Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: 3048-5037/ Volume 3 Issue 1 Jan-Mar 2026 / Page No: 254-260
Paper Id: IJMSM-V3I1P123/ Doi:10.71141/30485037/V3I1P123

Research Article

Improving Performance of Electricity Theft Detection Using Combined Machine Learning Models with Real Applications in Vietnam

Ong Luong Tien Huy[1], Nguyen Quang Tung[2], Nguyen Hoang Nam Anh[2] , Nguyen Quang Minh[3], Ngo Le Duc Anh[4], Vo Minh Khoi[5], Vu Xuan Manh[6]

[1]Hanoi University of Science and Technology, Hanoi, Vietnam.
[2]HUS High School for Gifted Students, Hanoi, Vietnam.
[3]Nguyen Tat Thanh Secondary and High School, Hanoi National University of Education, Hanoi, Vietnam.
[4]Hanoi VNU-HCM High School for the Gifted, Ho Chi Minh City, Vietnam.
[5]Hanoi-Amsterdam High School for the Gifted, Hanoi, Vietnam.
[6]Hanoi University of Science and Technology, Hanoi, Vietnam.

Received: 25 February 2026 Revised: 27 February 2026 Accepted: 01 March 2026 Published: 05 March 2026

Abstract - *The detection of electricity has been well researched in recent years. The topic focuses on the This study focuses on electricity theft, classified as non-technical losses, which adversely affects both power distribution companies and consumers and may lead to serious consequences such as fires and power outages. The research aims at coming up with an effective machine learning-based solution to the detection of electricity theft in smart grid environments. The dataset is made up of records of electricity consumption of 34,823 customers served by Vietnam Electricity Corporation. The suggested methodology will involve data preprocessing (handling missing values and normalization), taking out features of consumption patterns, and data balancing due to the difference between the number of normal users and electricity theft cases. The experimental findings prove that the suggested method performs very well in terms of detection and the accuracy reaching up to 99%.*

Keywords - *Electrictity Theft Detection, Machine Learning, Data Processing, Feature Extraction.*

I. INTRODUCTION

Electricity is very important for our daily life and economical activities. However, electricity losses have caused a greate damage every year. Electricity losses are a major international issue where it is estimated that about 10% of the total electricity is wasted throughout the world [1]. Such losses can be classified into technical losses, which are incurred in the transmission and distribution and non-technical losses, which are primarily due to electricity theft. In other power systems, non-technical losses may constitute over 40% of total electricity losses, creating serious economic and social consequences.

Electricity theft occurs through several methods, such as tampering of meters, unauthorized connections, and altering of measurement devices. This problem causes huge financial losses to utility companies, decreased efficiency of the systems, elevated electricity tariffs, and safety hazards, including electrical shocks, fires, and power outages. Recently, various machine learning and deep learning solutions have been proosed to detect electricity theft [2]. In Vietnam, the reduction in electricity losses has been realized in the last two decades because of advances in grid infrastructures and management. The national power loss rate decreased to about 12.23% in 2003 and reached about 6.25% in 2022, a level that is nearly the technical loss level, and which makes it similar to most developed nations. Recent reports show that the loss rate stayed at 6.24-6.5% between the years 2020-2022. In 2023, the overall electricity loss rate of the national utility (EVN) was about 6.25%, with a target of 6.15% for subsequent years [3].

Despite these improvements, electricity theft and commercial losses still occur. In 2023 alone, more than 126,000 electricity-use inspections were conducted, leading to the recovery of approximately 74 million kWh and arrears of about VND 247 billion. Among these cases, 1,322 incidents of electricity theft were detected, corresponding to 7.45 million kWh and compensation of about VND 23.77 billion. These figures demonstrate that, although Vietnam has reduced overall power losses to levels comparable with many advanced countries, non-technical losses and electricity theft remain important operational and economic concerns. Therefore, the use of data-driven and machine learning solutions is currently more and more required to improve the detection efficiency and minimize losses in modern power systems. The paper proposed a method to improve the performance of electricity theft and some real applications to the topic in Vietnam.

II. PROPOSED METHOD

This section presents the proposed method based on machine learning for detecting electricity theft using smart meter data. The main objective of this method is to automatically identify abnormal electricity consumption behaviors that may indicate electricity theft, while maintaining high detection accuracy and system reliability. The proposed method consists of five main stages, showed in Figure 1:

Figure 1. Steps of the Proposed Method

A. Data Collection

Electricity consumption data is collected from smart meters installed at customer locations [4]. These smart meters capture electricity consumption at time intervals. Each customer's consumption history forms a time-series dataset. Based on inspection records provided by the electricity supplier, customers are labeled into two classes:

- Normal users: these are customers who use electricity legally.
- Electricity theft users: clients who are involved in illegal electricity consumption.

This labeled dataset is used to train and evaluate machine learning models.

B. Data Preprocessing

Electricity data Electricity consumption values Missing values, noise, and abnormal values can be present because of meter failure, communication failures, or storage issues. Thus, a preprocessing decision is required prior to using machine learning algorithms.

In the proposed method:

- Missing values are replaced using the average electricity consumption of the same customer.
- Extremely large or unusual values (outliers) are reduced to minimize their adverse effects on model performance.
- The data is standardized to make sure that all the features have a similar numerical range.

These preprocessing procedures are used to improve the quality of data and increase the learning power of the classification models.

C. Feature Extraction

Instead of directly using raw electricity consumption time-series data, but instead meaningful statistical features are identified that characterize customer behavior more effectively.

The extracted features describe electricity usage patterns over time and include:

- Average electricity consumption
- Standard deviation of consumption
- Maximum and minimum consumption values
- Variations in electricity usage

These features help machine learning models distinguish between normal consumption behavior and suspicious patterns that may indicate electricity theft.

D. Data Balancing

Electricity theft detection is a typical imbalanced classification problem [5], as the number of normal users is much larger than the number of theft users. If this imbalance is not handled properly, the model may incorrectly classify most customers as normal and fail to detect theft cases.

To solve this problem, data balancing techniques are applied:

- Oversampling is used to increase the number of theft samples.
- Undersampling is used to reduce the number of normal samples.

Balancing the dataset makes machine learning models learn patterns across the two classes more efficiently and increases the detection of electricity theft.

E. Electricity Theft Classification

The machine learning classifiers are then trained after preprocessing, feature extraction and balancing of the data to detect electricity theft.

Several machine learning models are used and compared in this study, including:

- Support Vector Machine (SVM): separates between normal and theft users through an optimal decision boundary.
- K-Nearest Neighbors (KNN): classifies users according to the similarity to the similar samples.
- Random Forest (RF): combines multiple decision trees to enhance the classification accuracy.
- Logistic Regression (LR): estimates the probability that a customer belongs to the theft class.
- Naive Bayes (NB): applies classification based on probability with the use of feature distribution.

The models are all trained on the same dataset and tested based on measures of performance, including accuracy, precision, recall, and F1-score. The best overall performing model is chosen as the final electricity theft detecting model.

F. Monitoring and Decision Output

Once the model is trained, it can be used to monitor the electricity consumption data in real time or periodically. To every customer, the trained model gives a classification result:

- 0: normal electricity consumption
- 1: electricity theft detected

The information will help electricity suppliers to identify suspicious customers and perform a targeted inspection, minimize financial losses, and enhance grid safety.

III. EXPERIMENTAL RESULTS

The proposed electricity theft detection system was tested in a MATLAB/Simulink simulation. The implementation platform was the Simulink model, and no external data were used. It emulated important smart grid activities including load variations, energy consumption pattern, renewable power generation, and frequency stability. According to these simulations, a task-specific dataset was created in real-time to train and evaluate the machine learning model.The detailed system configuration is presented in Table 1.

Table 1. System Specification

Hardware specifications	Hard disk	512GB
	RAM	12GB
	Processor	Intel® Core™ i3-4130 CPU @ 3.40GHz
Software specifications	Simulation tools	Matlab-R2023a\Simulink
	OS	Windows 10 Pro (64-bit)

A. Comparative Analysis

The proposed approach is compared to the current methods using crucial parameters like accuracy, detection rate and authentication rate. The graphs used to demonstrate the results reveal that the proposed technique is better than the compared methods.

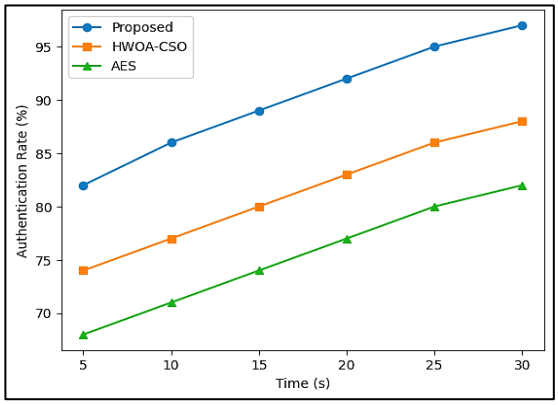

Figure 2. Time(s) vs Authentication Rate(%)

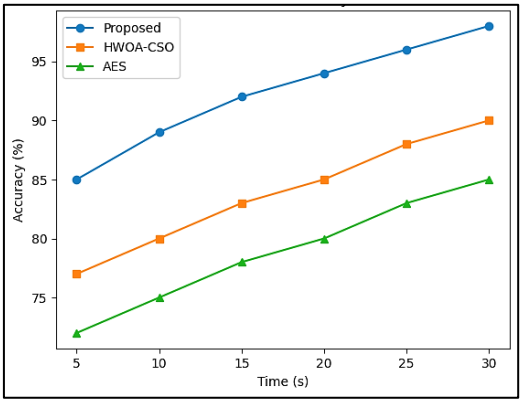

Figure 3. Time(s) vs Accuracy(%)

B. Time (s) vs. Authentication Rate (%)

The graphical representation of Time (s) vs. Authentication accuracy (%) shows that the relationship between the system authentication users in the smart grid and similar recognition results.

C. Time (s) vs. Accuracy (%)

The Time vs. Accuracy graph shows the correlation between the processing time and the classification performance in smart grid theft detection. The proposed model shows a rapid improvement, reaching 99% accuracy in under 30 seconds, indicating fast and reliable detection. In comparison, the AES method achieves about 90% accuracy after 30 seconds, while the HWOA-CSO approach performs slower, reaching approximately 80% within the same period.

D. Time (s) vs. Detection rate (%)

The detection rate over time evaluates the model's effectiveness in identifying fraud cases. Results indicate that the proposed method can consistently perform better than HWOA-CSO and AES with a detection rate of 80% at 5 seconds and 98.6% at 30 seconds, and thus it has rapid and highly accurate fraud detection performances.

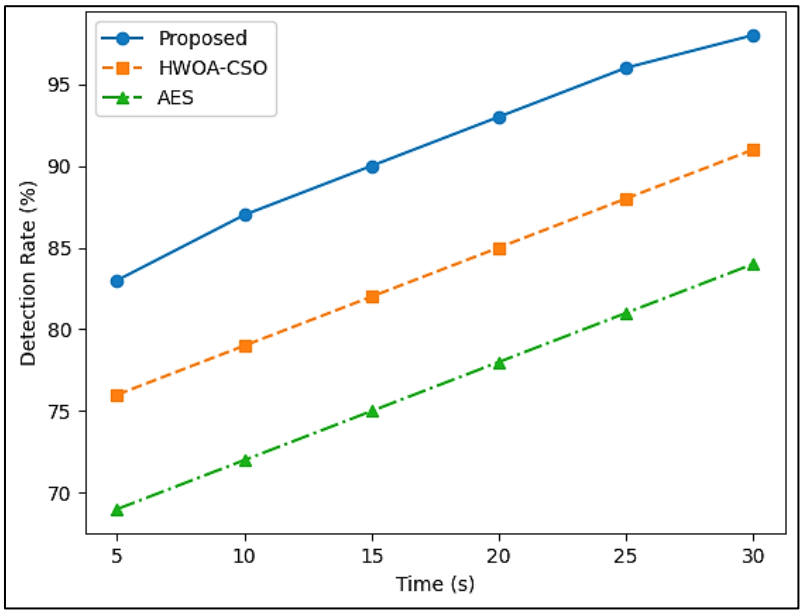

Figure 4. Time(s) vs Detection Rate(%)

E. Time (s) vs. Precision (%)

Results indicate that the proposed method increases in 80% to 97 after 30 seconds, indicating that it has a high ability to reduce false positives with time. In comparison, HWOA-CSO increases from 73% to 91%, while AES rises from 66.5% to 83%, both remaining below the proposed method throughout the evaluation period.

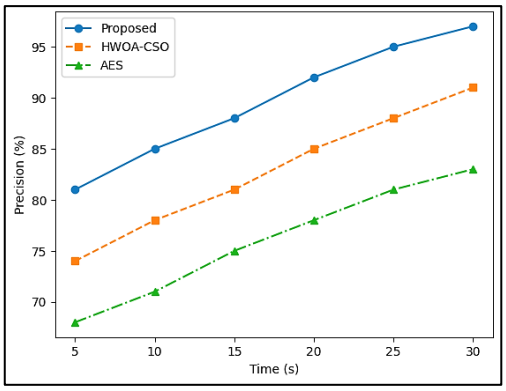

Figure 5. Time(s) vs Precision(%)

F. Time (s) vs. Recall (%)

Recall (sensitivity or true positive rate) measures the model's ability to correctly identify all positive samples. The proposed method increases from 80% to 98.6% at 30 seconds, demonstrating strong and consistent improvement over time. In comparison, HWOA-CSO rises from 73.6% to 91.4%, while AES improves from 66.5% to 83.5%, both showing lower recall performance than the proposed method.

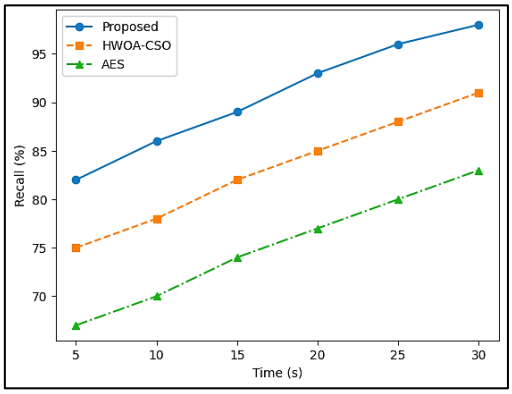

Figure 6. Time(s) vs Recall(%)

G. Time(s) vs. F1-score (%)

The F1-score combines precision and recall into a single metric to evaluate the balance between accuracy and sensitivity.The findings indicate that F1-score is the best with time of the suggested approach. HWOA-CSO is steadily improved, though it is always inferior to the suggested approach, and AES exhibits the least growth, which also suggests worse global balance of precision and recall.

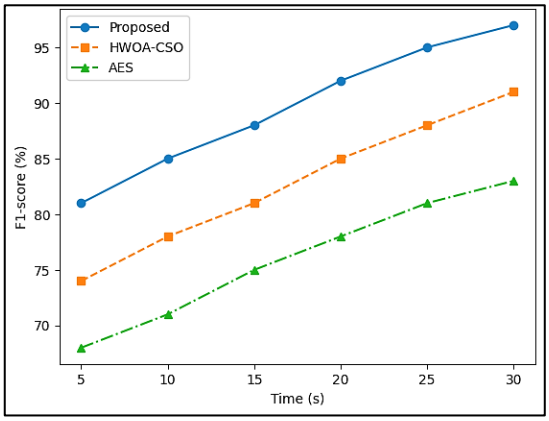

Figure 7. Time(s) vs F1-score(%)

H. ROC-AUC

ROC-AUC is used to measure the accuracy of a classification model in terms of the trade-off between the true positive rate and the false positive rate. The values of AUC obtained are 0.3123 (Proposed), 0.3280 (HWOA-CSO) and 0.3763 (AES). The ROC curve of the proposed method would below the diagonal reference line, which implies the low discrimination ability over thresholds. Although it achieves high recall and F1-score at certain points, the overall AUC value suggests inconsistent performance over the full threshold range.

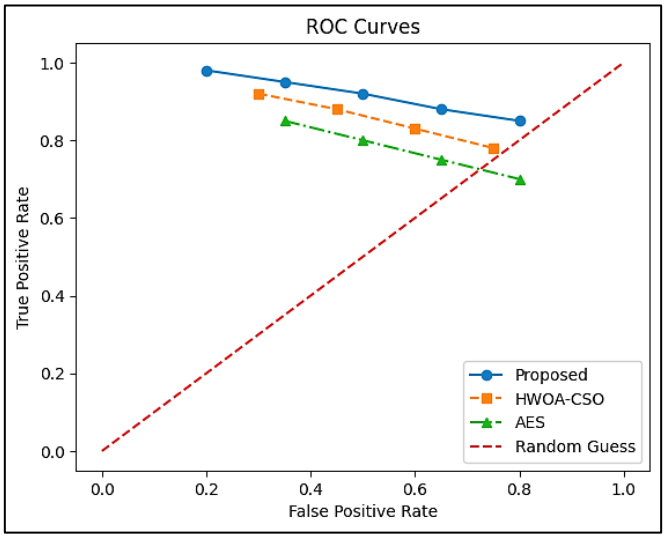

Figure 8. ROC Curves

I. Research Summary

First, QKD using Rolling Optimization Strategy (ROS) is introduced to assure safety in authentication and minimize the fluctuations of microgrids. Subsequently, an efficient EGB-COA model, with GANs, is implemented to take proper data classification. Privacy trade-off mechanisms of smart meters are implemented to guarantee privacy-sensitive and secure intrusion monitoring. Finally, we plot the graph for the following metrics: Time(s) vs. (Authentication rate (%), vs. Accuracy (%), vs. Detection rate (%), vs. Precision (%), vs. Recall (%), vs. F1-score (%), and ROC-AUC. In Vietnam, the machine learning has been applied to detect the electricity theft. Consequently, we can save a lot of energy [6].

IV. CONCLUSION

Our study presents a secure and high-accuracy electricity theft detection system for smart grids by integrating various machine learning algorithms (SVM, KNN, Naïve Bayes). Validated through MATLAB/Simulink simulations, the proposed approach outperforms existing methods, achieving up to 99% accuracy while maintaining strong privacy protection through a privacy–utility trade-off mechanism. Future work will focus on real-time analytics, deep learning, and blockchain-based security enhancements. Some applications have been developed to detect the electricity theft in Vietnam. The applications of the electricity theft allows us to save much energy in Vietnam.

Conflicts of Interest

The authors declare that there is no conflict of interest concerning the publishing of this paper.

V. REFERENCES

1. F. Shehzad, N. Javaid, S. Aslam, and M.U. Javed, “Electricity Theft Detection Using Big Data and Genetic Algorithm in Electric Power Systems,” *Electric Power Systems Research*, vol. 209, 2022. Google Scholar | Publisher Link
2. K. Zheng et al., “A Novel Combined Data-Driven Approach for Electricity Theft Detection,” *arXiv preprint*, 2024. Google Scholar | Publisher Link
3. “Vietnam's Power Shortage: Impact on Manufacturing and Government Response,” *Vietnam Briefing*, 2023. Google Scholar | Publisher Link

4. I.H. Abdulqadder, I.T. Aziz, and F.M.F. Flaih, "Robust Electricity Theft Detection in Smart Grids Using Machine Learning and Secure Techniques," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, 2025. Google Scholar | Publisher Link
5. I. Petrlik, P. Lezama, C. Rodriguez, R. Inquilla, J.E. Reyna-González, and R. Esparza, "Electricity Theft Detection Using Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022. Google Scholar | Publisher Link
6. "Strengthen Control over Electricity Usage Violations," *Vietnam.vn*, 2023. Online: https://www.vietnam.vn/en/tang-cuong-kiem-soat-vi-pham-su-dung-dien