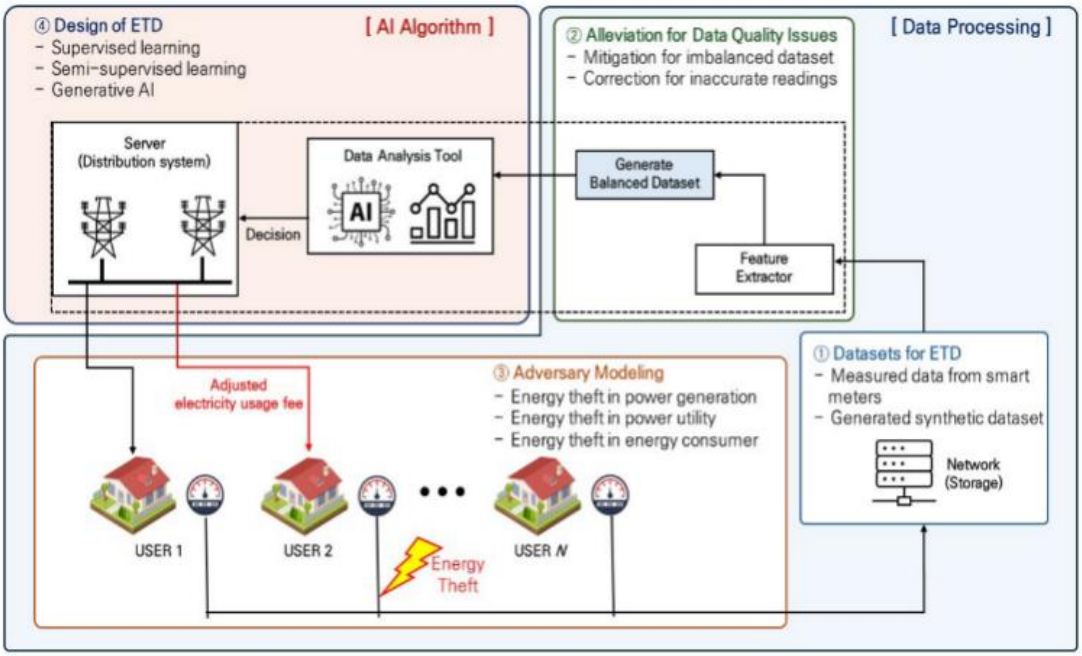

Figure 1. *Overall design process of data-driven ETD.*

Figure 2 presents a comprehensive classification of proposed data-driven ETD approaches, aimed at addressing existing challenges and improving methodological effectiveness [9]. The framework captures multiple dimensions of energy theft detection, including different theft mechanisms such as meter tampering, meter faults, cyber-attacks, feeder tapping, and billing anomalies. It also highlights key dataset-related challenges, including high dimensionality, class imbalance, erroneous measurements, and incomplete or missing labels. To tackle these issues, data-driven ETD techniques are broadly grouped into supervised learning, semi-supervised learning, and generative AI–based approaches. This is described as shown in Figure 2.

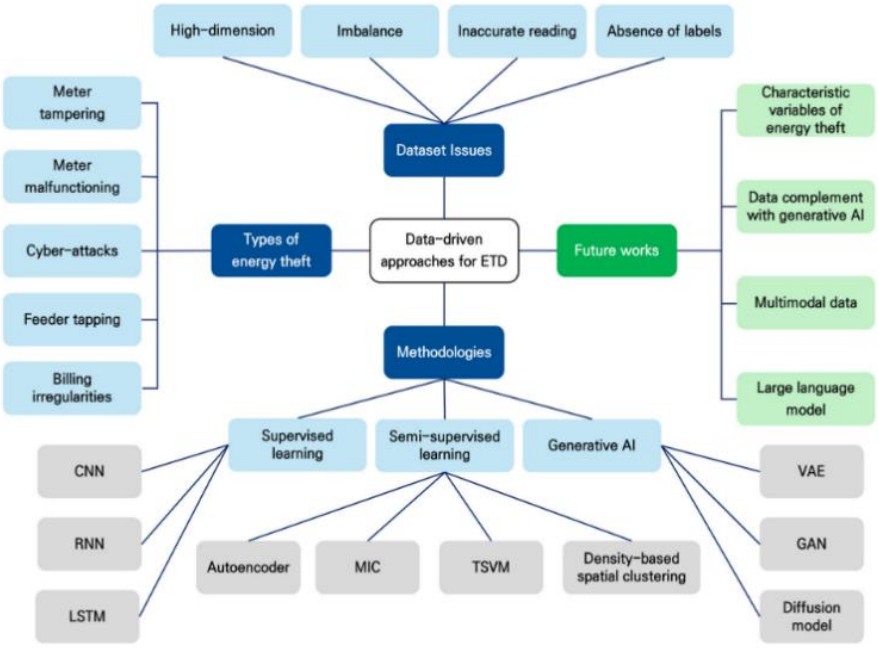

Figure 2. *Overview and categorization of data-driven approaches for ETD.*

2.1. Related Works of Literature

Over the past few decades, data mining techniques have been widely applied in electricity distribution systems for tasks such as tariff design [10], pricing strategy formulation [11], false data injection (FDI) attack detection, and customer classification [12]. Nevertheless, comparatively limited attention has been devoted to the detection of electricity leakage and theft. Several representative studies are summarized, [13] proposed a hybrid detection framework that combines Decision Tree (DT) and Support Vector Machine (SVM) classifiers to identify potential electricity theft users, achieving a detection accuracy of 92.50%. In another study, multiple machine learning algorithms, including Logistic Regression (LR), K-Nearest Neighbor (K-NN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), were evaluated for electricity theft prediction within a comparative modeling framework [14]. Their findings indicated that, in certain scenarios, relatively simple classifiers can still deliver competitive performance. The work by [15] investigated several deep learning architectures for NTL and ETD, including Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) networks, Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNN). However, the lack of efficient hyperparameter optimization in these models limits their generalization capability when addressing large-scale and diverse electricity theft scenarios. Similarly, [16] developed a Wide & Deep CNN-based model composed of a Wide component and a Deep CNN component, attaining a maximum MAP@100 value of 94.04%, which reflects strong detection performance.

The challenges associated with energy theft detection (ETD) and non-technical losses (NTLs) are escalating rapidly on a global scale, prompting researchers to explore mitigation strategies based on statistical methods, machine learning, and deep learning techniques [17, 18]. Various machine learning algorithms, including Random Forest (RF), Decision Trees (DT), Bagging Ensemble (BE), Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN), have been comparatively assessed for automated ETD in smart grid environments. Among these, RF demonstrated approximately 10% higher detection accuracy than the other evaluated methods [19]. The increasing prevalence of NTLs and electricity theft thus remains a critical challenge for distribution network operators. In [20], a neural network–based approach employing

Neural Architecture Search (NAS) was proposed for electricity theft analysis and detection under missing data conditions. By integrating density-based spatial clustering of applications with noise clustering techniques, the model achieved a strong performance, reporting an area under the curve of 0.926 for NTL and ETD tasks.

2.2. Contribution of the Work

Despite the promising results reported in the literature, the effectiveness of these methods remains limited when suitable and informative features are not properly extracted from meter readings. Consequently, further research is required to develop more robust, scalable, and feature-efficient electricity theft detection techniques. The contribution of this work is as follows:

(a) A systematic comparison of four widely used machine learning algorithms, DT, SVM with Error-Correcting Output Codes, RF, and kNN, was conducted. The study highlights their relative strengths and weaknesses in detecting theft-based, normal, and borderline consumption patterns.

(b) Unlike many prior studies that focus on binary classification (theft vs. non-theft), this work addresses a multi-class problem by distinguishing between Normal, Energy slightly higher, and Theft-based categories. This provides a more nuanced and realistic representation of consumption anomalies.

(c) Beyond reporting overall accuracy, the study incorporates confusion matrix evaluation to reveal misclassification trends. This dual-level analysis strengthens the reliability of conclusions and provides deeper insight into classifier behaviour, especially in borderline cases.

3. Materials and Methods

This study adopts a supervised machine learning–based comparative methodology to detect energy theft in a smart grid environment. The proposed framework involves data acquisition, feature selection, data partitioning, classifier training, performance evaluation, and comparative analysis. Four widely used classification algorithms, Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbors (kNN), are implemented and evaluated under identical conditions to ensure fairness and consistency. The overall workflow is illustrated through systematic stages implemented in the MATLAB environment.

To preserve the integrity of the dataset and ensure compatibility with the original column headers, the data are imported using MATLAB's readtable function with the VariableNamingRule set to preserve.

Two numerical features are extracted for analysis:

1) Energy used by theft (per unit)
2) Normal energy consumption (per unit)

These features represent abnormal and legitimate energy usage patterns, respectively, and serve as the input variables for model training. The target variable, referred to as Condition, represents the operational state of the consumer (e.g., normal or theft) and is converted into a categorical format to support multi-class classification. No artificial normalization or scaling is applied, as the features are already expressed in per-unit values, which inherently provide scale uniformity.

To evaluate the generalization capability of the classifiers, the dataset is divided into training and testing subsets using a hold-out validation strategy. Specifically:

80% of the data is used for training.

20% is reserved for testing.

The partitioning is performed using MATLAB's cvpartition function to ensure random and unbiased sample allocation. The testing dataset is strictly excluded from the training phase to avoid data leakage and optimistic bias in performance evaluation. Four supervised machine learning classifiers are implemented and trained using the same training dataset.

$$Accuracy = \frac{Number\ of\ Correction\ Predictions}{Total\ Number\ of\ Predictions\ Made} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (3)$$

$$F1\ Score = \frac{2\ X\ Precision\ X\ Recall}{Precison+Recall} \quad (4)$$

4. Results and Discussion

This section presents and discusses the results obtained from the comparative evaluation of four supervised machine learning classifiers: Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbors (kNN) for energy theft detection. The evaluation is conducted using a hold-out testing dataset comprising 20% of the total samples, which was not exposed to the models during training. Performance assessment is based on classification accuracy and confusion matrix analysis to ensure both quantitative and qualitative evaluation of the models.

Figure 3 compares the classification accuracy of four machine learning models used for energy theft detection. The DT model records the lowest accuracy of 73.33%, indicating limited robustness when used alone. The SVM achieves the highest accuracy (approximately 86%), highlighting its effectiveness in capturing non-linear consumption patterns. kNN both show competitive performance, each with 83.33% accuracy. Overall, the results demonstrate that advanced models, particularly SVM and Random Forest, outperform simpler classifiers, confirming their suitability for reliable energy theft detection in smart grid applications as depicted in Table 1.

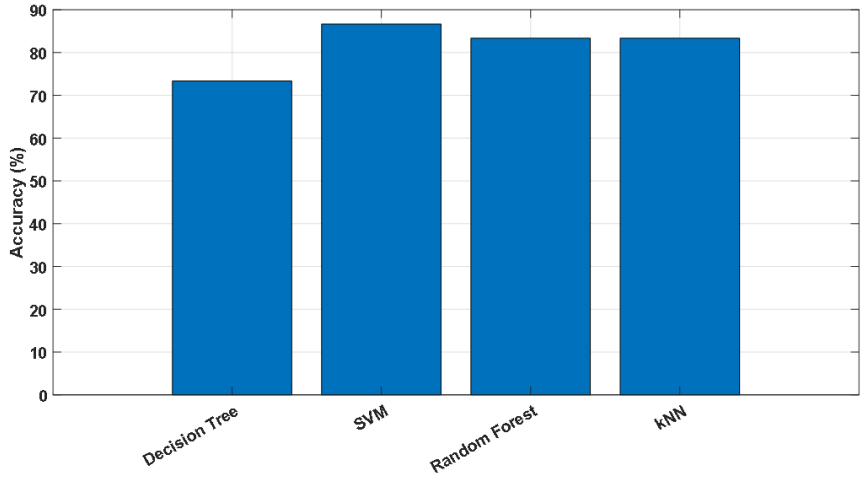

Figure 3. *Comparative analysis of classifiers for Energy theft Detection.*

Table 1. *Energy theft classifier.*

Classifier	Accuracy (%)
Decision Tree	73.33
SVM	86.67
Random Forest	83.33
KNN	83.33

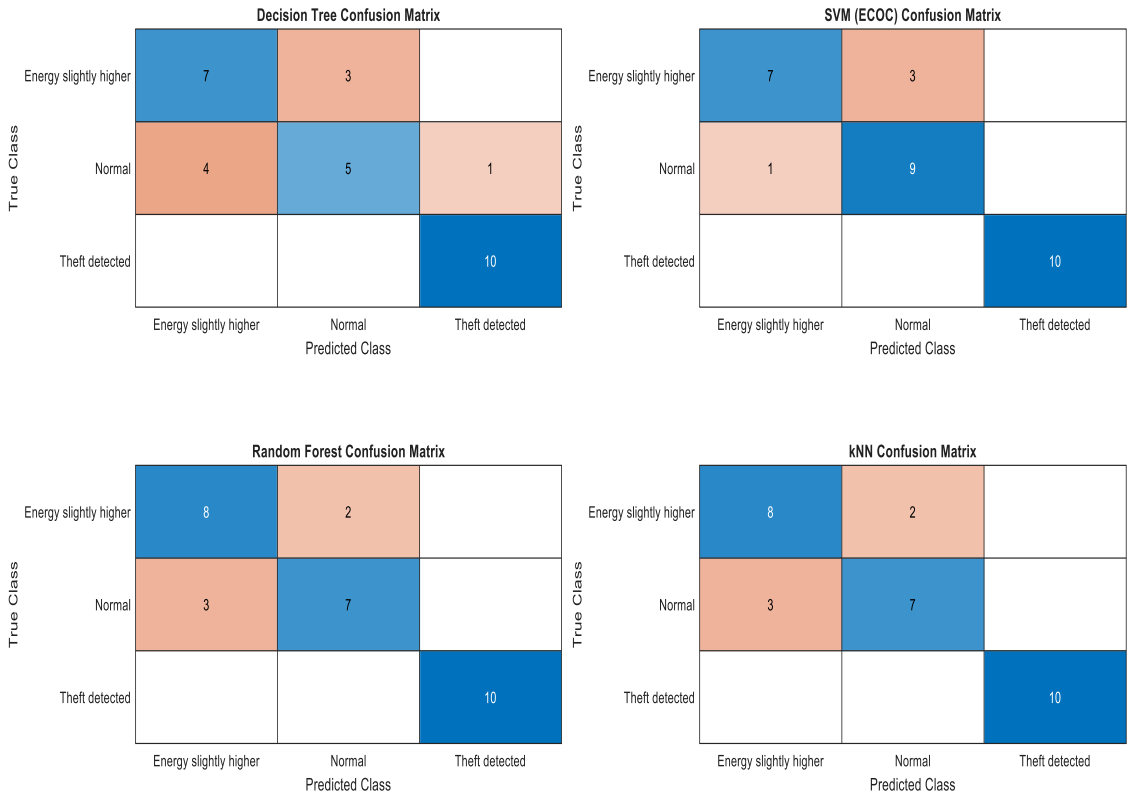

Figure 4. *Confusion matrix for the classifiers.*

Figure 4 presents the confusion matrices for four machine learning classifiers: DT, SVM using ECOC, RF, and kNN, applied to the multi-class problem of energy theft detection. Each matrix compares predicted labels with true labels across three categories: Energy slightly higher, Normal, and Theft-based. The results show that all classifiers consistently achieved perfect classification of Theft-based cases, reflecting the distinct separability of theft-related energy usage patterns. However, misclassifications were observed between Normal and Energy slightly higher conditions, which represent more subtle consumption differences. The Decision Tree correctly classified all theft cases but struggled with Normal instances, misclassifying several as Energy slightly higher. The SVM (ECOC) demonstrated the most balanced performance, with improved accuracy across all three classes and fewer misclassifications between Normal and Energy, slightly higher. The Random Forest achieved strong results overall, particularly for Normal cases, though some misclassifications occurred with borderline consumption patterns. The kNN classifier performed comparably to Random Forest but showed sensitivity to overlapping feature distributions, occasionally misclassifying Energy slightly higher as Theft-based.

The results of the study were compared with other existing works as shown in Table 2.

Table 2. *Comparison of results from existing literature.*

	Decision Tree	SVM	Random Forest	KNN
References	Accuracy (%)			
[21]	75.83	77.80	---	82.83
[22]	----	81.00	80.00	79.00
[23]	----	85.70	----	----
[24]	----	83.00	----	----
Present Study	73.33	86.67	83.33	83.33

The result of the study showed that the accuracy of the present study had superior performance to the other works reviewed in the literature.

5. Conclusions

This study has demonstrated the effectiveness of machine learning techniques in detecting energy theft within modern power systems. By applying four classifiers, DT, SVM with Error-Correcting Output Codes, RF, and kNN to a dataset of consumption patterns, comparative analysis revealed clear differences in predictive performance. The results showed that SVM and Random Forest consistently outperformed the other models, achieving accuracies of 86.67% and 83.33%, respectively, with balanced classification across all categories. Both models exhibited robustness in distinguishing subtle consumption anomalies, particularly between Normal and Energy slightly higher conditions. In contrast, Decision Tree, while interpretable, yielded the lowest accuracy (73.33%), and kNN, though competitive (83.33%), was more sensitive to overlapping feature distributions. Confusion matrix analysis further confirmed that theft-based cases were consistently identified across all models, highlighting the distinct separability of theft patterns in the feature space. These findings underscore the importance of selecting advanced, scalable algorithms for deployment in smart grid environments. Ensemble and kernel-based methods provide the necessary balance of accuracy, robustness, and generalization required for real-world applications. The study contributes to the growing body of research advocating for data-driven solutions in energy management, offering practical insights into how machine learning can enhance grid security, reduce economic losses, and improve operational efficiency.

Future work should focus on extending this framework to larger, more diverse datasets, incorporating feature engineering for improved anomaly detection, and exploring hybrid approaches that combine interpretability with predictive strength. Such advancements will further strengthen the scalability and reliability of energy theft detection systems, supporting the transition toward smarter, more resilient power infrastructures.

Abbreviations

ANN	Artificial Neural Network
DT	Decision Tree
ECOC	Error-correcting Output Codes
ETD	Energy Theft Detection
kNN	k-nearest Neighbor
LR	Logistic Regression
NAS	Neural Architecture Search
NTL	Non-technical Loss

RNN	Recurrent Neural Network
SVM	Support Vector Machine

Author Contributions

Olushola Akintola: Conceptualization, Data Curation, Formal Analysis, Software, Methodology, Writing – original draft.

Babatunde Adetokun: Conceptualization, formal analysis, validation, supervision, Writing – review & editing.

Oghenewvogaga Oghorada: Formal Analysis, validation, supervision, writing – review & editing.

Funding

This work is not supported by any external funding.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] F. Dewangan, A. Y. Abdelaziz, and M. Biswal, "Load Forecasting Models in Smart Grid Using Smart Meter Information: A Review," Energies. 2023, vol. 16, pp. 1–55. https://doi.org/10.3390/en16031404

[2] J. Chen, Y. A. Nanehkaran, W. Chen, Y. Liu, and D. Zhang, "Data-driven intelligent method for detection of electricity theft," International Journal of Electrical Power & Energy Systems. 2023, vol. 148, p. 108948. https://doi.org/10.1016/j.ijepes.2023.108948

[3] R. Razavi and M. Fleury, "Socio-economic predictors of electricity theft in developing countries: An Indian case study," Energy for Sustainable Development. 2019, vol. 49, pp. 1–10. https://doi.org/10.1016/j.esd.2018.12.006

[4] C. L. Athanasiadis, T. A. Papadopoulos, G. C. Kryonidis, and D. I. Doukas, "A review of distribution network applications based on smart meter data analytics," Renewable and Sustainable Energy Reviews. 2024, vol. 191, p. 114151. https://doi.org/10.1016/j.rser.2023.114151

[5] O. Yakubu, N. Babu C., and O. Adjei, "Electricity theft: Analysis of the underlying contributory factors in Ghana," Energy Policy. 2018, vol. 123, pp. 611–618. https://doi.org/10.1016/j.enpol.2018.09.019

[6] S. Kim et al., "Data-Driven Approaches for Energy Theft Detection: A Comprehensive Review," Energies. 2024, 17(12). https://doi.org/10.3390/en17123057

[7] M. Z. Gunduz and R. Das, "Smart Grid Security: An Effective Hybrid CNN-Based Approach for Detecting Energy Theft Using Consumption Patterns," Sensors. 2024, 24, 1148. https://doi.org/10.3390/s24041148

[8] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," Artif. Intell. Rev. 2024, 57(6), pp. 1-51. https://doi.org/10.1007/s10462-024-10759-6

[9] X. Gong, B. Tang, R. Zhu, W. Liao, and L. Song, "Data Augmentation for Electricity Theft Detection Using Conditional Variational Auto-Encoder," Energies. 2020, 13(17), pp. 1-14. https://doi.org/10.3390/en13174291

[10] S. Chandrasekaran, "Multiobjective optimal power flow using interior search algorithm: A case study on a real-time electrical network," Comput. Intell. 2020, 36(3), pp. 1078–1096. https://doi.org/10.1111/coin.12312

[11] H. Shahinzadeh, A. Mahmoudi, J. Moradi, H. Nafisi, E. Kabalci, and M. Benbouzid, "Anomaly Detection and Resilience-Oriented Countermeasures against Cyberattacks in Smart Grids," Proceedings - 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS), Tehran, Iran, 2021, pp. 1-7. https://doi.org/10.1109/ICSPIS54653.2021.9729386

[12] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," Electric Power Systems Research. 2021, vol. 192, p. 106904. https://doi.org/10.1016/j.epsr.2020.106904

[13] A. Hirsi et al., "HSF: A Hybrid SVM-RF Machine Learning Framework for Dual-Plane DDoS Detection and Mitigation in Software-Defined Networks," IEEE Access. 2025, vol. 13, pp. 112303–112323. https://doi.org/10.1109/ACCESS.2025.3583712

[14] A. P. Taruna, G. Arisona, D. Irwanto, A. B. Bestari, and W. Juniawan, "Electricity Theft Detection Using Machine Learning in Traditional Meter Postpaid Residential Customers: A Case Study on State Electricity Company (PLN) Indonesia," IEEE Access. 2025, vol. 13, pp. 7167–7191. https://doi.org/10.1109/ACCESS.2025.3526764

[15] N. M. Elshennawy, D. M. Ibrahim, and A. M. Gab Allah, "An efficient electricity theft detection based on deep learning," Scientific Reports. 2025 15(1), pp. 1-15, https://doi.org/10.1038/s41598-025-93140-z

[16] Z. Jiang, X. Liu, and L. Zhang, "Wide and Deep Learning-Aided Nonlinear Equalizer for Coherent Optical Communication Systems," Photonics. 2024, 11(2), p. 141. https://doi.org/10.3390/photonics11020141

[17] V. K. Jaiswal, H. K. Singh, and K. Singh, "Arduino GSM-based Power Theft Detection and Energy Metering System," 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 448–452. https://doi.org/10.1109/ICCES48766.2020.9138085

[18] E. Stracqualursi, A. Rosato, G. Di Lorenzo, M. Panella, and R. Araneo, "Systematic review of energy theft practices and autonomous detection through artificial intelligence methods," Renewable and Sustainable Energy Reviews. 2023, vol. 184, p. 113544. https://doi.org/10.1016/j.rser.2023.113544

[19] S. Zidi, A. Mihoub, S. Mian Qaisar, M. Krichen, and Q. Abu Al-Haija, "Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment," Journal of King Saud University - Computer and Information Sciences. 2023, 35(1), pp. 13–25. https://doi.org/10.1016/j.jksuci.2022.05.007

[20] K. Fei, Q. Li, and C. Zhu, "Non-technical losses detection using missing values' pattern and neural architecture search," International Journal of Electrical Power & Energy Systems. 2022, vol. 134, p. 107410. https://doi.org/10.1016/j.ijepes.2021.107410

[21] Y.-C. Tsao, D. Rahmalia, and J.-C. Lu, "Machine-learning techniques for enhancing electricity theft detection considering transformer reliability and supply interruptions," Energy Reports, vol. 12, pp. 3048–3064, 2024. https://doi.org/10.1016/j.egyr.2024.08.068

[22] I. Petrlik, P. Lezama, C. Rodriguez, R. Inquilla, J. Elizabeth Reyna-González, and R. Esparza, "Electricity Theft Detection using Machine Learning," International Journal of Advanced Computer Science and Applications. 2022, 13(12), pp. 420–425, 2022. https://doi.org/10.14569/IJACSA.2022.0131251

[23] N. G. Ezeji, K. I. Chibueze, and N. H. Nwobodo-Nzeribe, "Developing and Implementing an Artificial Intelligence (AI)-Driven System for Electricity Theft Detection," ABUAD Journal of Engineering Research and Development (AJERD). 2024, 7(2), pp. 317–328. https://doi.org/10.53982/AJERD.2024.0702.31-J

[24] S. A. Abro, G. L. Hua, J. A. Laghari, M. A. Bhayo, and A. A. Memon, "Machine learning-based electricity theft detection using support vector machines," International Journal of Electrical and Computer Engineering (IJECE). 2024, 14(2), pp. 1240–1250. http://arxiv.org/abs/1809.03006