IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Smart Electricity Theft Detection Using Behavioural Feature Engineering and Machine Learning

Aamna Zaheer Khan[1] and Ms. Yogita Kadbane[2]
[1]Student, Master of Computer Applications
[2]Assistant Professor, of Computer Applications
Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

Abstract : *Electricity theft is a common problem for electricity distribution companies and can result in financial losses and difficulties in managing electricity supply. Traditional methods of detecting electricity theft mainly depend on manual inspection, meter checking and predefined rules. These methods may take more time and may not identify unusual electricity consumption behaviour effectively.*

In this study, Machine Learning is used to identify possible electricity theft based on consumer electricity consumption patterns. The study uses the SGCC Electricity Theft Detection dataset containing 42,372 consumer records with electricity consumption information over different dates. Behavioural features such as average consumption, consumption variation, zero-consumption days, nonzero consumption and missing-consumption patterns are created from the original data. Five Machine Learning models, namely Decision Tree, Random Forest, Support Vector Machine, XG Boost and Multi-Layer Perceptron, are developed and compared..

Keywords: Electricity Theft Detection, Machine Learning, Behavioural Feature Engineering, SMOTE, Random Forest, XG Boost, Smart Grid

I. INTRODUCTION

Electricity plays an important role in homes, businesses and industries. Electricity theft means using electricity illegally or changing the meter or connection in a way that causes the actual electricity consumption to be incorrectly recorded. It can create financial losses for electricity distribution companies and can also affect proper electricity management. Electricity theft detection is difficult because consumer electricity usage can change depending on their daily activities and requirements. Traditional detection methods mainly depend on manual checking, customer complaints and fixed rules. These methods may not always identify unusual consumption patterns, especially when a large number of consumers have to be monitored .

Machine Learning provides a useful approach for analysing large amounts of electricity consumption data. It can study previously collected consumption patterns and learn differences between normal and suspicious consumer behaviour. However, electricity consumption datasets may contain a large number of daily readings, missing values and an unequal number of normal and theft records. In this research, behavioural features are created from electricity consumption data to represent the usage pattern of each consumer. SMOTE is used to handle the class imbalance, and five Machine Learning models are developed and compared. Feature importance analysis is also performed to understand which consumption patterns are useful for identifying possible electricity theft.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

OBJECTIVE OF STUDY

The objective of this study is to analyse consumer electricity consumption behaviour and identify patterns related to possible electricity theft. The study focuses on preparing the electricity consumption data, creating meaningful behavioural features and developing Machine Learning models for classification. The study also aims to handle the imbalance between normal and theft consumers using SMOTE and compare Decision Tree, Random Forest, SVM, XG Boost and MLP models using different classification measures. In addition, feature importance is analysed to understand which consumption patterns contribute more to the model's prediction. The overall aim is to develop a simple and useful Machine Learning-based approach that can support electricity theft investigation.

LITERATURE REVIEW

Many researchers have studied the use of Machine Learning for electricity theft detection.

[1] Razavi et al. (2019) proposed a practical feature-engineering framework for electricity theft detection in smart grids. Their study used electricity demand data from more than 4,000 households and created new features from consumption behaviour. A Gradient Boosting Machine was then used for classification. The study showed that feature engineering can improve electricity theft detection and also examined different types of theft, including irregular and unseen attacks

[2] Gunturi and Sarkar (2021) studied ensemble Machine Learning models for detecting energy theft using customer electricity consumption patterns. They compared several ensemble algorithms including Random Forest, Extra Trees and different boosting methods. The study also considered the class imbalance problem and used techniques such as SMOTE and Near-Miss. Their results showed that ensemble models can provide good detection performance and reduce false-positive predictions .

[3] Appiah et al. (2023) proposed an Extremely Randomised Trees model for electricity theft detection. Their study focused on problems commonly found in electricity consumption data, such as missing values and class imbalance. The researchers used SMOTE-Tomek to handle class imbalance and Grid Search for model optimisation. Their results showed that the Extra Trees approach achieved strong performance for electricity theft detection

[4] Iftikhar et al. (2024) developed a Machine Learning-based electricity theft detection system using data from the Chinese National Grid Corporation. Their research addressed the high dimensionality and class imbalance problems in electricity consumption data. The proposed approach used preprocessing, a k-means SMOTE technique and a combination of GRU and MLP models. The study showed that balancing the data and using Machine Learning can improve the detection of suspicious electricity consumption behaviour.

[5] Alfaverh et al. (2026) studied electricity theft detection using electricity, gas and water measurements. Their research used the TDD2022 dataset and compared several Machine Learning models, including Random Forest, Decision Tree, XGBoost, LightGBM, CatBoost, Extra Trees and Logistic Regression. SMOTE was used to handle class imbalance, while LASSO and ReliefF were used for feature selection. The study further shows the importance of combining suitable features, balancing methods and different Machine Learning algorithms for electricity theft detection.

Research Gap

From the reviewed studies, several research gaps are identified. Electricity consumption datasets generally contain a large number of daily readings, which can make direct analysis difficult.

Therefore, meaningful behavioural features are required to represent the consumption pattern of each consumer.

Another important issue is class imbalance because normal consumers are much higher in number than theft consumers. This can affect the ability of Machine Learning models to identify the minority theft class. Different Machine Learning models may also show different performance depending on the characteristics of the dataset and the selected features.Previous studies have also used different combinations of features and algorithms, but there is a need for a systematic comparison of multiple Machine Learning models using the same behavioural feature set. In addition, identifying which consumption patterns are most useful for the model can provide better understanding of the detection

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

process. Based on these gaps, the present study focuses on behavioural feature engineering, SMOTE-based class-imbalance handling, comparison of five Machine Learning models and identification of important electricity consumption patterns.

RESEARCH METHODOLOGY

The proposed research methodology is designed to identify possible electricity theft using consumer electricity consumption behaviour and Machine Learning. The methodology consists of data collection, data selection, preprocessing, behavioural feature engineering, train-test splitting, class-imbalance handling, feature scaling, model development, model evaluation and feature-importance analysis.

The overall process starts with the SGCC electricity consumption dataset. The raw consumption information is first cleaned and prepared. Behavioural features are then created to represent the consumption pattern of each consumer. The prepared data is divided into training and testing data, and SMOTE is applied to the training data to balance the normal and theft classes. Five Machine Learning models are then trained and evaluated using different performance measures.

Data Collection and Feature Selection

The SGCC Electricity Theft Detection dataset is used as the primary data source for this research. The dataset contains 42,372 consumer records and 1,036 columns. The dataset includes electricity consumption values for different dates, consumer identification information and a target variable named FLAG.

The FLAG variable represents the class of the consumer. A value of 0 represents a normal consumer, while a value of 1 represents a theft consumer. The dataset contains 38,757 normal records and 3,615 theft records. This shows that the dataset has a considerable difference between the two classes.

The consumer identification column CONS_NO is not used as a predictive feature because it identifies the consumer rather than describing electricity consumption behaviour. The daily consumption values are used to create meaningful behavioural features for the Machine Learning models.

Data Preprocessing

The selected dataset was processed before model development. The consumer identification column was removed from the input features and the consumption columns were identified separately. Columns containing very high levels of missing values were removed to reduce the effect of incomplete information.The consumption values were converted into numerical form and infinite values were handled as missing values. Missing values were handled during feature preparation. After preprocessing and feature engineering, the original large set of daily consumption readings was represented using 16 behavioural features.These preprocessing steps helped prepare the dataset for Machine Learning and reduced the complexity of the original consumption data.

Behavioural Feature Engineering

The original dataset contains a large number of daily electricity consumption readings for each consumer. Instead of directly using all these readings, behavioural features were created to describe the general electricity usage pattern of each consumer.The features include mean consumption, median consumption, standard deviation, minimum consumption, maximum consumption, consumption range, zero-consumption days, missing days, missing percentage, nonzero days, nonzero percentage, quartile values, interquartile range, coefficient of variation and zero percentage.These features provide information about the level, variation and behaviour of electricity consumption. For example, standard deviation and consumption range describe how much the consumption changes, while zero-consumption and missing-consumption features describe unusual gaps or patterns in electricity usage.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

Risk-Level / Class Classification

The target variable in this research is the FLAG column. The value 0 represents normal electricity consumption and the value 1 represents electricity theft.

The dataset contains considerably more normal consumers than theft consumers. Therefore, simply training a model on the original data may make it easier for the model to learn the majority class. This class imbalance was considered during the model development process.

Handling Class Imbalance

The data was divided into training and testing sets using an 80:20 stratified split. Before applying SMOTE, the training data contained 31,005 normal records and 2,892 theft records. SMOTE was applied only to the training data to increase the representation of the minority theft class. After applying SMOTE, the training data contained 31,005 normal records and 31,005 theft records.

The test data was kept in its original distribution so that the models could be evaluated using the actual class distribution.

Model Training and Machine Learning Models

The prepared dataset was divided into 80% training data and 20% testing data. Feature scaling was performed using Standard Scaler before training the models. Five supervised Machine Learning algorithms were developed and compared.

Decision Tree was used as a simple tree-based classification model. It makes decisions using a series of conditions based on the input features.

Random Forest combines multiple decision trees and uses their combined results for classification. It can identify different relationships between the behavioural features.

Support Vector Machine attempts to find a suitable boundary between normal and theft classes. It was included to compare a different classification approach with the tree-based models.

XG Boost is a boosting-based Machine Learning algorithm that builds trees sequentially. Each new tree attempts to improve the errors made by previous trees. It was included because it can capture complex relationships between features.

Multi-Layer Perceptron is a neural-network-based model that learns relationships between input features and the target class. It was included to compare a neural-network approach with the other Machine Learning models.

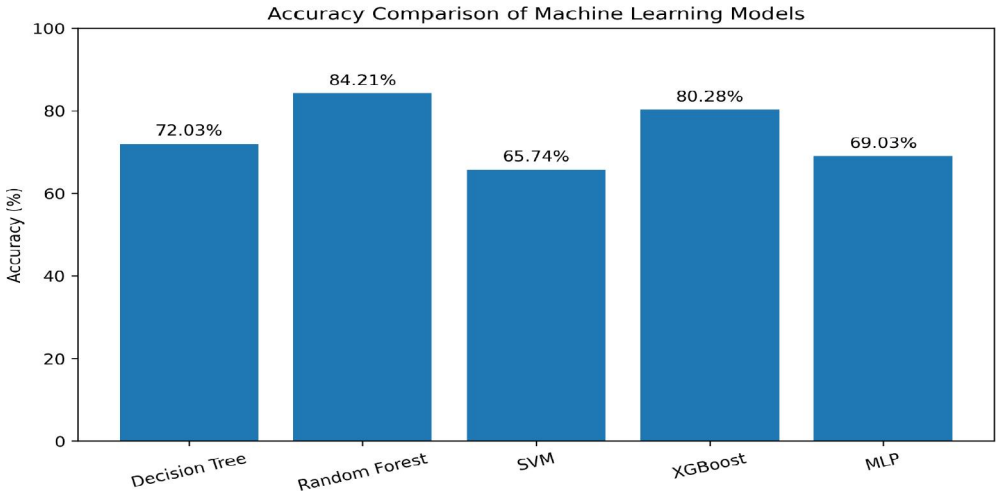
Model Evaluation

The models were compared using Accuracy, Precision, Recall, F1-score and ROC-AUC. These measures were selected because electricity theft is the minority class and accuracy alone may not provide enough information about theft detection performance.

Accuracy represents the overall percentage of correct predictions. Precision shows how many of the consumers predicted as theft were actually theft cases. Recall shows how many of the actual theft consumers were successfully detected by the model. F1-score provides a balance between Precision and Recall. ROC-AUC shows how well the model can distinguish between normal and theft consumers.

Confusion matrices were also used to examine correct and incorrect classifications and to understand false-positive and false-negative predictions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2



Experimental Results

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	72.03	17.20	58.51	26.46	72.00
Random Forest	84.21	24.08	40.66	29.56	74.70
SVM	65.74	14.97	64.45	24.30	71.88
XG Boost	80.28	22.57	52.28	31.10	77.80
MLP	69.03	15.64	63.07	24.43	74.70

Random Forest gives the highest reported Accuracy and Precision among the five models. SVM gives the highest Recall, which means it identifies a larger proportion of the actual theft records in the reported results. XGBoost gives the highest F1-score and ROC-AUC, showing a stronger balance between Precision and Recall and stronger class discrimination according to these measures

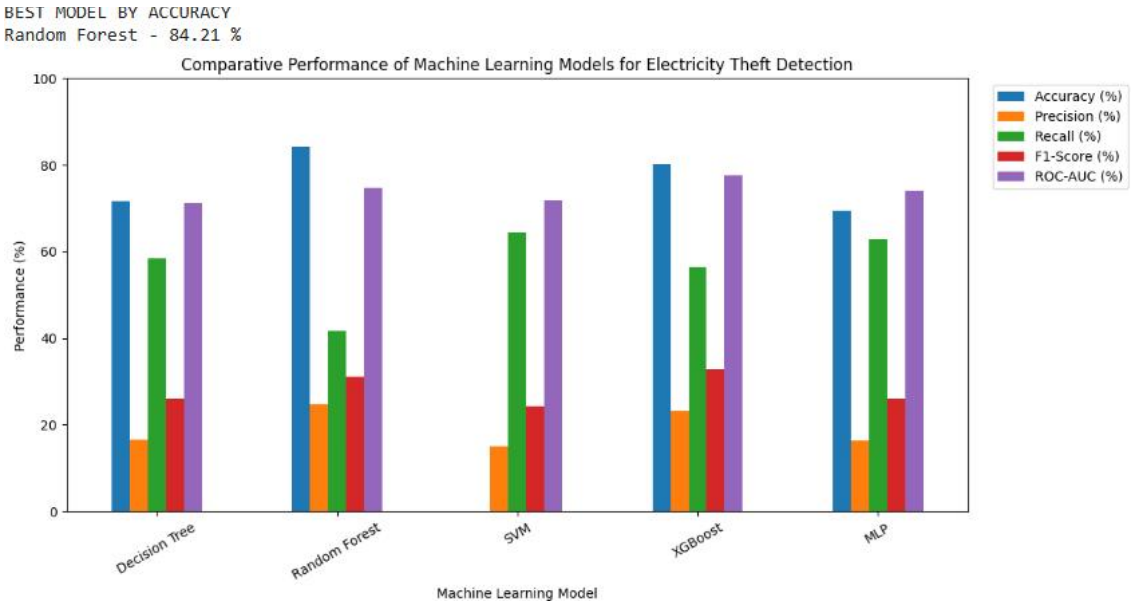
Feature Importance Analysis

After comparing the models, feature importance was analysed using the XG Boost model to understand which behavioural features were useful for the prediction .The analysis shows that Standard Deviation, Zero Percentage, Consumption Range, Missing Percentage, Zero Consumption Days and Missing Days are among the more influential features in the model .This indicates that the amount of variation in electricity consumption, zero-consumption behaviour and missing-consumption behaviour provide useful information for distinguishing the classes.

Feature importance represents the usefulness of the variables to the trained model and does not by itself prove that these behaviours directly cause electricity theft.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40733

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 2, November 2026 Impact Factor: 8.2



Practical Prototype

An interactive prototype can be developed to demonstrate the basic idea of electricity consumption screening. The prototype can accept consumer number, average monthly consumption, previous month consumption and current month consumption as input.

The system can then display a simple consumption status such as potentially suspicious, average consumption or normal consumption. For example, if a consumer normally uses around 300 units and the current consumption is considerably lower, the system can highlight the pattern for further investigation.

This prototype is only a demonstration of the research concept. It should not be treated as a fixed rule for confirming electricity theft because real electricity theft detection requires detailed analysis and investigation

LIMITATIONS

This research has certain limitations. The study is based on the publicly available SGCC electricity consumption dataset, and the results may not fully represent electricity consumption patterns in different regions or electricity distribution systems.

The dataset contains a considerable class imbalance between normal and theft consumers. Although SMOTE was used during training, the models can still produce false-positive and false-negative predictions.

The research uses behavioural features created from the available consumption information. Other factors such as geographical information, consumer type, meter condition, connection information and external electricity-system information are not considered.

The developed system is also not directly connected to real-time electricity meters. Therefore, the current work should be considered a research and experimental system rather than a complete real-world electricity theft monitoring system.

Finally, a Machine Learning prediction only indicates a suspicious pattern. It should not be considered final proof of electricity theft without further investigation.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40733

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, November 2026
Impact Factor: 8.2

II. CONCLUSION

In this study, consumer electricity consumption behaviour was used to identify possible electricity theft using Machine Learning. The original electricity consumption dataset contained a large number of daily readings, so behavioural feature engineering was used to create 16 meaningful features representing consumer electricity usage.

The study compared five Machine Learning models: Decision Tree, Random Forest, SVM, XGBoost and MLP. SMOTE was applied to the training data to handle the imbalance between normal and theft consumers. The results show that the models have different strengths. Random Forest achieved the highest reported Accuracy, while SVM achieved the highest Recall.

XGBoost achieved the highest F1-score and ROC-AUC in the reported results.

REFERENCES

[1] R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," Applied Energy, vol. 238, pp. 481–494, 2019. doi: 10.1016/j.apenergy.2019.01.076.

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

[3] S. Y. Appiah, E. K. Akowuah, V. C. Ikpo, and A. Dede, "Extremely randomised trees machine learning model for electricity theft detection," Machine Learning with Applications, vol. 12, p. 100458, 2023. doi: 10.1016/j.mlwa.2023.100458.

[4] H. Iftikhar, N. Khan, M. A. Raza, G. Abbas, M. Khan, M. Aoudia, E. Touti, and A. Emara, "Electricity theft detection in smart grid using machine learning," Frontiers in Energy Research, vol. 12, 2024, article 1383090. doi: 10.3389/fenrg.2024.1383090.

[5] W. Zhang and Y. Dai, "A multiscale electricity theft detection model based on feature engineering," Big Data Research, vol. 36, p. 100457, 2024. doi: 10.1016/j.bdr.2024.100457.

[6] F. Alfaverh, H. Gan, V. Miroshnyk, Z. Bin Saeed, I. Blinov, P. Shymaniuk, P. Tarassodi, and I. Mporas, "Electricity Theft Detection from Electricity, Gas and Water Measurements using Machine Learning," Energies, vol. 19, no. 9, p. 2045, 2026.