ISSN: 2251-9246
EISSN: 2345-6221

pp. 59:66

A Machine Learning Approach to Detect Energy Fraud in Smart Distribution Network

Mahdi Emadaleslami[1] ,Mahmoud-Reza Haghifam[2]*

[1]Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran. mahdi.emadaleslami@modares.ac.ir
[2]Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran, haghifam@modares.ac.ir

Abstract

Electricity utility have long sought to identify and reduce energy fraud as a significant part of non-technical losses (NTL). Generally, to determine customer's honesty in consumption on-site inspection is vital. Since, inspecting all customers is expensive, utilities look for new ways to reduce inspection's range to cases with a higher probability of fraud. One way to reduce the scope of inspection is to use machine learning (ML) algorithms to analysis consumption pattern. But, their performance is not satisfactory due to insufficiency of fraudulent customers. In this paper, a new two-stage ML-based model is presented to detect fraud in distribution network. In the first stage, an Artificial Neural Network (ANN) is trained to model fraudulent customers, which is used to predict theft scenarios for normal consumers to handle data insufficiency. In the second stage, a Support Vector Machine (SVM) classifier is trained to distinguish normal and suspicious consumers. Assessment and comparison of the proposed algorithm to those of conventional models on a real data set with more than 5000 customers shows its high performance.

Keywords: Energy theft, Fraud detection, Classification of suspicious customers, Load forecasting, Data mining
Article history: Received 27-Jun-2021; Revised 04-Jul-2021; Accepted 12-Jul-2021. Article Type: Research Paper

http://dx.doi.org/10.30495/ijsee.2021.683248

1. Introduction

Energy losses is one of the main problems in the electricity industry and is categorized as technical loss and NTL [1]. Technical losses usually happen due to energy dissipation in the power grid line, which carry energy from generation to customers. Aside from technical losses, not all energy that flows through the network and is consumed by consumers is quantifiable and paid. This unpaid share of energy losses is referred to as non-technical [2]. It is estimated that NTL cost over $89.7 billion as a worldwide business in 2014. These costs, especially energy theft, are not only applied to customers through additional bill payment but are also applied on utilities by increasing distribution network usage risk [3]. Therefore, having a reliable energy fraud detector is advantageous to both customers and utilities.

From past to present, fraudulent consumers have tried to shift high usage to low-tariff periods and change metered values in order to reduce their bill [4]. These cases can only be identified through on-site inspection. However, due to a variety of problems such as customer numbers, expenses, and network complexity, it is impossible to inspect and verify all customers' consumption data. Hence, utilities are constantly striving to limit the inspection to cases with a higher likelihood of theft. From the academic point of the view, many works have been done to reduce the inspection range by detecting energy fraud. These works are generally based on, state estimation, game theory, and classification [۵]. State-estimation methods ,as proposed in [6, 7], might be able to detect areas of high fraud probability but might not be able to identify individual cases. Moreover, special devices, such as wireless sensors and distribution transformers, and thus high costs, affect the detection accuracy of state estimation methods. Game theory-based methods, as proposed in [8], attempt to formulate theft detection problems as a game between the utility and the theft. Although this approach is low-cost and fair [4], modeling the behavior of all players, including regulators, thieves, and distributors, remains a complex and sophisticated challenge.

These challenges, faced by state-estimation and game-theory approaches, have led to the use of new methods such as data mining and artificial intelligence (AI) to classify customers. Moreover, the acceptability and effectiveness of these methods have not only increased over time but also along with advances in technology most studies of load monitoring, theft detection, and other power-system issues are treated with a focus on data mining and artificial intelligence.

The proposed method in this paper is based on classification. Thus, we emphasize on existing works that have used classification algorithms to find energy fraud. Authors in [4] selected areas with high differences between total energy consumed and energy reported by the distribution transformer meter as theft-prone areas. Then, assuming that suspicious customer patterns can be expressed as a function of normal customers, a synthetic dataset was created and an SVM was implemented via clustering to identify abnormalities and inconsistencies in the collected data. In a subsequent study [9], a combination of a decision tree and SVM was used to detect theft. The decision tree estimates a customer's expected consumption based on features such as the numbers of household appliances and members and external temperature. Then the SVM determines if the customer’s consumption pattern is normal or abnormal. authors in [10] used optimum path forest clustering to classify consumption patterns. A consumer was identified as an anomaly based on the load profile's distance from the cluster centroid compared to threshold value, whereas another study [11] used each consumer's normal and abnormal consumption profiles rather than load profiles to train a classification system for each customer individually. Authors in [8] transferred consumer consumption data into the frequency domain and then by modelling the characteristics of a reference period, determined whether the consumption pattern of a target period in the future is normal or abnormal.

More recently, authors in [12] proposed two methods to overcome the problem of data imbalance affecting NTL detection. Abnormal consumers were classified using the KNN algorithm and a synthetic dataset was generated based on the difference between two consumers who had the shortest distance to each other. Finally, a combination of CNN and LSTM algorithms was used to detect theft.

Following the methodology described by [4] and relying on the fact that the consumption behavior has a weekly cycle, a combination of two DNNs has been used to extract features describing consumption behavior and detect electricity fraud in [12]. in another study [13], the CNN algorithm was first trained to learn features at different hours of the day and different days from the consumption data of various customers. Then, the random forest algorithm was used to detect fraudulent customers based on the features obtained. authors in [14], combined consumption data over ten years and an auxiliary database to describe the geographical and technical characteristics of customers who had been inspected at least once, and used the results to train KNN, XG Boost, and logistic regression algorithms. The best performing options based on accuracy and training time were used to provide a list of customers with a probability of anomalies in their meters. [15], generated a dataset of malicious samples based on the consumption patterns of normal consumers and proposed an SVM-based solution to detect fraudulent customers by taking advantage of the predictability of customer consumption patterns.

Given that capturing the periodicity of electricity consumption using one dimensional(1-D) electricity consumption data is known to result in low detection accuracy, a wide and deep convolutional neural network (CNN) model to study theft detection in smart grids has also been developed [16]. Subsequently, a combination of data mining and clustering technique was proposed to identify the abnormal behavior of thieves, which has become more diverse via the tampering with and hacking of smart meters [15]. In this scheme, the Maximum Information Coefficient (MIC) was used to detect thefts that appear normal in shapes. Then, clustering was used to detect abnormal users among thousands of load profiles.

From the system operator (SO) perspective, load monitoring, energy management, and billing can be obtained using fine-grained power consumption readings gathered by smart meters (SMs), which are usually installed at the consumer side. However, fraudulent consumers reduce their bills illegally by tampering with SMs and reporting false readings, which causes financial losses and adversely affects grid performance. Therefore, recent studies [16, 17] taking into account system operator (SO) opinion also pay attention to bill computation and load monitoring alongside theft detection under consumers' privacy. By designing a CNN model based on stable multi-party computing protocols using arithmetic and binary circuits, a privacy-preserving scheme has been proposed in [16] that helps the SO identify fraudulent customers who steal electricity. In an online/interactive session, the SO to test the CNN model using reported power consumption readings conducts these protocols.

In this paper, assuming that theft patterns are predictable through analyzing actual and metered consumption of fraudulent customers a Two-Stage ML-based model to detect energy fraud is presented.

In the first stage, the consumption patterns of normal and fraudulent Customers are clustered, then using real and metered consumption of fraudulent customers, an ANN network is trained to predict various types of theft patterns. The trained ANN network is then used to create a synthetic malicious dataset. Finally, a SVM classifier is trained using historic and malicious datasets to separate normal and suspicious customers.

The primary research contributions of this paper can be summarized as follows:

- Customer’s Livelihood and income data is used alongside energy consumption data to analyze consumption behavior.
- The consumption patterns of customers are extracted and classified by considering the correct number of group.
- An ANN model is trained to predict theft pattern of fraudulent customers.
- A SVM model is trained to classify customers and detect fraud.

The rest of this paper is organized as follows: Section 2 discusses threat models for energy fraud; Section 3 describes methodology and modelling; Section 4 evaluates the proposed method's performance; and finally Section 5 presents conclusions and suggestions.

2. Threat Model

Generally, the primary goal of energy thieves is to manipulate information recorded by smart meters to reduce bills while consuming more energy. Techniques used to manipulate smart meters are categorized into cyber, physical, and data methods, as mentioned in [١٨]. Physical methods try to disable or damage the meter in order to avoid consumption registration. On the other hand, cyber and data methods try to target the metering values within the smart meters or over the communication link to utility without damaging the meter. Identifying these cases is only possible through on-site inspection, and utilities are continually seeking to limit the scope of inspection to cases with a higher probability of fraud. One way to do this is to utilize AI-based methods. Nevertheless, an important challenge in AI-based methods is data imbalance in the ratio of fraudulent to normal customers. A traditional and reliable NTL detection method has always involved the prediction of theft behavior based on the analysis of the characteristics and consumption pattern of consumers. In the past, experts generally proposed a set of simple criteria for distinguishing between normal and suspicious consumers based on a general understanding of their behavior. However, changes in consumer behavior and more diverse types and time of energy usage make it impossible to use such simple criteria to distinguish normal from suspicious consumers. Therefore, criteria that are more complex are required. One way to identify these is to assume that suspicious behavior can be expressed as a function of normal customer and try to derive this function. The derivation of functions describing behavior associated with electricity theft has now received increasing attention, especially in the simulation of different scenarios of energy fraud against the smart grid. However, the functions proposed by researchers, such as [18], are very simple.

In summary, the purpose of energy theft is to manipulate the consumption and specifically the load profile. The algorithm presented in this paper is designed to detect fraud based on the suspicious behavior of fraudulent consumers who have manipulated their consumption using either cyber or databased methods.

3. Methodology and Modeling

A) Modeling

Considering the customer's real and metered consumption within a specified period, as in Figure 1. The real consumption may be slightly different, usually due to measurement error in the meter, but is equal to the metered values if and only if the customer is honest about consumption and the meter has not been manipulated, and vice versa if the metered value is not equal to real consumption. The consumer can then be a candidate of suspicious behavior and is placed on the utility company's inspection list.

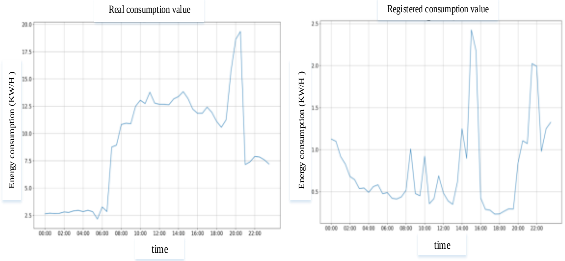


Fig. 1. Real and Metered consumption curve of a sample customer

However, since there is no viable information about the real consumption of fraudulent consumers whose unauthorized usage has been approved by the electricity company, and only their manipulated consumption is available, it is not possible to compare real and metered consumption. However based on the last inspection of fraudulent consumers, in which the real consumption was metered, it is possible to forecast the actual amount of consumption in the period studied by using a short load forecasting (STLF). This procedure expresses what the consumption pattern would be like if the customer's behavior was the same as in the past.

Since load forecasting is an entirely different field, and several methods exist for this purpose [19, 20], the present study merely assumes that an STLF model is available with acceptable accuracy, and that the consumption of suspicious consumers in the study period has been estimated. Any further reference in this paper to real consumption means the value estimated by the STLF module.

The mathematical explanation of the described content can be expressed as in equations 1–4:

$$\vec{X}_i = [x_i, x_2, ..., x_j] \quad (1)$$

$$\vec{Y}_i = [y_1, y_2, ..., y_j] \quad (2)$$

$$\vec{Y}_i = \vec{X}_i + \varepsilon \quad (3)$$

$$\vec{Y}_i = h(\vec{X}_i + \varepsilon) \quad (4)$$

where $\vec{X}_i$ Represents the real consumption vector of consumer i in the period of j, x_j represents the real consumption of the consumer during the time j, $\vec{Y}_i$ is the consumption vector registered and sent by consumer $i\,'s$ meter in the period j, y_j represents the consumption sent to the utility by the meter for the time j, while ε is the error due to meter's measurement error or data transmission as determined by the meter's accuracy class and which is usually less than 2% [21]. Equation 3 shows a normal consumer's behavior whose metered consumption is considered equal to the real consumption plus or minus the error. $h(\vec{X}_i \pm \varepsilon)$ is the mathematical expression of a suspicious consumer. Based on the various analytic scenarios of the effect of energy theft on measured values simple form of $h(\vec{X}_i \pm \varepsilon)$ can be extracted. As an example, the simplest form can be extracted as equation 5 by ignoring meter error and considering that the purpose of the thief is to report less consumption than the real value:

$$h(x) = \alpha x, 0 < \alpha < 1 \quad (5)$$

Where α express the level of fraudulence in reporting consumption. So, for instance, α equal to 0.8 means that the utility receives a reading that represent only 80% of real consumption and 20% of consumption has not been reported via various means. A detailed explanation of other simple forms of h(x) is available in [4].

It is important to note that not all possible forms of $h(x)$, and especially those which are complex and take non-linear forms, can be extracted based on experience and a variety of scenarios; But assuming that the consumption data of some suspicious consumers is available, then the relationship $y = h(x)$ can be expressed as a matrix relation by combining together the vectors $\vec{X}_i$, $\vec{Y}_i$ of these customers:

$$\begin{bmatrix} x_{1n} & \dots & x_{nn} \\ \dots & \dots & \dots \\ x_{1m} & \dots & x_{nm} \end{bmatrix} \cdot \begin{bmatrix} \theta_{11} & \dots & \theta_{1n} \\ \dots & \dots & \dots \\ \theta_{n1} & \dots & \theta_{nn} \end{bmatrix} = \begin{bmatrix} y_{1n} & \dots & y_{nn} \\ \dots & \dots & \dots \\ y_{1m} & \dots & y_{nm} \end{bmatrix} \quad (6)$$

m : Number of suspicious consumers
n : Period of study
x_{mn} : Real consumption value of consumer m at the time of n
θ_{nn} : Impact factor of N-th hour on N-th registered consumption.
y_{mn} : Registered consumption value of consumer m at the time of n

The coefficients of Matrix θ are unknown, and the values of y, x matrix are known. The degree of Equation is determined by n and due to its nonlinearity, as n increases (when considering weekly or monthly consumption), the determination of θ is very difficult and sometimes impossible. One way to estimate Matrix θ, which expresses the relationship between y and x, is to model equation (6) as a multivariate regression problem and use a machine-learning model to estimate these coefficients. Using a neural network model, the real consumption values (matrix X) are considered as input, and the metered values (matrix Y) are considered as output; therefore, the trained model represents the unknown coefficients of matrix θ based on available samples. Figure 2 demonstrates the concept.

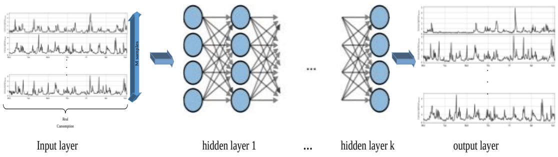


Fig. 2. Regression problem modelling

An important issue that needs to be addressed before training a model is distinguishing the different consumption patterns of suspicious consumers, where each may present a different pattern depending on the method of theft used. It is necessary to identify and categorize the different consumption patterns of suspicious consumers and then train a model based on the patterns categorized. Failure to pay attention to different patterns and their effects will undoubtedly lead to many errors in training and thus reduced model accuracy. One method to extract different consumption patterns based on customer characteristics is the use of clustering algorithms such as the K-shape [22]. Since K-shape is a non-deterministic algorithm, the correct number of clusters must be defined beforehand. One method used to evaluate the correct number of clusters is the Silhouette-score, which measures how similar an instance is to its own cluster compared to other clusters. It is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\left(b(i), a(i)\right)} \quad (7)$$

where a (i) represents the average dissimilarity of i with other samples with other samples in the same cluster, and measures how well an instance is assigned to its cluster; the smaller the value, the better the assignment. Meanwhile b (i) represents least average dissimilarity of i to any other instances in different clusters; the larger the value, the better the separation.

B) *Methodology*

In summary, the proposed model is consisted of 4 step as Figure 3 demonstrates the flowchart.

In Step one, which deals with data preprocessing, each input data is processed to Handle missing values, feature engineering, and data normalization.in handling of missing values, we choose to delete samples with missing values to maintain data integrity. In the feature engineering, since several factors can affect load pattern apart from energy consumption, type of heating and cooling system, number of rooms, Stochastic attributes like maximum, minimum and average consumption are also calculated for each customer. Finally, after preprocessing step, a collection of 60 features including 24-hour consumption data, stochastic, and livelihood features is provided for each customers.

In the second step, K-shape clustering is performed with different values of k on consumption pattern based features provided in step one. For each value of k, silhouette score is also calculated to determine the best number of group in dataset.

In the third step, based on the modeling described in section 3.1, Fraudulent consumption pattern are divided into training and test set with a ratio of 80 to 20%, and several ANN models with different parameters such as number of layers, number of neurons, and activation function with mean square error (MSE) as cost function, batch size equal to 8, and Adam optimizer are trained to predict potential theft patterns for the next 24 hours. Then, using the trained model a synthetic malicious dataset is generated, which contains possible theft scenario for normal customers and is used to train a SVM classifier in the next step.

In the final Step, A SVM model is trained with l' classes, l' is equal to the number of clusters determined in the second step. Energy fraud is detected when a new Sample is identified as suspicious for M times. M, namely fraud limit threshold is determined by utility. M is considered to be 10 in this paper.

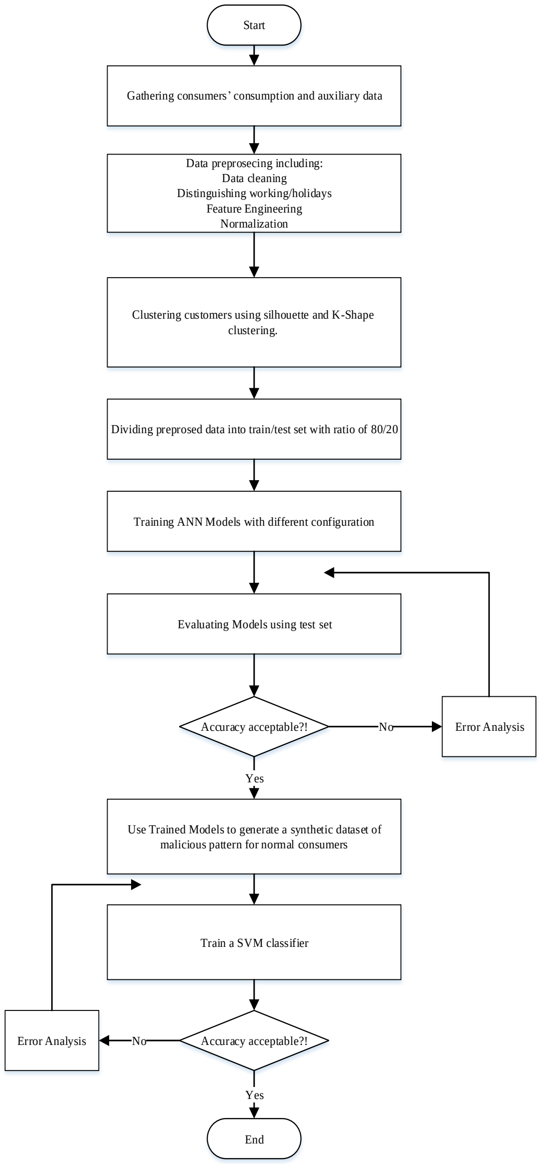

Start
Gathering consumers' consumption and auxiliary data
Data preprosecing including:
Data cleaning
Distinguishing working/holidays
Feature Engineering
Normalization
Clustering customers using silhouette and K-Shape clustering.
Dividing preprosed data into train/test set with ratio of 80/20
Training ANN Models with different configuration
Evaluating Models using test set
Accuracy acceptable?!
No
Error Analysis
Yes
Use Trained Models to generate a synthetic dataset of malicious pattern for normal consumers
Train a SVM classifier
Accuracy acceptable?!
No
Error Analysis
Yes
End

Fig. 3. Flowchart of the proposed algorithm

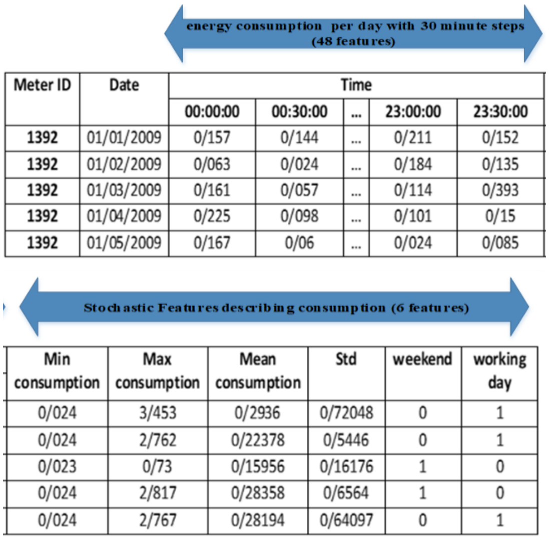
4. Evaluation

To measure the performance of the proposed algorithm, smart meter data from the Irish Smart Energy Trial [23] have been used in this paper. This dataset was released to study and understand the performance and potential risks of smart meters for home consumers and small businesses. This data set contains half-hourly consumption reports of 6445 consumers during 2009-2010 and, for each customer, 48 values per day are registered by meters. In addition to consumption information, two surveys have been provided to collect livelihood and income information from customers. Accessibility, the livelihood and income information, a high and diverse number of consumers, and a two-year period of measurement make this dataset an excellent source for research into load analysis. The present study involves a combination of data mining and artificial intelligence, with the Python programming language alongside Pandas, and Sklearn libraries.

In the first step, preprocessing is conducted and the following results are obtained:

- The input data were provided in the form of 6 .txt files and coded in 5 digit format. Where, the first digit indicates the date, the next two digits indicate the time digits, and the final two indicate consumption. Since, meters are programmed to record daily consumption at 30-minute intervals, the time-digits must be a integer value between 0-48, where zero indicates consumption in the period 00:00-00:30 and 48 indicates consumption in the period 23:30-00:00. Any violation of the above range is excluded from the dataset as in incorrect format.
- Through exploratory data analysis, several consumers consumptions data on 21/04/2010 and 20/04/2010 was found to be erroneous in format and had time-digit values of 49 and 50; this information was deleted to maintain data integrity.
- On 15/09/2009 none of the customers' consumption data were registered between 00:30-1:30. Given the mentioned period is in relation to midnight, it can be assumed that, on this date, either the meters or utility's communication system were updated and consequently zero consumption was registered for all consumers. These information was also deleted to maintain data integrity.

Finally, after preparation, the input data are converted to the proper format as shown in Figure 4.



energy consumption per day with 30 minute steps (48 features)

Meter ID	Date	Time				
		00:00:00	00:30:00	...	23:00:00	23:30:00
1392	01/01/2009	0/157	0/144	...	0/211	0/152
1392	01/02/2009	0/063	0/024	...	0/184	0/135
1392	01/03/2009	0/161	0/057	...	0/114	0/393
1392	01/04/2009	0/225	0/098	...	0/101	0/15
1392	01/05/2009	0/167	0/06	...	0/024	0/085

Stochastic Features describing consumption (6 features)

Min consumption	Max consumption	Mean consumption	Std	weekend	working day
0/024	3/453	0/2936	0/72048	0	1
0/024	2/762	0/22378	0/5446	0	1
0/023	0/73	0/15956	0/16176	1	0
0/024	2/817	0/28358	0/6564	1	0
0/024	2/767	0/28194	0/64097	0	1

Fig. 4. Sample customer data after preprocessing

In the next step, in order to extract different consumption patterns of suspicious consumers, K-shape clustering was performed with different values of k, and each time a Silhouette-score was calculated. According to Silhouette-scores presented in Figure 5, the most appropriate numbers of clusters is 6. Hence, the clustering results for k=6 are presented in Figure 6.

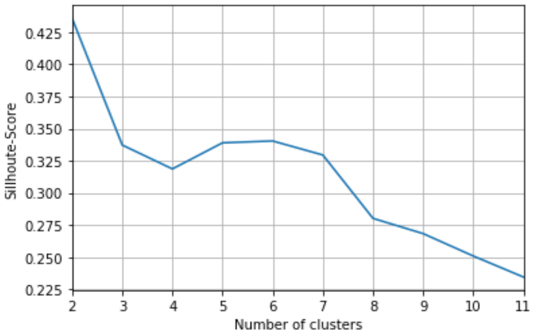


Fig. 5. Silhouette-scores for different values of k

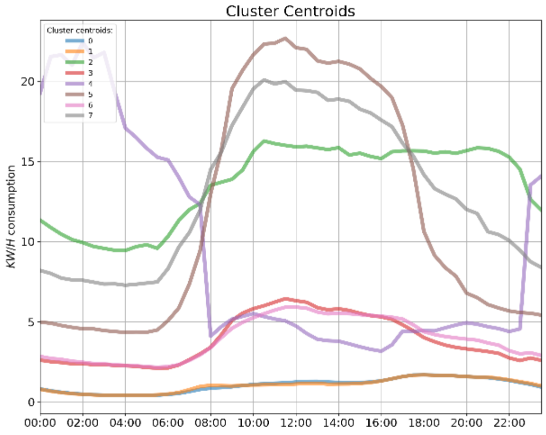


Fig. 6. Consumption pattern clustering results by k-mean

In the next step, based on clustered consumption patterns and given that the training data has 70 features, Several ANN models with different configuration are trained as presented in Table 1. According to Table 1, the mean MSE for the training data after 100 iteration is approximately 4.3% and for the test data is about 8.3%. According to Figure 7, the trained models are almost capable to follow the trend and in some cases are not able to predict high consumption in load pattern. These drawbacks can be explained in two ways. Firstly, since the input data has not been considered as a consecutive time series in the training process, the model is not able to learn the trend; and secondly, due to the small amount of data, one cannot expect very accurate performance from the model. After training several ANN models, the best ones are used to generate a synthetic malicious dataset. This dataset and historic data are then used to provide a SVM classifier. the SVM classifier is trained to detect energy fraud, will considering various theft and normal patterns. This means In addition to detecting whether the input sample is suspicious or not, the group label to which the pattern belong is also determined.

Table 2 demonstrate a comparison among the proposed method and the most recent models in the literature. As one can see the proposed method has a better performance compared to those in the literature even though its structure is much simpler than deep learning model presented in [23].Use of Livelihood data, alongside energy consumption, and clustering technic has helped the algorithm to achieved great performance based on DR and FPR.

Table.1.
Result of ANN models

No. Layer	*No. neurons*	*activation function*	*No. iteration*	*Batch size*	*Train-MSE*	*Test-MSE*
3	(70,60,48)	Relu	100	8	1.30	2.01
4	(70,45,50,48)	Relu	150	16	1.24	1.83
3	(70,64,48)	Tanh	100	8	1.15	1.615

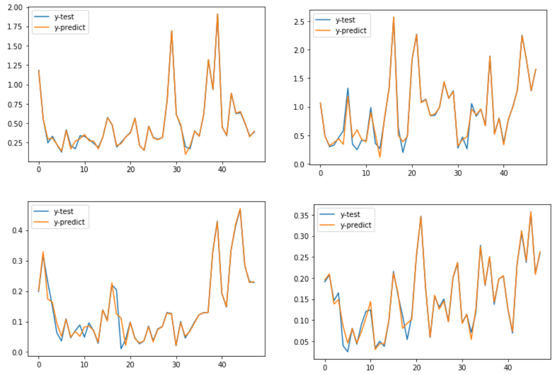


Fig. 7. Sample output of ANN model and real consumptions patterns

Table.2.
Comparison among recent models

Model	DR(%)	FPR(%)	Accuracy(%)
ANN+SVM	98	5.15	95.1
DAE+DSN[23]	95.04	5/14	-
DAE+DNN[23]	97.27	5.56	-
SVM[4]	94	11	-
SVM+K-Means[14]	96	9	-
DT+SVM[9]	-	5.12	92.50

5. Conclusion

In this paper, a two-stage Machine Learning-based model is presented to detect fraud in the smart distribution network. In the first stage, K-shape clustering and Silhouette-score are used to identify and cluster different consumption patterns of customers. Then, using clustered patterns and relying on the predictability of real and metered consumptions pattern, several ANN network are trained to model fraudulent customers behavior, which are then used to generate a synthetic malicious dataset for normal consumers to handle data insufficiency. In the second stage, a SVM classifier is provided to separate normal and suspicious consumers. Assessment of the proposed model on a real data set with more than 5,000 customers and comparison with other state-of-the-art models have shown the superiority of the proposed model from the perspectives of high detection rate, low false positive rate and satisfactory accuracy.

In future research, it is suggested that a combination of LSTM and CNN networks is used to overcome the minor drawbacks of the proposed model mentioned in section 4, since these deep networks have a better performance in the time-series regression.

6. References

[1] T. Baricevic, M. Skok, S. Zutobradic, and L. Wagmann, "Identifying energy efficiency improvements and savings potential in Croatian energy networks," CIRED-Open Access Proceedings Journal, vol. 2017, no. 1, pp. 2329-2333, 2017.

[2] C. W. Groupe, Reduction of Technical and Non-Technical Losses in Distribution Networks. 2017.

[3] Y. He, Y. Chen, Z. Yang, H. He, and L. Liu, "A review on the influence of intelligent power consumption technologies on the utilization rate of distribution network equipment ",Protection and Control of Modern Power Systems, vol. 3, no. 1, pp. 1-11, 2018.

[4] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in AMI using customers’ consumption patterns," IEEE Transactions on Smart Grid, vol. 7, no. 1, pp. 216,۲۰۱۵ .

[5] P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," International Journal of Computational Intelligence Systems, 2016.

[6] S.-C. Huang ,Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," IEEE Transactions on Power Systems, vol. 28, no. 3, pp. 2959-2966, 2013.

[7] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 1023-1032, 2016.

[8] A. A. Cárdenas, S. Amin, G. Schwartz, R. Dong, and S. Sastry, "A game theory model for electricity theft detection and privacy-aware control in AMI systems," in 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1830-1837,2012.

[9] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1005-1016, 2016.

[10] L. A. P. Júnior et al., "Unsupervised non-technical losses identification through optimum-path forest," Electric Power Systems Research, vol. 140, pp. 413-423, 2016.

[11] A. Nizar, Z. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine

method," IEEE Transactions on Power Systems, vol. 23, no. 3, pp. 946-955, 2008.

[12] M. Hasan, R. N. Toma, A.-A. Nahid ,M. Islam, and J.-M. Kim, "Electricity theft detection in smart grid systems: a CNN-LSTM based approach," Energies, vol. 12, no. 17, p. 3310, 2019.

[13] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," Journal of Electrical Computer Engineering, vol. 2019, 2019.

[14] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 2661-2670, 2018.

[15] A. A. Korba, "Smart Grid Energy Fraud Detection Using SVM," presented at the international Conference on Networking and Advanced Systems (ICNAS), 2019.

[16] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," IEEE Transactions on Industrial Informatics, vol. 14, no. 4, pp. 1606-1615, 2017.

[17] G. Figueroa, Y.-S. Chen, N. Avila, and C.-C. Chu, "Improved practices in machine learning algorithms for NTL detection with imbalanced data," in 2017 IEEE Power & Energy Society General Meeting, 2017: IEEE, pp. 1-5.

[18] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz" ,A multi-sensor energy theft detection framework for advanced metering infrastructures," IEEE Journal on Selected Areas in Communications, vol. 31, no. 7, pp. 1319-1330, 2013.

[19] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841-851, 2017.

[20] F. Fahiman, S. M. Erfani, S. Rajasegarar, M. Palaniswami, and C. Leckie, "Improving load forecasting based on deep learning and K-shape clustering," in 2017 International Joint Conference on Neural Networks (IJCNN), 2017: IEEE, pp. 4134-4141.

[21] EPRI, "Accuracy of Digital Electricity Meters,," Electric Power Research Institute, 2010.

[22] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1855-1870.

[23] Irish Social Science Data Archive. [Online]. Available:http://www.ucd.ie/issda/data/commissionforenergyregulationcer/