Synthetic Minority Oversampling Technique Enhanced Machine Learning Models for Energy Theft Detection

Md Saiful Islam Sajol[1], Imtiaz Ahmed[2], Quazi Sanjid Mahmud[3]
[1] Louisiana State University, [2] New Mexico Institute of Mining and Technology, [3] University of California Riverside.
Email: [1] msajol1@lsu.edu, [2] imtiaz.ahmed@student.nmt.edu, [3] qmahm002@ucr.edu

***Abstract*—Electricity theft poses significant challenges to utility companies worldwide, resulting in substantial financial losses. This study addresses the problem by leveraging machine learning algorithms to detect energy theft in smart grids. The insufficiency of data on theft conditions and the imbalance of datasets have always hindered the precise identification of fraudulent activity. To mitigate these challenges, we curated a dataset from the Open Energy Data Initiative, which encompasses sixteen consumer categories and six theft conditions. Our approach focuses on using the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance by generating synthetic samples for minority classes. We conducted a comparative analysis of various machine learning based classification algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest (RF), Bagging with RF, and Ensemble Learning, and observed the results before and after the implementation of SMOTE on the dataset. We find that SMOTE demonstrates its most significant impact on classifying the most challenging classes within the dataset. In particular, it shows improvements of 57.00%, 37.88%, and 36.88,% for Class 6, Class 1, and Class 3, respectively, with the KNN algorithm. Other algorithms also indicate significant increments in terms of accuracy, kappa, F1-score, and AUC metrics in detecting fraudulent activity. Overall, this research contributes to advancing energy security by highlighting the importance of robust theft detection frameworks for safeguarding energy distribution systems.**

***Index Terms*—Electricity Theft, Machine Learning, Smart Grids, Ensemble, Algorithm, Class Imbalance**

I. Introduction

The widespread issue of electricity theft, characterized by unauthorized consumption bypassing payment, imposes substantial financial burdens on utility companies globally, with estimated annual losses surpassing $96 billion [1]. The advent of smart grids and Advanced Metering Infrastructure (AMI) presents an opportunity to utilize machine learning (ML) techniques for analyzing smart meter data to detect anomalies indicative of electricity theft [2].

Deep learning models and optimization techniques have shown promising results in analyzing power systems [3] and renewable energy [4]. So involving deep learning in Energy theft detection can help in detecting anomalies and patterns that traditional methods might miss. It begins with the preprocessing and feature engineering of smart meter data to extract attributes distinguishing standard from abnormal usage patterns. To mitigate the disparity in data distribution between legitimate and fraudulent instances, various sampling strategies are employed. Key contributions of this research include aggregating a dataset from publicly available smart meter readings representing diverse consumer categories, employing tailored data preprocessing, feature engineering, and sampling techniques for electricity theft detection, and assessing both machine learning and ensemble algorithms for classifying electricity theft across multiple categories. Model performance is comprehensively evaluated using rigorous metrics such as accuracy and F1-score.

The remainder of the paper is organized as follows: Section 2 provides a literature overview on ML-based electricity theft detection. Section 3 details the dataset, preprocessing steps, and methodologies employed. Section 4 analyzes model performance across various metrics. The paper concludes in Section 5, presenting conclusions and avenues for future research.

The rise of smart grids marks a transformative shift in the energy sector toward greater efficiency and reliability. However, this digital evolution also brings challenges such as energy theft, leading to significant economic losses and grid inefficiencies [5]. Traditional theft detection approaches often fall short of addressing the complex data patterns and anomalies inherent in smart grid systems. This study aims to overcome these challenges by harnessing ML techniques to enhance the accuracy of theft detection and fortify system integrity. Notably, our investigation reveals that integrating the K-Nearest Neighbors algorithm with the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance demonstrates exceptional proficiency in identifying abnormal consumption behaviors indicative of energy theft [6].

II. Background Study

The integration of Machine Learning (ML) techniques in electricity theft detection within smart grid systems addresses the limitations of traditional methods while leveraging advanced data analytics [7] [8]. Smart meters within the Smart Grid framework enable precise monitoring and management of energy consumption, enhancing grid efficiency and resilience [9] [10]. However, the proliferation of smart metering systems has amplified the challenge of energy theft, resulting in substantial financial losses and compromising grid safety and reliability [11][12].

To address these challenges, there is a growing emphasis on leveraging ML techniques for more effective theft detection. Traditional methods, such as statistical techniques [13] and conventional data analysis algorithms, have shown limitations in adapting to the dynamic and complex nature of smart grid data, often leading to low accuracy in theft detection [14]. In contrast, Supervised [15] and Unsupervised ML algorithms, including extreme learning machines (ELM), and deep learning approaches like Convolutional Neural Networks (CNN) [16], Multimodel Transformers [17] and Long Short-Term Memory (LSTM) networks [18], have demonstrated superior performance by effectively capturing and analyzing intricate patterns associated with energy theft [19].

The application of CNNs and RNNs extends beyond traditional uses, showcasing effectiveness in analyzing time-series data and pattern recognition within energy systems, highlighting their adaptability and potential for energy-efficient computing [20] [21]. Similarly, Random Forest (RF) algorithms, particularly when combined with Synthetic Minority Over-sampling Technique (SMOTE), have proven effective in handling imbalanced datasets, a common challenge in theft detection, by improving classification performance through the generation of synthetic samples in the minority class [22] [23].

Despite these advancements, challenges persist, including the reliance on large, labeled datasets and substantial computational resources for deep learning models [24], and variability in the effectiveness of techniques like SMOTE based on dataset characteristics [25]. Other approaches, such as Stacked Machine and Deep Learning models, Adaptive Boosting, and Adaptive Random Forest, offer unique advantages but are also subject to limitations related to feature extraction, hyperparameter optimization, and the choice of optimization algorithms [26].

In conclusion, the transition towards ML-based theft detection in smart grids represents a significant advancement in addressing energy theft challenges. While each ML technique has unique strengths and limitations, careful algorithm selection should consider dataset characteristics and desired performance metrics. Future research should focus on overcoming existing barriers, exploring hybrid model fusion, and enhancing the adaptability and efficacy of ML techniques in practical scenarios.

III. Methodology

A. Dataset Description

The dataset utilized in this study was obtained from the Open Energy Data Initiative (OEDI) platform, a renowned repository of energy research datasets sourced from various programs, offices, and national laboratories under the U.S. Department of Energy [27]. This dataset encompasses comprehensive energy consumption data across diverse consumer types, collected over a year at hourly intervals.

1) Origin and Composition: The dataset, originally sourced from OEDI, has been refined to facilitate machine learning-based detection of fraudulent activities within smart grids[28].

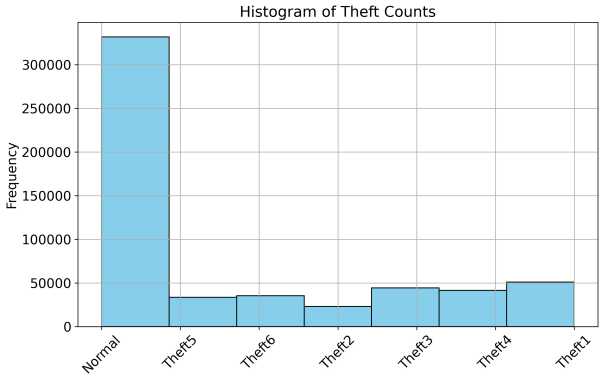

Fig. 1. Histogram of normal condition and different types of theft count. As can be seen from the figure, compared to the number of normal instances, the number of different types of theft instances is significantly low

It includes six distinct fraud scenarios mirroring different methods of electricity theft and covers sixteen consumer categories, representing various establishments such as restaurants, warehouses, hospitals, and schools.

2) Feature Representation: The dataset is feature-rich, capturing nuanced aspects of energy consumption patterns through numerical and categorical attributes. Features include electricity consumption metrics across facility components like cooling, heating, lighting, and equipment.The dataset contains a total of 560,640 instances distributed across 12 columns. Among these columns, 11 represent distinct features, with 10 being numerical and one categorical. The dataset displays 16 distinct consumer types, such as restaurants, hospitals, hotels, offices, schools, retail outlets etc., with an equal distribution of instances per consumer type, with each type containing 35,040 instances.

3) Fraud Detection Context: Fraud-related scenarios, labeled as theft1 through theft6, are integrated into the dataset to facilitate advanced anomaly detection and fraud mitigation strategies within smart grid environments. In summary, Researchers can leverage this dataset to uncover insights and develop robust frameworks for combating fraudulent activities within energy distribution networks.

B. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling approach in which the minority class is oversampled by creating "synthetic" examples rather than by oversampling with replacement [28] . It can be used for imbalanced datasets due to several reasons :

1) Addressing Class Imbalance: In imbalanced datasets, where one class (majority class) significantly outweighs the other (minority class), biased models favoring the majority class may result. SMOTE resolves this issue by generating synthetic samples for the minority class, balancing the distribution, and enhancing model performance [28].

2) Improved Generalization: By creating synthetic samples via interpolation among existing minority class instances,

TABLE I
DIFFERENT ALGORITHMS FOR KNOWN CONSUMER 7 CLASSES

	KNN			DT			RF			Bagging			Ensemble Learning		
	Before	After	Change	Before	After	Change	Before	After	Change	Before	After	Change	Before	After	Change
Class 0	93.12	100	6.88	97.03	95.08	-1.95	94.23	94.03	-0.2	93.87	94.11	0.24	93.78	95.69	1.91
Class 1	60.23	98.11	37.88	74.21	92.19	17.98	75.16	95.80	20.64	72.10	94.86	22.76	74.76	97.84	23.08
Class 2	100.00	100.0	0	100	100	0	100	100	0.00	100	100	0	100	100	0
Class 3	61.24	98.12	36.88	69.09	91.05	21.96	78.01	92.02	14.01	75.89	90.58	14.69	74.93	95.76	20.83
Class 4	74.35	98.65	24.3	88.13	96.11	7.98	87.31	96.97	9.66	84.78	94.68	9.90	86.82	98.65	11.83
Class 5	88.43	98.75	10.32	96.80	100.00	3.2	97.11	100.00	2.89	95.21	100.00	4.79	96.45	99.68	3.23
Class 6	38.23	95.23	57.00	55.24	88.89	33.65	28.19	91.79	63.6	40.16	92.67	52.51	46.23	94.86	48.63

TABLE II
ACC, F1, AUC FOR FOR KNOWN CONSUMER 7 CLASSES

	Accuracy			F1 Score			Kappa			AUC		
	Before	After	chnage	Before	After	chnage	Before	After	chnage	Before	After	chnage
KNN	84.85	97.68	12.83	82.78	97.68	14.9	74.57	97.79	23.22	.90.87	.98.79	7.92
DT	89.40	94.69	5.29	89.41	94.68	5.27	82.96	93.80	10.84	90.59	96.90	6.31
RF	87.90	95.75	7.85	87.05	95.72	8.67	80.23	95.04	14.81	96.72	99.73	3.01
Bagging	88.67	94.78	6.11	86.78	94.78	8.00	81.69	94.88	13.19	95.45	99.85	3.4
Ensemble	89.21	97.7	8.49	88.14	97.70	9.56	82.23	97.32	15.09	97.30	99.94	2.64

SMOTE increases dataset diversity, fostering better model generalization and reducing the risk of overfitting to the minority class [28].

3) Preservation of Information: SMOTE generates synthetic samples based on feature space, focusing on feature relationships to preserve underlying dataset information while balancing class distribution [28].

4) Enhanced Performance: Studies demonstrate that SMOTE significantly boosts classifier performance on imbalanced datasets, making it invaluable for handling such challenges [28].

5) Versatility: SMOTE's success spans various domains and has inspired the development of new approaches for tackling class imbalance. Its simplicity, robustness, and effectiveness have made it a popular choice among researchers and practitioners working with imbalanced datasets[28].

In summary, SMOTE is favored for addressing class imbalance, improving generalization, enhancing performance, and offering versatility in handling data challenges.

IV. EXPERIMENTS AND RESULTS

All the experiments are done using sklearn packages of Python programming language. TABLE I compares the classification accuracy using different algorithms such as the K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Bagging, RF, and Ensemble Learning method before and after the application of SMOTE. We noted that before applying SMOTE, all the algorithms struggle to detect Class 1, 3, and 6 to classify properly. Class 6 is the most challenging scenerio, followed by Class 1 and Class 3. After using SMOTE we observe drastic improvement of accuracies of for Class 6 across all the methods such as 57.00% for KNN, 63.00% for RF, and so on. Overall, the performance improves significantly for all of the given classes except Class 0 which is the Normal Condition, and the model was already capable of detecting it properly both in pre and post SMOTE condition.

We found that KNN achieves the biggest improvement 57.00% to identify for class 6, followed by 37.88% for Class 1 and 36.88% for Class 3. We also observe 63% performance improvement for Class 6 with the RF algorithm. For Class 6 we see a rise of 33.65%, 52.51%, and 48.63% improvement with DT, Bagging, and Ensemble Learning methods respectively.

TABLE II shows the overall performance for the given ML models based on accuracy, F1 score, kappa, and AUC considering all the classes. Though the accuracy of KNN and Ensemble is the highest and approximately close, KNN outperforms other models, as it takes the least amount of time to run and has the highest performance improvement after using SMOTE. Besides, all the performance of the models shows an accuracy of well over 90% after making the data balanced by SMOTE.

We observe that the F1 scores is also improved, ranging from 5.27% to 14.9% after the utilization of SMOTE. The Kappa score quantifies the agreement between predicted and actual class labels, where higher scores indicate a better agreement. The AUC is calculated for each class against all other classes combined. AUC represents the capacity of the classifier to discriminate between classes. Both Kappa and AUC demonstrate a substantial increase in performance after applying SMOTE, which again proves the enhanced model performance in correctly predicting class labels.

V. CONCLUSION & FUTURE WORK

This research underscores the efficacy of machine learning methods, particularly the dataset augmented with SMOTE, in elevating electricity theft detection within smart grids. Our approach to use SMOTE to make the dataset imbalance to balance has demonstrated remarkable proficiency in identifying theft through after sophisticated data preprocessing and algorithm selection, presenting a significant protective measure for utility companies and consumers alike. Looking ahead,

the field faces challenges such as the dependency on large labeled datasets and computational resources, and the variable effectiveness of techniques like SMOTE. Future research should navigate these hurdles by exploring hybrid models that combine the strengths of various machine learning techniques. This pursuit aims to enhance the adaptability and efficiency of these models for real-world application, ensuring that smart grid systems remain resilient against electricity theft.

REFERENCES

[1] A. Kawoosa. "Using machine learning ensemble method for detection of energy theft in smart meters". In: *Iet Generation Transmission and Distribution* 17 (21 2023), pp. 4794–4809.

[2] X. Fang et al. "Smart grid — the new and improved power grid: a survey". In: *Ieee Communications Surveys and Tutorials* 14 (4 2012), pp. 944–980.

[3] Mahdi Khodayar et al. "Deep learning in power systems research: A review". In: *CSEE Journal of Power and Energy Systems* 7.2 (2020), pp. 209–220.

[4] A S M Jahid Hasan et al. "Electricity Cost Optimization for Large Loads through Energy Storage and Renewable Energy". In: *2023 International Conference on Information and Communication Technology for Sustainable Development*. 2023, pp. 46–50.

[5] L. Lepolesa, S. Achari, and L. Cheng. "Electricity theft detection in smart grids based on deep neural network". In: *Ieee Access* 10 (2022), pp. 39638–39655.

[6] Jason Brownlee. *SMOTE for Imbalanced Classification with Python*. 2021. URL: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[7] G. Rausser, W. Strielkowski, and D. Štreimikienė. "Smart meters and household electricity consumption: a case study in ireland". In: *Energy and Environment* 29 (1 2017), pp. 131–146.

[8] S. Kim et al. "Security issues on smart grid and blockchain-based secure smart energy management system". In: *Matec Web of Conferences* 260 (2019), p. 01001.

[9] Y. Zhao, L. Zheng, and M. Zhao. "Design of energy consumption simulation model and development of support platform based on energy supply system". In: *Journal of Physics Conference Series* 2229 (1 2022), p. 012026.

[10] I. Machorro-Cano et al. "Hems-iot: a big data and machine learning-based smart home system for energy saving". In: *Energies* 13 (5 2020), p. 1097.

[11] R. Jiang et al. "Energy-theft detection issues for advanced metering infrastructure in smart grid". In: *Tsinghua Science and Technology* 19 (2 2014), pp. 105–120.

[12] J. Khan, F. Siddiqui, and R. Khan. "Survey: ntl detection in electricity energy supply". In: *International Journal of Computer Applications* 155 (9 2016), pp. 18–23.

[13] A S M Jahid Hasan et al. "Data Driven Energy Theft Localization in a Distribution Network". In: *International Conference on Information and Communication Technology for Sustainable Development*. 2023, pp. 388–392.

[14] J. Pitchaimani et al. "Smart grid security enhancement by detection and classification of non-technical losses employing deep learning algorithm". In: *International Transactions on Electrical Energy Systems* 30 (9 2020).

[15] Faisal Bin Ashraf, SM Maksudul Alam, and Shahriar M Sakib. "Enhancing breast cancer classification via histopathological image analysis: Leveraging self-supervised contrastive learning and transfer learning". In: *Heliyon* 10.2 (2024).

[16] M. Alazab et al. "A multidirectional lstm model for predicting the stability of a smart grid". In: *Ieee Access* 8 (2020), pp. 85454–85463.

[17] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. "Multimodal transformer for material segmentation". In: *arXiv preprint arXiv:2309.04001* (2023).

[18] M. Hasan et al. "Electricity theft detection in smart grid systems: a cnn-lstm based approach". In: *Energies* 12 (17 2019), p. 3310.

[19] S. Bhattacharya et al. "A novel pca-firefly based xgboost classification model for intrusion detection in networks using gpu". In: *Electronics* 9 (2 2020), p. 219.

[20] A. Yona et al. "Decision technique of solar radiation prediction applying recurrent neural network for short-term ahead power output of photovoltaic system". In: *Smart Grid and Renewable Energy* 04 (06 2013), pp. 32–38.

[21] A. Monteiro et al. "Embedded application of convolutional neural networks on raspberry pi for shm". In: *Electronics Letters* 54 (11 2018), pp. 680–682.

[22] C. Velasco et al. "Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data". In: *International Journal of Emerging Technologies in Learning (Ijet)* 18 (04 2023), pp. 120–155.

[23] L. Ma and S. Fan. "Cure-smote algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests". In: *BMC Bioinformatics* 18 (1 2017).

[24] M. Hasan et al. "Electricity theft detection in smart grid systems: a cnn-lstm based approach". In: *Energies* 12 (17 2019), p. 3310.

[25] A. Kawoosa. "Using machine learning ensemble method for detection of energy theft in smart meters". In: *Iet Generation Transmission and Distribution* 17 (21 2023), pp. 4794–4809.

[26] I. Khan et al. "A stacked machine and deep learning-based approach for analysing electricity theft in smart grids". In: *Ieee Transactions on Smart Grid* 13 (2 2022), pp. 1633–1644.

[27] Salah Zidi et al. “Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment”. In: *Journal of King Saud University - Computer and Information Sciences* 35.1 (2023), pp. 13–25. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2022.05.007. URL: https://www.sciencedirect.com/science/article/pii/S1319157822001562.

[28] Alberto Fernández et al. “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary”. In: *Journal of artificial intelligence research* 61 (2018), pp. 863–905.