Deceptive Maneuvers: Subverting CNN-AdaBoost Model for Energy Theft Detection

Santosh Nirmal, Pramod Patil

***Abstract*— As deep learning models become more prevalent in smart grid systems, ensuring their accuracy in tasks like identifying abnormal customer behavior is increasingly important. As its use is increased in smart grids to detect energy theft, crafting adversarial data by attackers to deceive the model to get the desired output is also increased. Evasion attacks (EA) attempt to evade detection by misclassifying input data during testing. The manipulation of data inputs is done so that it is not noticeable to humans but can cause the machine learning (ML) model to produce incorrect results. Electricity theft has become a major problem for utility companies that need to be dealt with effectively. Convolutional Neural Network (CNN) and AdaBoost hybrid model have been developed that promise to detect electricity theft with high accuracy. However, this model is also vulnerable to evasion attacks that can render it ineffective.**

In this paper, to make the detection system more robust, we present a generative method to create evasion attacks against a hybrid model combining Convolutional Neural Network and Adaboost (CNN-Adaboost). Generated adversarial data from the proposed algorithm is crafted on the model to test its performance. Our proposed attack is validated with State Grid Corporation of China (SGCC) dataset. We test the CNN-Adaboost energy theft detection model and other models' performance under 5% and 10% evasion attacks. Our findings reveal model performance degradation under our proposed generative evasion attack ranging from 96.35% to 89.23%. With the defence mechanism, we successfully increased adversarial accuracy by up to 97% and decreased the attack success rate (ASR) by up to 3%. These adversaries are useful for designing robust and secure machine learning models, offering an improved solution compared to previous work in this area. We tested the model with varying percentages of adversarial data to analyze its behavior effectively. These adversaries are useful for designing robust and secure ML models. The proposed attack and defence can be utilized to test energy theft detection (ETD) models in industrial and commercial settings.

***Index Terms*—Adversarial examples, energy theft detection, evasion attack, smart grids.**

Original Research Paper
DOI: 10.53314/ELS2428046N

Manuscript received on June 5th, 2024. Received in revised form on November 29th, 2024. Accepted for publication on December 9th, 2024. Accepted for publication 09 December 2024.

Santosh Nirmal is with CSIR- Unit for Research and Development of Information Product and Dr.D.Y.Patil Institute of Technology Maharashtra, India

Pramod Patil is with Dr.D.Y.Patil Institute of Technology, Pune, Maharashtra ,India (nirmalsantosh@gmail.com, pdpatiljune@gmail.com)

I. Introduction

Machine learning techniques are used in several applications, mainly image identification [1], face recognition [2], energy theft detection [3], banking fraud detection [4,5], and natural language processing. [6,7] etc. Though it is an emerging technology that is capturing every aspect of human life, it has its own drawbacks. It can be easily malfunctioned by attackers who can craft adversarial examples and get the desired result, which may pay a huge cost. Especially in auto-driving cars kind of applications, which may be a threat to human life.

Due to the increase of machine learning techniques in energy sector, lots of developments in this sector have been seen over the decade, several advanced methods have been deployed to collect the data, to process, to manage the data and which results in get cost effective, reliable power system [8]. Though there are lot of benefits to the use of machine learning in SMRAT gird systems, there are significant challenges also like attacks on the system including poising attacks, evasion attacks [9], etc.

There are different kinds of attack based on the outcome, knowledge and stage. For outcome category attack the goal is to mislead the classifier to predict. These are target and non-target-based attacks.

There are different types of attacks on ML depending on the goal of the attacker like espionage, sabotage or fraud and at which stage it is executed. The attacks performed on machine learning in their training stage are known as poising, and the attacks performed in the production phase are known as evasion attacks, which is also known as false negative evasion. In the poising approach, logic corruption, data manipulation, and data injection can be used to force the model to classify the data wrongly.

Statistical span filters are useful and effectively work on span emails; in some of the earliest studies like [10], Daniel Lowd et al. used good word attacks on these filters. This is basically known as data poisoning, a major problem in machine learning. As research is done in this field, new techniques are found to overcome these problems, and different attacking methods have also arisen in the last 15-20 years.

A vast research has been carried out and it is found that DL models are also prone to data poisoning attacks such as data manipulation training [11] and how data poising techniques like black gradient optimization [12] are used across different applications and corrupted the algorithms' logic which results in performance degradation in models. There is no proper literature on attacks on tabular data like bank or electricity consumption

data; whatever the literature is, it mainly focuses on images and text domains. Adversarial examples are widely used in ML to test the models' robustness. In smart grid, there are several studies have been conducted in which adversarial attacks are proposed and studied which result in getting the limitation of the ETD models and helps them to overcome the limitation which ultimately increase its robustness, accuracy and the revenue of the utility companies.

The following is the paper's organization. An overview of related work is discussed in section II; section III explains the proposed algorithm; section IV for experimental setup and a conclusion is given in section V.

II. Related work

Finding and preventing the theft of electricity is becoming a significant challenge, and most researchers are exploring new methods and innovations to overcome the threat of attack and make the machine learning model more and more robust, which ultimately benefits society by saving electricity. This section explains the related work which had been carried out for the last few years; it includes different ETD techniques and adversarial attacks on different ML models.

Several studies have shown that deep learning models are highly vulnerable to adversarial examples in the computer vision domain [13]-[17]. Deep learning models can be deceived by adversarial attackers who insert well-crafted perturbations into legitimate inputs, leading to wrong predictions. In [18]-[20], attacks on machine learning models used in power systems are well demonstrated.

The role of smart grid system plays a major role in various sectors, and as the popularity of deep learning (DL) use in smart grids increased, the compromise of the model for personal benefit by attackers is also increased. In [3], authors studied energy ETD and proposed a CNN-Adaboost model, which shows markable accuracy in the prediction, but still, it is vulnerable and needs to test its performance under attack. Bin Sulaiman et al. [5] analyzed multiple ML techniques for credit card fraud detection and compared their accuracy and data confidentiality implications and also proposed a hybrid model for ETD and enhancing privacy. In [6] author studied various phases of natural language processing (NLP) and spam filters. Kotyan et al. [21] studied various attacks on machine learning models and different defense techniques and their limitations in detail. In [22], author Ashrapov et al. have done a details study on uneven data distribution and different generative adversarial network (GAN) techniques. The author [23] studied the impact of evasion attacks on ETD and proposed evasion attack, and claimed the performance of the tested models had decreased. In [24] a new adversarial nets framework with two generative and discriminative models. In their article [25], Su Wang demonstrates the versatility and potential of using GAN in Natural Language Processing as a neural network architecture.

Ballet et al. [26] claim that adversarial examples were first introduced in the tabular domain, providing the feature importance, and based on that it is decided how much percentage of perturbation can be applied to different features, authors also claim that more perturbations are required for less important features, features change with the help of perturbation, the impact of features changes on models, were also studied in [27] and [28]. These changes are made to obtain the expected output.

In [29], J. Chan et al. proposed a novel algorithm called HopSkipJumpAttack, which estimates the gradient direction using binary information at the decision boundary, the author also claims that it requires fewer model queries than boundary attack. Liu, Yugeng, et al in [30] proposed and also claimed that the first backdoor attack against the machine learning model via a malicious dataset distillation service provider. Author Atef Bondok et al. raise the privacy concern [31] and propose a federated learning (FL) approach having a global and consumer model to investigate the effectiveness of adversarial training (AT); three different attacks are introduced by authors in this paper, namely Distillation, No-Adversarial-Sample-Training, and False-Labeling. The impact of EA on Household energy forecasting [32] or state estimation [33] models have been studied in details by authors. These studies have concluded that EAs can threaten the accuracy of machine learning-based detectors. Vincent Ballet et al. [34] proposed an algorithm to generate imperceptible adversarial examples in the tabular domain; our work is also referred to [34] for creating adversarial examples. In [35], Cartella et al. studied different adversarial algorithms used for tabular data, image classification and compared their performance. Author also introduced constraints in the perturbation. Singh et al. [36] generated adversarial examples using white-box attacks for residential houses. Two novel attacks on Deep Reinforcement Learning (DRL) agents, namely critical point and antagonist attacks are proposed [37] by Sun, Jianwen, et al. in which the authors inject adversarial examples in order to perform the damage to the agent. The Authors also perform comprehensive evaluations on their effectiveness on different algorithms. Author [38] conducted a study to measure the impact of small perturbations on GRU-RNN models. They then performed adversarial training to increase the robustness against these attacks. The researcher [39] utilized a hybrid NN to detect energy theft attacks in the distributed generation domain.

Nowroozi, Ehsan, et al [40] performed eight different evasion attacks on DL models and proposed a novel strategy based on feature randomization techniques to resist these attacks. The authors also focused on preventing adversarial transferability to make the target network more secure.

The primary distinction between our research and the contributions mentioned in this section lies in the approach. While studying previous work in this area, it was observed that the ETD model's behavior was not tested with different percentages of adversarial data, which is crucial for analysing behavior patterns and determining when performance improves or degrades. Additionally, some studies have utilized surrogate models for black-box attacks, while only a few have proposed defense mechanisms to safeguard models against adversarial attacks. In contrast, we presented an evasion attack by generating the adversarial data using the algorithm. This algorithm can

be used for tabular data of other domains, and we also proposed a defense system against the attack aimed at enhancing the robustness of the model. This ultimately ensures system efficiency and reliability while promoting fair access to electricity.

III. Proposed Evasion Attack

This section discusses the ETD model and the adversarial algorithm that generates the adversarial data.

This study uses the SGCC electricity consumption (EC) dataset to train and test the model. It contains data for 34 months, with 40,256 consumers, of which 3,579 are fraudulent consumers.

Fig. 1 below shows how the electricity consumption (EC) pattern forms for legitimate and fraudulent consumers. When there is low consumption for a long time for a particular consumer, it comes under surveillance because of the low consumption. There are several reasons that should be taken into account before considering it as a fraud consumer. The house might be closed for a long time, the meter might be faulty, etc. It's important to keep in mind that a customer with lower-than-normal consumption doesn't always mean they are a fraudulent user. Therefore, it's crucial to analyze the data thoroughly before making any assumptions or taking action.

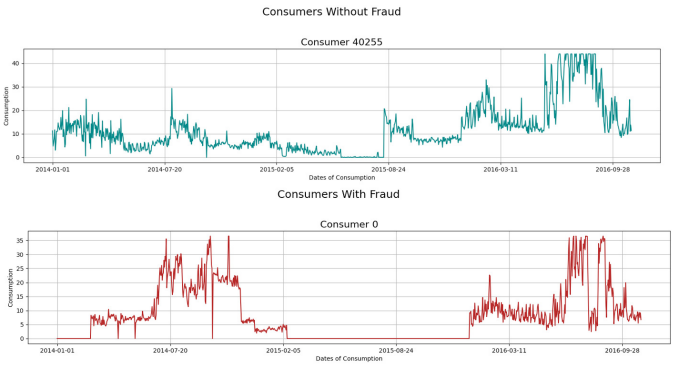


Fig. 1. EC pattern of fraud and legitimate user

A. Data Acceptance Conditions

There are several conditions that need to be considered before sending the data to the model. Power consumption increases and decreases at different time intervals; it increases in the morning and decreases at midnight. The consumption is higher on weekends, and it is less on working days for residential consumers, but for industrial and commercial consumers it is more on weekdays and less on the weekends due to the holiday. A database required for Algorithm 1 is made with consumers which are treated as legitimate consumers by the ETD model. One year of consumer data was considered for generating the adversarial dataset. The parameters like sudden changes in the consumption, gradual changes are analyzed and we separated them and other meter data in which there is no markable hike or drop in consumption is considered for the algorithms.

Table I
Number of Meters and their Variations in Energy Uses in Different Conditions

Months	Faulty Meter	Festival	Seasonal changes	Outstation	No Remarkable Changes
1-3	310	713	502	105	4312
4-6	201	630	237	72	3870
7-9	193	426	182	360	4828
10-12	227	900	1526	470	3250

In Table I, it is found that those consumers having a markable change in their consumption when tested with the ETD model, they may be treated as fraud consumers, especially those who are out of the station, as there has been no consumption of those meters for a long time.

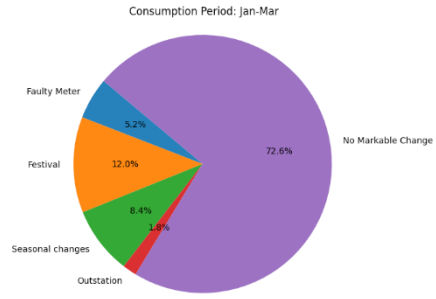


Fig. 2. EC for Jan-Mar

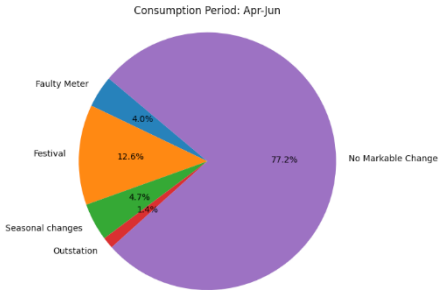


Fig. 3. EC for Apr-Jun

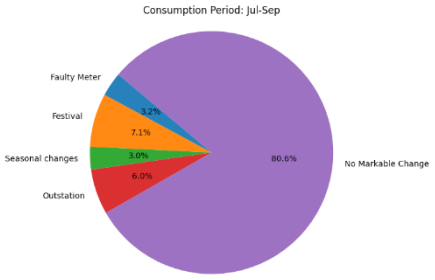


Fig. 4. EC for Jul-Sep

Figs. 2, 3, 4, and 5 show the quarterly electricity consumption; From those Figures and Table I show that there are variations in electricity consumption throughout the year. The purpose of these categories is to get the new dataset, which is stable and further used to generate adversarial examples in such a way that it remains close to the actual reading and easily treated be-

nign samples by the ETD algorithm. So now the dataset is generated from no markable changes column.

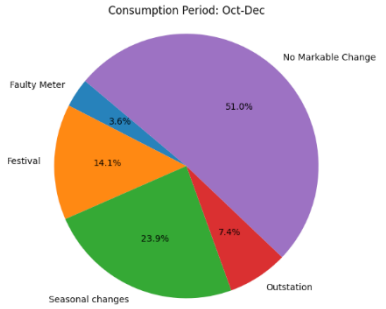


Fig. 5. EC for Oct-Dec

Fig. 6 shows the evasion attack on CNN and AdaBoost hybrid model. In the proposed model, CNN is used, and its hyperparameters, including activation functions (AF), are represented in Table II.

TABLE II
HYPERPARAMETERS OF CNN

Layer	Units	AF
Input	48	Linear
Conv1D	64	Relu
Conv1D	128	Relu
Dense	128	Relu
Dense	64	Relu
Dense	64	Relu
Dense	2	Softmax

We compare the model's performance before and after the attack. It is a hybrid model in which CNN is used to extract the features from the processed data, and AdaBoost is used to perform the classification of real and fraud consumers.

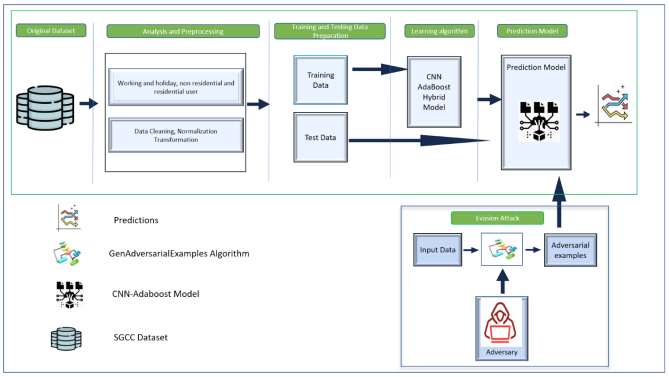


Fig. 6. CNN-Adaboost ETD under Evasion attack

B. Data Preprocessing

The ETD model in Fig. 2 performs well on SGCC dataset and gives 96% accuracy. but it is not tested with adversarial data.

The SGCC dataset contains both normal and malicious consumers. Before transferring the data to the training phase, it is necessary to preprocess it. Several steps have been carried out, including data cleaning, which involves imputation, removal of duplicates, normalization, and outlier handling. Special cases such as holidays and faulty meters have also been considered.

An adversarial attack is proposed in this paper to test the model's performance. We generate adversarial data using GenAdversarialExamples algorithm and test the accuracy of the model by crafting this adversarial data during the testing phase.

C. An Algorithm to Generate Adversarial Examples

The adversarial algorithm is used by an attacker to manipulate the consumption and submit it to the ETD model, By doing so, the submitted consumption is accepted as legitimate, even though it is fraudulent. This is achieved by retrieving the optimal changes that need to be applied to the consumption. When the attacker crafts the adversarial data at the testing phase it is known as an evasion attack, which forces the ETD model to misclassify the data. Algorithm 1 is used to do optimal changes in the consumption and then it is tested with the model, this change is done in such a way that it comes near to the actual consumption but it is not legitimate. In this experiment, as no malicious samples are publicly available, we use benign samples that slightly modify and generate malicious samples then test them on different models, including CNN Adaboost, to determine if the algorithm can successfully identify malicious samples.

Algorithm 1 : GenAdversarialExamples

Input: Sample c, target label t, loss function L,
scaling factor α.
Output: adversarial example c'
1: $r \leftarrow [0,0,\ldots\ldots 0]$
2: $c_o \leftarrow c$
3: for i in 0…..N-1 do
4. $p_i \leftarrow -\nabla r\,(L(c_i,t)$
5. $p \leftarrow p + \alpha\, p_i$
6. $c_i' \leftarrow c_i + p$
7. set all negative elements in c_i' to 0.
8.end for
9. return c_i'

In the above Algorithm 1, we generate perturbation p and add a scaling factor α to generate adversarial measurements, which are close to the benign samples. Loss of the model for c is calculated by using L(c,t). N represents the number of iterations, On the other hand, by minimizing the loss L(c+p,t), we ensure that the perturbation leads the perturbed sample towards the target class in an iterative manner. All values in ci' set to zero to generate non negative examples.

Now we have generated adversarial examples which we craft on the models and test it's performance.

IV. Experiments and result analysis

Performance evaluation of the CNN-Adaboost algorithm and other state-of-the-art models in both pre and post-evasion attacks are done in this section. The models are run on the SGCC dataset in a Python 3.9 environment to calculate performance metrics.

A. Evaluation Metrices

The Detection Rate (DR), which is also known as True Positive Rate (TPR) is a measure that helps to determine the proportion of actual positive instances identified correctly by the classifier. In contrast, a true negative (TN) denotes a truly detected benign sample. The False Alarm (FA) Rate, also known as False Positive Rate (FPR), is a measure that helps to determine the proportion of negative instances that are incorrectly classified as positive by the classifier. Lastly, a false negative (FN) refers to a malicious sample falsely detected as benign. Accuracy is a ratio of correct prediction to all prediction; it shows how correctly the machine predicts the outcome.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN).$$

B. Performance Evaluation

Simulation results of the performance of the CNN-Adaboost and other ML models under the evasion attack are shown in Tables III, IV and V. In this attack, adversarial examples which are created using Algorithm 1 are injected and is shown in the Tables. In this experiment, different evasion percentages are used to calculate the Recall, F1-score, DA, FA, and accuracy of the models.

Table III
Impact of Evasion Attacks on the CNN-Adaboost and other Methods
(Evasion Percentage = 0%)

Method	Recall	F1(%)	DR	FA	Accuracy
LR	86.08	95.08	83.99	16.76	84.05
SVM	86.79	95.28	90.92	10.01	90.88
DT	85.21	91.19	89.55	10.08	90.05
RF	83.34	95.21	90.12	9.13	90.98
CNN-Adaboost	93.78	95.67	96.13	3.12	96.35

Table IV
Impact of Evasion Attacks on the CNN-Adaboost and other Methods
(Evasion Percentage = 5%)

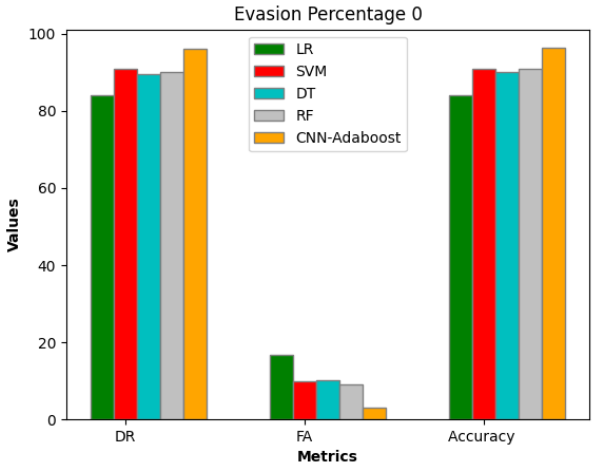
Method	Recall	F1(%)	DR	FA	Accuracy
LR	84.10	91.26	80.48	18.73	80.48
SVM	84.76	90.09	85.90	13.59	85.90
DT	82.45	88.42	86.78	12.32	86.78
RF	81.03	91.78	87.05	12	87.05
CNN-Adaboost	91.78	92.67	93.12	5.94	93.12

Table V
Impact of Evasion Attacks on the CNN-Adaboost and other Methods
(Evasion Percentage = 10%)

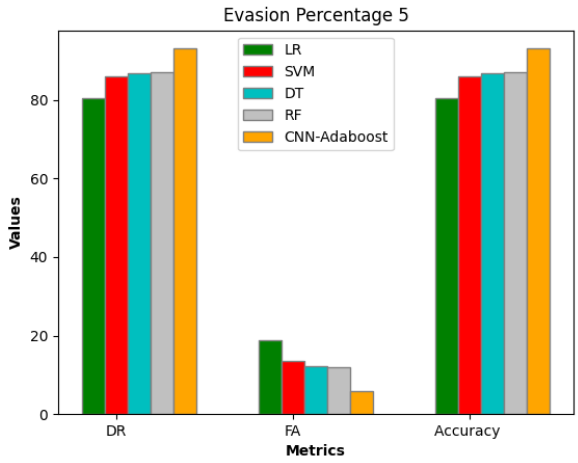
Method	Recall	F1(%)	DR	FA	Accuracy
LR	81.45	88.32	75.00	23.32	76.43
SVM	82.65	87.43	80.12	18.50	79.12
DT	80.32	85.01	81.05	17.76	81.05
RF	79.87	89.02	82.98	16.98	83.98
CNN-Adaboost	88.43	90.97	89.78	11.12	89.23

It is evident that when we carefully craft adversarial examples created in Algorithm 1 on CNN-Adaboost and other state-of-the-art methods used as ETD models, We observed a decrease in accuracy rate as the percentage of adversarial examples increased. We tested the models on 5% and 10% crafted data and found the result as shown in Tables IV and V. The average performance deterioration rates for models vary between 3% and 10% with evasion attacks. We demonstrated the damaging impact of evasion attacks on the performance of CNN-Adaboost and other ETD models. It is found that CNN-Adaboost is more robust as compared to the other models.

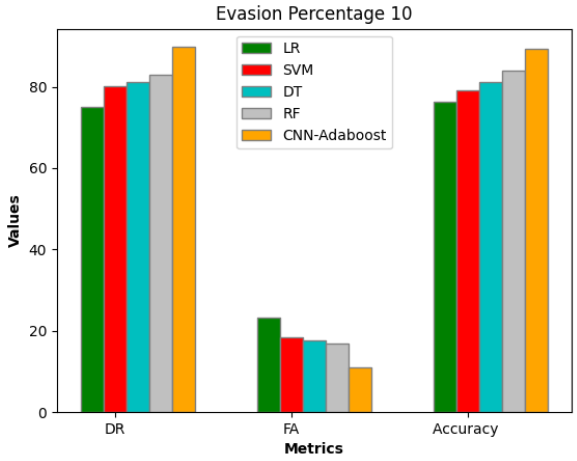
In Fig. 7, a comparison of the performance of different ML models under evasion attacks is shown in bar charts from left for 0%, 5%, and 10%, respectively.



a)



b)



c)

Fig. 7. Impact of Evasion Attacks with 0%, 5%, and 10% adversarial data: a) impact of Evasion Attacks without adversarial data, b) impact of Evasion Attacks with 5%, c) impact of Evasion Attacks with 10%

In Figs. 7a, 7b, and 7c the performance of different ML models under 0%, 5% and 0% adversarial examples when we craft on different ML models. As the crafting percentage increases, performance degrades. According to the line bar chart in Fig. 8, it can be observed that the accuracy of CNN-Adaboost has experienced a decline with an increase in evasion percentage.

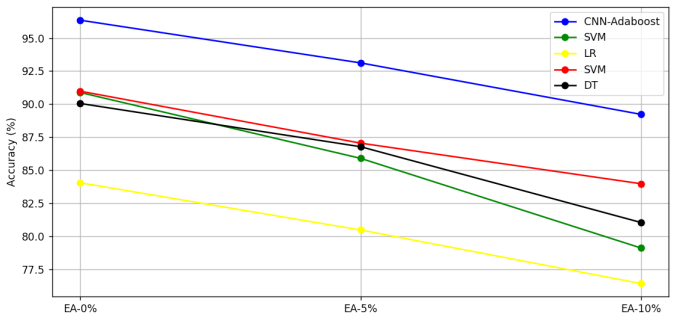


Fig. 8. Accuracy Comparison

The accuracy for the hybrid model is 93.12% for 5% and 89.23% for 10%. Despite this decrease, the accuracy of this model is still comparatively better than that of the other models.

From Fig. 8, it can be concluded that CNN-Adaboost remains robust and performs well compared to other models.

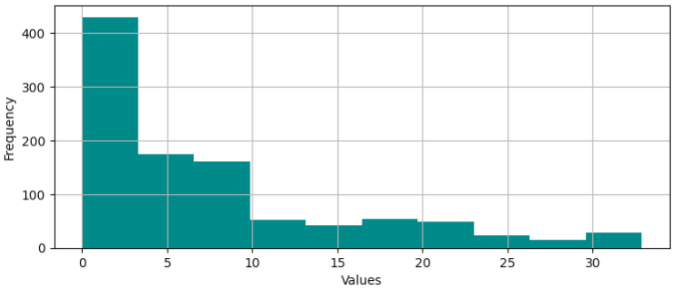


Fig. 9. Histogram for Benign Samples

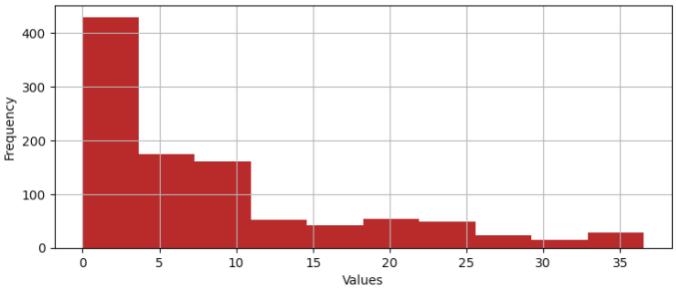


Fig. 10. Histogram for Malicious Samples

In Figs. 9 and 10, histograms for benign and malicious samples are plotted respectively, showing a close match between histograms of benign and malicious samples. There is a minor difference between EC. For consumers, in order to get more profit, consumption can be reduced more but it can be easily detected by the model so by keeping minimum profit, more and more fraud consumers can easily force the model to treat them as legitimate consumers. From the histogram, it shows that the Algorithm 1 generates adversarial measurements which are close to the benign samples.

C. Defence mechanism with adversarial training

It is necessary to build a mechanism to protect the model against such attacks and ensure the robustness of the ETD model. Reduced performance of the model due to the adversarial attack can be improved using adversarial training (AT). It is achieved by augmenting the training process by introducing adversarial examples into the training data and testing its performance. In the training dataset, the benign samples are denoted as 0, and the malicious and adversarial samples are labeled as 1. We found that attack success rate is significantly reduced when the model is trained with the adversarial data.

Suppose the adversarial samples are represented by E. Adversarial training on the model is performed using 20% and 40% adversarial samples, and the performance of the proposed model is evaluated. It is evident that as E increases, the attack success rate decreases while adversarial accuracy improves, as shown in Table VI with 20% and 40% samples. Adversarial accuracy measures how adversarial examples are classified by the model correctly.

We tested the adversarial accuracy of the proposed model along with other methods and found that for some methods, adversarial accuracy is increased as the percentage of E used in adversarial training is increased, and for some methods like SVM and DT, only slide increase in the adversarial accuracy as compared to adversarial accuracy increased in CNN-Adaboost. This happened because of the strength of the adversarial training, the model's capacity to generalize and a mechanism to adopt to adversarial data directly.

Table VI presents the adversarial accuracy for 20% and 40% of adversarial data across different methods. It is observed that the proposed model achieves a higher adversarial accuracy rate compared to the other methods

TABLE VI
ADVERSARIAL ACCURACY OF CNN-ADABOOST AND OTHER METHODS

Method	Adv Accuracy on Adv Sample (E) = 20%	Adv Accuracy on Adv Sample (E) = 40%
LR	82.05	83.48
SVM	84.82	87.03
DT	82.53	84.06
RF	86.54	90.56
CNN-Adaboost	90.15	97.42

The same is illustrated in the chart shown in the Fig. 11.

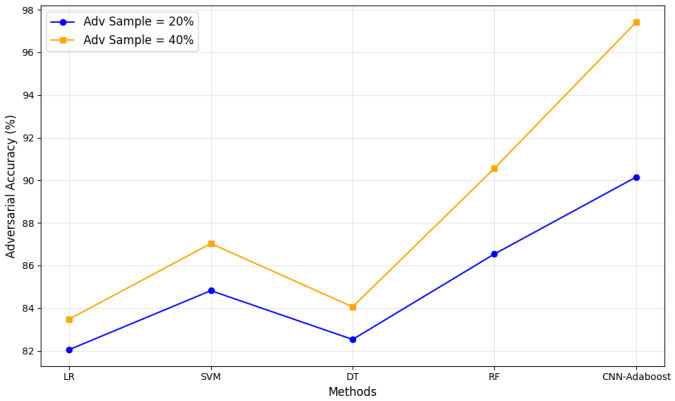


Fig. 11. Adversarial accuracy for 20% and 40% adversarial data

ASR is also calculated for the CNN-Adaboost model for a 20% and 40% adversarial sample, which is approximately 5% and 3%, respectively.

V. CONCLUSION

The adversarial vulnerability of the theft detection modes is extensively explored and paves the way for improving their robustness for secure real-world deployment.

This paper examines the resilience of electricity theft detection systems to adversarial attacks and evaluates their accuracy. We use the CNN-Adaboost model to test its performance with different percentages of adversarial data, and the same data is also applied to the other models and compare their performance. The algorithm used to generate adversarial data to test the performance of electricity theft detection models can also be used to generate data for different applications.

We generated adversarial examples using an algorithm and crafted them on the different models. We evaluate the models' performance on 5% and 10% adversarial data. We observed that as the percentage of EA increased, the performance of the models decreased. Based on our simulation results, we also observed that the CNN-Adaboost hybrid model is more robust than others.

We also proposed a defence system to mitigate the vulnerability and tested the model's performance; we found that it greatly reduced the impact of adversarial attacks and improved the ability of our model to defend against the attack to a certain extent. But in the future, we need to test how effectively it works on other models with different evasion attacks.

This work establishes a foundation for future exploration of countermeasures to enhance the security and resilience of the CNN-Adaboost model, specifically in the context of detecting electricity theft and defending against adversarial attacks.

In the future, the emphasis will be on maintaining the accuracy of the CNN-Adaboost model, New methodologies will be developed to prevent such adversarial attacks and offer a more stable model performance. We hope the proposed research opens multiple research threats both towards finding the vulnerabilities of the tabular classifier and improve their robustness.

Data availability statement

- Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou. (Sep. 30, 2021). Electricity Theft Detection, [Online]. Available:https://github.com/henryRDlab/ElectricityTheftDetection

REFERENCES

[1] Liu, Lijuan, Yanping Wang, and Wanle Chi. "Image recognition technology based on machine learning." IEEE Access (2020).
[2] Hashmi, Sameer Aqib. "Face Detection in Extreme Conditions: A Machine-learning Approach." arXiv preprint arXiv:2201.06220 (2022).
[3] Santosh Nirmal, Pramod Patil, Jambi Ratna Raja Kumar, CNN-AdaBoost based hybrid model for electricity theft detection in smart grid, e-Prime - Advances in Electrical Engineering, Electronics and Energy,Volume 7,2024, 100452, ISSN 2772-6711, https://doi.org/10.1016/j.prime.2024.100452.
[4] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. "A machine learning based credit card fraud detection using the GA algorithm for feature selection." Journal of Big Data 9.1 (2022): 1-17.
[5] Bin Sulaiman, Rejwan, Vitaly Schetinin, and Paul Sant. "Review of machine learning approach on credit card fraud detection." Human-Centric Intelligent Systems 2.1-2 (2022): 55-68.
[6] Khurana, Diksha, et al. "Natural language processing: State of the art, current trends and challenges." Multimedia tools and applications 82.3 (2023): 3713-3744.
[7] Abdallah, Emad E., and Ahmed Fawzi Otoom. "Intrusion Detection Systems using supervised machine learning techniques: a survey." Procedia Computer Science 201 (2022): 205-212.
[8] Liang, Hao, et al. "Decentralized economic dispatch in microgrids via heterogeneous wireless networks." IEEE journal on Selected Areas in communications 30.6 (2012): 1061-1074.
[9] Badr, Mahmoud M., et al. "A Novel Evasion Attack Against Global Electricity Theft Detectors and a Countermeasure." IEEE Internet of Things Journal (2023).
[10] Daniel Lowd, Christopher Meek, "Good Word Attacks on Statistical Spam Filters," Semantic Scholar, https://www.semanticscholar.org/paper/Good-Word-Attacks-on-Statistical-Spam-Filters-Lowd-Meek/16358a75a3a6561d042e6874d128d82f5b0bd4b3
[11] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, Tom Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," in Proceedings of 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montr´eal, Canad, https://papers.nips.cc/paper/7849-poison-frogs-targeted-clean-label-poisoning-attacks-on-neural-networks.pdf
[12] Luis Mu˜noz-Gonz´alez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, Fabio Roli, "Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization," August 29, 2017, https://arxiv.org/abs/1708.08689
[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.
[15] A. Rozsa, E. M. Rudd, and T. E. Boult, “Adversarial diversity and hard positive generation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–32, 2016.
[16] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582, 2016.
[17] N. Carlini and D. Wagner, “Towards evaluating the robustness of neuralnetworks,” in 2017 ieee symposium on security and privacy (sp), pp. 39–57, IEEE, 2017.
[18] Y. Chen, Y. Tan, and D. Deka, “Is machine learning in power systems vulnerable?,” in 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6, IEEE, 2018.
[19] Y. Chen, Y. Tan, and B. Zhang, “Exploiting vulnerabilities of load forecasting through adversarial attacks,” in Proceedings of the Tenth ACM International Conference on Future Energy Systems, pp. 1–11,2019.
[20] J. Li, Y. Yang, J. S. Sun, K. Tomsovic, and H. Qi, “Conaml: Constrained adversarial machine learning for cyber-physical systems,” arXiv preprint arXiv:2003.05631, 2020.
[21] Kotyan, Shashank. “A reading survey on adversarial machine learning: Adversarial attacks and their understanding.” arXiv preprint arXiv:2308.03363 (2023).
[22] Ashrapov, Insaf. “Tabular GANs for uneven distribution.” arXiv preprint arXiv:2010.00638 (2020).
[23] Takiddin, Abdulrahman, Muhammad Ismail, and Erchin Serpedin. “Robust data-driven detection of electricity theft adversarial evasion attacks in smart grids.” IEEE Transactions on Smart Grid 14.1 (2022): 663-676.
[24] Goodfellow, Ian, et al. “Generative adversarial nets.” Advances in neural information processing systems 27 (2014).
[25] Brownlee, Jason. “A gentle introduction to generative adversarial networks (GANs).” Machine Learning Mastery 17 (2019).
[26] Ballet, V.; Renard, X.; Aigrain, J.; Laugel, T.; Frossard, P.; and Detyniecki,M. 2019. Imperceptible Adversarial Attacks on Tabular Data. arXiv preprint arXiv:1911.03274 .
[27] Grosse, K.; Manoharan, P.; Papernot, N.; Backes, M.; and Mc- Daniel, P. 2017. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280
[28] Papernot, N.;McDaniel, P. D.; Jha, S.; Fredrikson,M.; Celik, Z. B.; and Swami, A. 2016b. The Limitations of Deep Learning in Adversarial Settings. In IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbr¨ucken, Germany, March 21-24, 2016, 372–387. IEEE
[29] Chen, Jianbo, Michael I. Jordan, and Martin J. Wainwright. “Hopskipjumpattack: A query-efficient decision-based attack.” 2020 ieee symposium on security and privacy (sp). IEEE, 2020.
[30] Liu, Yugeng, et al. “Backdoor attacks against dataset distillation.” arXiv preprint arXiv:2301.01197 (2023).
[31] Bondok, Atef H., et al. “Novel Evasion Attacks against Adversarial Training Defense for Smart Grid Federated Learning.” IEEE Access (2023).
[32] G. R. Mode and K. A. Hoque, ‘‘Adversarial examples in deep learning for multivariate time series regression,’’ in Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR), Oct. 2020, pp. 1–10.
[33] J. Tian, B. Wang, Z. Wang, K. Cao, J. Li, and M. Ozay, ‘‘Joint adversarial example and false data injection attacks for state estimation in power systems,’’ IEEE Trans. Cybern., vol. 52, no. 12, pp. 13699–13713, Dec. 2022.
[34] Ballet, Vincent, et al. “Imperceptible adversarial attacks on tabular data.” arXiv preprint arXiv:1911.03274 (2019).
[35] Cartella, Francesco, et al. “Adversarial attacks for tabular data: Application to fraud detection and imbalanced data.” arXiv preprint arXiv:2101.08030 (2021).
[36] Singh, Abhijit, and Biplab Sikdar. “Adversarial attack and defence strategies for deep-learning-based IoT device classification techniques.” IEEE Internet of Things Journal 9.4 (2021): 2602-2613
[37] Sun, J., Zhang, T., Xie, X., Ma, L., Zheng, Y., Chen, K., & Liu, Y.”Stealthy and efficient adversarial attacks against deep reinforcement learning.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 04. 2020.
[38] Ezeddin, Maymouna, et al. “Efficient deep learning based detector for electricity theft generation system attacks in smart grid.” 2022 3rd International Conference on Smart Grid and Renewable Energy (SGRE). IEEE, 2022.
[39] M. Ismail, M. F. Shaaban, M. Naidu, and E. Serpedin, “Deep learning detection of electricity theft cyber-attacks in renewable distributed generation,” IEEE Transactions on Smart Grid, 2020
[40] Nowroozi, Ehsan, et al. “Resisting deep learning models against adversarial attack transferability via feature randomization.” IEEE Transactions on Services Computing (2023).
[41] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou. (Sep. 30, 2021). Electricity Theft Detection, [Online]. Available: https://github.com/henryRDlab/ ElectricityTheftDetection