Journal of King Saud University – Computer and Information Sciences 35 (2023) 13–25

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment

Salah Zidi [a], Alaeddine Mihoub [b,*], Saeed Mian Qaisar [c], Moez Krichen [d], Qasem Abu Al-Haija [e]

[a] Hatem Bettaher Laboratory (IRESCOMATH), University of Gabes, Gabes, Tunisia
[b] Department of Management Information Systems and Production Management, College of Business and Economics, Qassim University, P.O. Box: 6640, Buraidah 51452, Saudi Arabia
[c] Department of Electrical and Computer Engineering, Effat University, 22332 Jeddah, Saudi Arabia
[d] Faculty of CSIT, Al-Baha University, Saudi Arabia and ReDCAD Laboratory, University of Sfax, Tunisia
[e] Department of Computer Science/Cybersecurity, Princess Sumaya University for Technology (PSUT), Amman 11941, Jordan

ARTICLE INFO

Article history:
Received 21 February 2022
Revised 12 April 2022
Accepted 11 May 2022
Available online 17 May 2022

Keywords:
Smart meter data
Energy consumption
Theft detection
Theft generator
Machine learning

ABSTRACT

Smart meters are key elements of a smart grid. These data from Smart Meters can help us analyze energy consumption behaviour. The machine learning and deep learning approaches can be used for mining the hidden theft detection information in the smart meter data. However, it needs effective data extraction. This research presents a theft detection dataset (TDD2022) and a machine learning-based solution for automated theft identification in a smart grid environment. An effective theft generator is modelled and used for obtaining a multi-class theft detection dataset from publicly available consumer energy consumption data, owned by the "Open Energy Data Initiative" (OEDI) platform. This is an important and interesting phase to explore in the smart grid field. The proposed dataset can be used for benchmarking and comparative studies. We evaluated the proposed dataset using five different machine learning techniques: k-nearest neighbours (KNN), decision trees (DT), random forest (RF), bagging ensemble (BE), and artificial neural networks (ANN) with different evaluation alternatives (mechanisms). Overall, our best empirical results have been recorded to the theft detection-based RF model scoring an improvement in the performance metrics by 10% or more over the other developed models.

1. Introduction

The smart grid is a compelling mix of information and communication technologies (ICTs). It comprises sensors and smart meters connected with the energy servers or cloud via the wireless or cable-based network. The electric power can be better managed in the case of a smart grid as compared to the conventional grid (Gul et al., 2020; Adil et al., 2020; Mujeeb and Javaid, 2019; Nazari-Heris et al., 2020). To attain an efficient use of resources in smart grid, framework analysis and dynamic load scheduling are realized implemented (Marzband et al., 2018; Jadidbonab et al., 2020). In Gholinejad et al. (2020) authors presented a hierarchical energy management system for reducing peak hours and trading more electricity at cheaper costs. To decrease the impact of the intermittent nature of renewable energies, an information-gap decision theory-based approach is proposed (Mian Qaisar, 2020). Because of the high cost of obtaining energy and the limited quantity of energy resources available, efficient and effective use of energy resources is a critical element of every country's social and economic growth. The smart grid has emerged as a vital element of making the most of future energy monitoring. The smart grid system is a comprehensive electrical network that includes power system architecture and computers for managing and monitoring energy consumption, as well as an intelligent monitoring system that monitors the usage pattern and mode of action of all customers linked to the system (Khan et al., 2020; Hasan et al., 2019). By combining contemporary digital technology with the current electrical infrastructure, the smart grid allows utilities and consumers to monitor, manage, and anticipate energy consumption.

The bidirectional energy and information exchange is an essential Energy Internet (EI) element (Cao et al., 2018; Wang et al., 2018). It is an advanced version of the smart grid (Zheng et al.,

* Corresponding author.
E-mail address: a.mihoub@qu.edu.sa (A. Mihoub).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.05.007
1319-1578/

2018). The EI is mainly based on the Advanced Metering Infrastructure (AMI) (Karnouskos et al., 2007; Jiang et al., 2014). The AMI provides a high granularity of energy consumption data to the power utilities. It is realized by smartly deploying the smart meters for accurate modeling of the user consumption behavior (Zheng et al., 2018), forecasting of load (Wang et al., 2016), and demand response (Sun et al., 2018). There are two types of losses in the transmission and distribution of electricity: technical and non-technical. Technical losses are the energy losses that occur in the equipment necessary for performing electricity transmission and distribution (Henriques et al., 2020). The Non-technical losses (NTL) happen due to power theft, utility worker misconduct, and billing inconsistencies (Savian et al., 2021). According to a study, the NTL costs utilities roughly US$96 billion each year globally (Hussain, 2021; Northeast Group LLC, xxxx).

Because of the enormous economic loss (Jamil and Ahmad, xxxx; Arango et al., 2017), power suppliers, engineers, and researchers are attempting to reduce NTL using a variety of innovative and efficient ways. The deployment of smart meters-based Energy Internet is one of the most effective ways to combat energy theft. If there is any questionable behavior, such a method may remotely monitor and record the consumers' usage statistics and immediately transmit the information to the utility. Despite their numerous advantages, smart meters are not practical for nations with grave economic problems due to the high costs connected with their deployment and operation. Furthermore, rising cyber risks must be addressed effectively before such devices may be widely used. Because of the specific properties of AMI, securing the EI's information flow is not straightforward. The malicious users can tamper with smart meter data by using intrusion techniques. Consequently, the power thefts on the EI are different from ones, happened in the conventional grid and were mainly the physically evading or extinguishing the mechanical meters (Zheng et al., 2018). Artificial Intelligence (AI) algorithms can allow to automatically monitor the users' energy consumption habits. It can lead towards a reliable identification of power thieves while analyzing the smart meters data.

Renowned agencies such as the US Federal Bureau of Investigation and the Fujian Daily have documented cases of organized energy theft (Zheng et al., 2018; Daily, 2013). These are based on the manipulation of tools and tactics against smart meters and resulted in significant NTL. The effective approaches for electricity theft detection, based on EI, are required to tackle the NTL problem effectively. Effective strategies for electricity theft detection based on EI are needed to tackle the NTL problem effectively since traditional detection methods such as deploying technical personnel or video monitoring are burdensome and require rigorous labor.

Theoretical, hardware, and non-hardware-based methods are the three primary kinds of NTL detection methodologies (Viegas et al., 2017). To combat NTLs, theoretical techniques link socioeconomic and demographic aspects (Yurtseven, 2015). To identify electrical thefts, the hardware-based or state-based techniques employ extra measures in the distribution network (Neto and Coelho, 2013; Leite and Mantovani, 2016). The intruders are prohibited from interfering with network measurements. Therefore, conflicts among the smart meter data and system states will originate. A high theft detection accuracy can be reached but at the cost of deploying extra hardware. Due to increased maintenance and sensor deployment expenses, these approaches are not practical for many power companies.

Non-hardware-based energy-theft detection measures, unlike hardware-based solutions, do not require extra NTL detecting equipment. The AI-based and game theory-based approaches are the two main types of these methods (Jokar et al., 2015). Methods based on the game theory are founded on developing a game among the service provider and fraudulent consumers to detect the NTL (Cárdenas et al., 2012; Amin et al., 2015). Even though these techniques are less expensive, they face problems in defining the essential roles of participants, offenders, regulatory agencies, and distributors, making them too difficult to execute.

On the other hand, AI-based solutions are more viable and utilize machine learning techniques such as classification and clustering (Ahuja et al., 2020) to evaluate consumer load profiles to identify anomalous users since fraudulent users' consumption habits are thought to differ from benign customers. The clustering is the foundation of unsupervised learning and is applicable on an unlabeled dataset (Zanetti et al., 2017; Sun et al., 2016). On the other hand, supervised learning-based classification requires a labeled dataset (Zheng et al., 2017; Ahmad et al., 2018).

To the best of our knowledge, this research is the first study to achieve an automated identification of theft detection as stated above. This contribution is based on data analysis techniques. Using data consumption energy of different consumers, various machine learning techniques are applied and compared to learn and detect abnormal consumption behavior. In this paper, an AI-based approach is used. Specifically, the major contributions of this paper are:

- We propose an effective theft generator that can help analyze energy consumption behaviour in smart grid environments.
- We present a multi-class theft detection dataset for classifiers performance evaluation and benchmarking.
- We develop an intelligent autonomous detection system involving six different types of theft.
- We provide extensive simulation results characterizing the performance of five different ML techniques (KNN, DT, RF, Bagging, ANN).

The remaining part of this paper is structured as follows: the related works are presented in Section 2. In Section 3, information about the proposed dataset is provided. Moreover, classification approaches and the performance assessment measures are also discussed in Section 3. Section 4 presents the results as well as the discussion of the main findings, and the conclusion is made in Section 5.

2. Related works

To address the increasing problem of electricity fraud, numerous Non-Technical Losses detection approaches have been adopted. Non-Technical Losses detection solutions incorporate contributions from various knowledge domains, ranging from hardware to data-driven solutions. Hardware solutions use devices (Henriques et al., 2014)to monitor grid system characteristics like power, current, and voltage. The main disadvantage of this strategy is that it necessitates additional equipment, making it costly.

The advent of Advanced Metering Infrastructure in the smart grid has resulted in a high level of data availability. As a result, data-driven strategies for detecting NTL have lately been favored. Game theory (Cárdenas et al., 2012; Amin et al., 2015), Statistical methods (Singh et al., 2018; Tao and Michailidis, 2020), and machine learning methods (Adil et al., 2021; Maamar and Benahmed, 2018) are examples of such methods. Compared to hardware alternatives, these methods are easier to implement and less expensive. In Jiang et al. (2014), the energy-theft techniques were classified as supervised, semi-supervised and unsupervised. In Nizar and Dong (2009), these techniques were classified as game theory-based, state-based and classification-based.

An unbalanced class exists in most electrical theft datasets, with anomalous thieves much smaller than typical users (Maamar and

Benahmed, 2018). The trained model will cover the most frequent categories and ignore the less frequent ones because of the unbalanced dataset. The authors of Pereira and Saraiva (2020) conducted a comparative study of several strategies for balancing data sets and applied several machine learning techniques to determine which machine learning and data handling techniques produce the best results for simulations related to the problem of electricity theft detection. When comparing the balancing strategies for the same machine learning method and comparing these combinations between themselves, the results demonstrated that some varieties could get much better values than others.

Prior fabrication modeling approaches play a huge role in replicating various solar electrical systems. The survey paper reported in Ahmad et al. (2018) focuses on the different modeling approaches for identifying and predicting non-technical losses. This research focused on modeling methodologies, which save time and protect the financial investment in the electrical system. This review study also discusses the advantages and upcoming opportunities of modeling approaches. In addition, to properly detect Non-Technical Losses, data-driven strategies must capture the behavior of both normal and abnormal consumption patterns. However, datasets with real anomalous cases are unavailable for training and verifying detection models. As a result, a method of constructing Non-Technical Losses attack models to capture the characteristics of real-world assault scenarios was proposed by a lot of scientists (Chen et al., 2019; Yip et al., 2017; Messinis et al., 2019; Nabil et al., 2018; Jokar et al., 2016).

The authors of Bohani et al. (2021) proposed an important comparison study on supervised learning approaches for detecting electricity theft. In this research, performance comparisons of numerous supervised learning approaches (such as AdaBoost, decision trees, artificial neural networks, and deep artificial neural networks) are presented and studied. This analysis relied on a publicly available dataset from China's State Grid Corporation. According to the research results, the deep artificial neural networks outperformed other supervised learning classifiers.

Similarly, the authors of Chuwa and Wang (2021) proposed an interesting survey in which they explored the forms and reasons of attacks in the advanced metering infrastructure system that result in non-technical losses. They also looked at attacks that resulted from non-technical losing attack models. They also discussed numerous features and feature-engineering methodologies (Punmiya and Choe, 2019; Razavi et al., 2019; Zhang et al., 2020) and their ability to separate normal and attack samples into distinct classes. The results of several learning models in detecting different attacks were also investigated in this study. The authors also provided a summary and recommendations for improving non-technical loss attack detection.

Compared to existing works, the originality of our work mainly lies in the following aspects:

1. We generated a new dataset based on a real dataset and not a purely synthesized one (See Section 3.1). The original dataset is extracted from smart-grids using smart meters.
2. The generated dataset has a considerable size (560, 640 instances) and important features considering existing datasets (see Table 1).
3. We considered a variety of models for electricity thefts (See Section 3.2 and Section 3.3) with various types of electricity consumers (See Table 1).
4. We applied several machine-learning classification techniques for predicting electricity thefts (See Section 3.4).

3. Materials and methods

This section describes our proposed dataset with its features and statistics, the proposed theft generator algorithm, the detection system development approach, the employed classification models, and the performance evaluation measures used to evaluate the system performance for all alternatives and mechanisms.

3.1. The dataset

We collected the data used in this work from the Open Energy Data Initiative (OEDI) platform. It is a centralized repository of high-value energy research datasets aggregated from the U.S. Department of Energy's Programs, Offices, and National Laboratories (Leite and Mantovani, 2016). This big data collection project is based on the data lake concept described in Fig. 1.

Table 1
Summarized description for the information and statistics of the generated dataset.

General Statistics		List of consumer types			
Item	Numbers	SN*	Customer Type	SN*	Customer Type
Total number of instances	560,640	1	FullServiceRestaurant	9	Warehouse
Number of columns	12	2	Hospital	10	SecondarySchool
Number of features	11	3	LargeHotel	11	SmallHotel
Number of numerical features	10	4	LargeOffice	12	SmallOffice
Number of categorical features	1	5	MediumOffice	13	Stand-aloneRetail
Label encoding technique	IE *	6	MidriseApartment	14	StripMall
Number of consumer types	16	7	PrimarySchool	15	SuperMarket
Number of instances per consumer type	35,040	8	OutPatient	16	QuickServiceRestaurant

Features Information		**Statistical Numbers for Each Class**		
Feature Name	**Type**	**CN***	**Class name**	**Number of instances**
Electricity:Facility [kW](Hourly)	Float	1	Normal	331,824
Fans:Electricity [kW](Hourly)	Float	2	Theft1	51,083
Cooling:Electricity [kW](Hourly)	Float	3	Theft2	22,958
Heating:Electricity [kW](Hourly)	Float	4	Theft3	44,349
InteriorLights:Electricity [kW](Hourly)	Float	5	Theft4	41,460
InteriorEquipment:Electricity [kW](Hourly)	Float	6	Theft5	33,553
Gas:Facility [kW](Hourly)	Float	7	Theft6	35,413
Heating:Gas [kW](Hourly)	Float	* IE stands for Integer Encoding Technique.		
InteriorEquipment:Gas [kW](Hourly)	Float	* CN stands for the class number.		
WaterHeater:WaterSystems:Gas [kW](Hourly)	Float	* SN stands for the sequence number.		
ConsumerType	String			

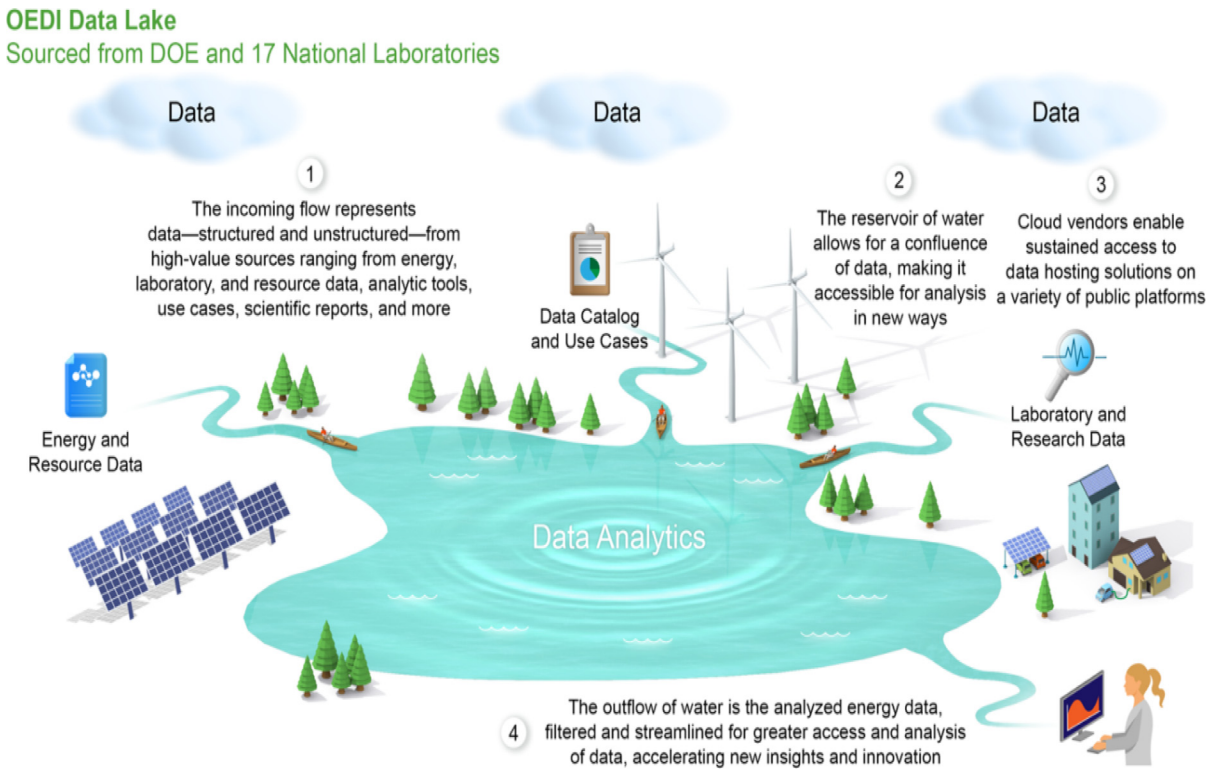

Fig. 1. The OEDI Data Lake.

It consists of curated and diverse datasets built to accelerate accessibility and collaboration. The information in this data lake comes from various sources, including private industry, laboratories, institutions, etc. The dataset contains energy consumption for 16 different types of consumers. The original data includes several energy consumption measurements for several customers for one year (12 months). Measurements are taken every hour throughout the day (24 h).

Table 1 provides further information and statistics about the generated dataset. To make it short and for better readability, the important details about the dataset are summarized in summarized items.

3.2. The theft generator

As described in other research studies (Jokar et al., 2015; Cárdenas et al., 2012; Amin et al., 2015), and (Ahuja et al., 2020), we consider six different types of frauds. They consist of different types of thefts that some consumers can cause. The first type of theft consists of a considerable reduction of electricity consumption during the day. This reduction is calculated by multiplying the consumption by the randomly chosen value between 0.1 and 0.8. In the second type of theft, electricity consumption drops to zero at random and during an arbitrary period. The third type of theft is similar to the first type, but each consumption value (each hour) is multiplied by a random number. A random fraction of the mean consumption is generated for the fourth type of theft. The fifth type reports the mean consumption, and the last type of theft (i.e., the sixth type) reverses the order of readings. We developed a theft generator that enabled us to generate these six types of theft as described previously randomly. Consequently, we developed a python program to generate the examples of thefts randomized and store them in the energy consumption database.

The proposed algorithm to generate the six types of electricity thefts can be formally stated as follows: Suppose the daily electricity consumption vector (X) is given as:

$X = \{x_1; x_2; x_3; \cdots\cdots\cdots x_{24}\}$ where x_i represents the hourly consumption for i = 1…0.24, then the theft types can be generated as:

Algorithm: Theft Types Generation

Inputs: X, Output: $TheftN$, $where\, N = 1, 2, \cdots 6$
Begin:
$Theft1(x_i) = \alpha.x_i, \alpha = random(0.1, 0.8)$
$Theft2(x_i) = \beta_i.x_i,$
$\beta_i = \begin{cases} 0 t_{start} < i < t_{end} \\ 1 otherwise \end{cases}$
$t_{start} = random(0, 23 - t_{off})$
$duration = random(t_{off}, 24)$
$t_{end} = t_{start} + duration$
$t_{off} \geq 4$
$Theft3(x_i) = \gamma_i.x_i, \gamma_i = random(0.1, 0.8)$
$Theft4(x_i) = \gamma_i.mean(x), \gamma_i = random(0.1, 0.8)$
$Theft5(x_i) = mean(x),$
$Theft6(x_i) = x_{24-t},$

All the six types of theft can be found in other works such as some research studies referenced at the beginning of this section. In the different theft types, there is the random aspect in the choice of the theft period as well as in the choice of the reduction factor. We think this may describe most existing types of thieves' behaviors. Their objective is always to reduce consumption over a random period of time. Sometimes it's just a question of stopping the

counter (meter) such as in the case of theft2. For this research work, we applied these different types in a random but fair way on the dataset. For your information the generated dataset related to this article can be found at http://dx.doi.org/10.17632/c3c7329tjj.1 an open-source online data repository hosted at Mendeley Data (Salah et al., 2022).

3.3. Detection Modelling Approach

As described above, the data has been compiled for 16 different consumers (listed in Table 1). To properly study and choose the best methodology, our strategy was based on a succession of learning tests according to four validation mechanisms as depicted in Fig. 2. We started with a first mechanism (P7C) which consists of classifying the data according to 7 classes (6 thefts and normal cases) using as inputs the consumption features and the consumer type, i.e., that the consumer type is known for the classifier. In the second mechanism (P6C), we considered only six classes (5 thefts and normal cases). We noticed that the detection of theft6 is very difficult, as shown in the confusion matrix (Table 8), and thus we proposed an additional mechanism that deals with six classes only. We also apply a classification with a known type of consumer in this mechanism. For the other two mechanisms (P7U and P6U), the classification is carried out independently of the type of consumers. In these two mechanisms, the types of consumers are unknown, and we have also considered seven classes for the P7U mechanism and only six classes for the P6U mechanism. For these protocols, only the ten consumption features are used as inputs. Other classifiers have been also tested which are the individual models, as shown in Fig. 3.

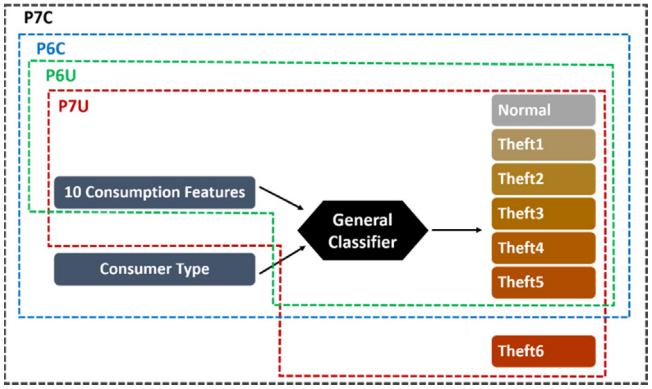

Fig. 2. The four validation protocols of theft detection.

The intuition behind these models is to train a specific model for each particular consumer exclusively with that consumer's data. Since we have 16 consumer types, we end up with a total number of 16 classifiers. The theft detection process begins with a consumer type distinction based on a vector of 10 energy consumption attributes at the input. This parameter is considered significant for the theft detection system. However, this significance will be validated or neglected later in this article when we compare the different protocols described previously. From the full generated base, a split is carried out in a first step according to the type of the consumer. Then the learning process starts on the data of each type. All results are verified for 6 and 7 classes.

3.4. Classification methods

In this section, we present the principle of different machine learning techniques used in this work for theft detection.

A. The k-Nearest Neighbor (KNN)

kNN algorithm is a supervised machine learning technique used to address classification and regression problems. KNN performs a categorization of any object based on a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (Guo et al., 2003; Kataria and Singh, xxxx). For example, in Fig. 4, the yellow point will be assigned to the red class if k = 3 since it is the winning class (2 vs. 1), but it will be assigned to the green class if k = 7 (4 vs. 3).

B. Decision Trees (DT)

DT is a supervised machine learning algorithm that uses a tree to build the predictive decision model to address the classification of regression tasks (Kotsiantis, 2013; Myles et al., 2004). This predictive model is based on a traversing of a decision tree (Fig. 5) where the branches represent observations about an item (conjunctions of features that lead to the class labels), and the leaves represent conclusions about the item's target value (class label).

C. Random Forest (RF)

RF algorithm is a supervised machine learning technique. In the training phase, this approach consists of constructing many decision trees (A random forest guided tour, 2022; Shaik and Srinivasan, 2019). For classification tasks, the output of this algorithm is the class selected by the majority of trees, as described in Fig. 6.

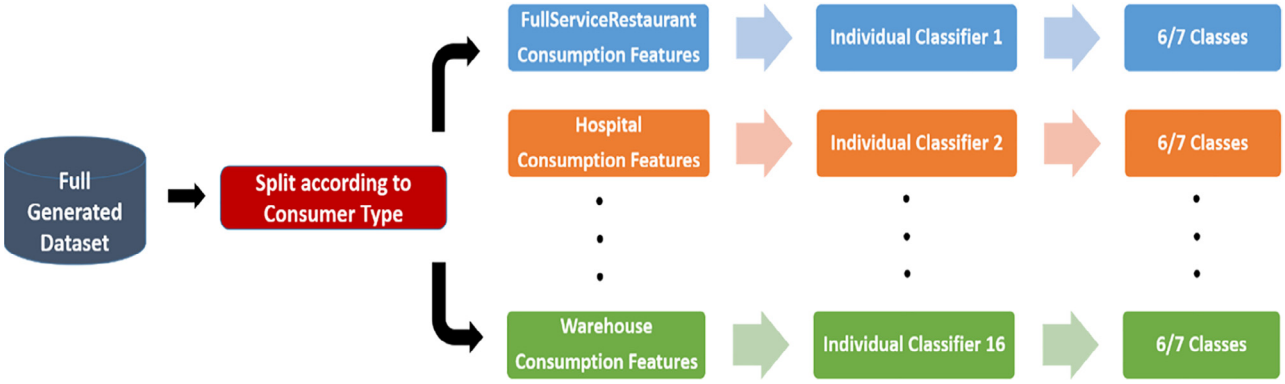

Fig. 3. The full detection process with known types of consumers.

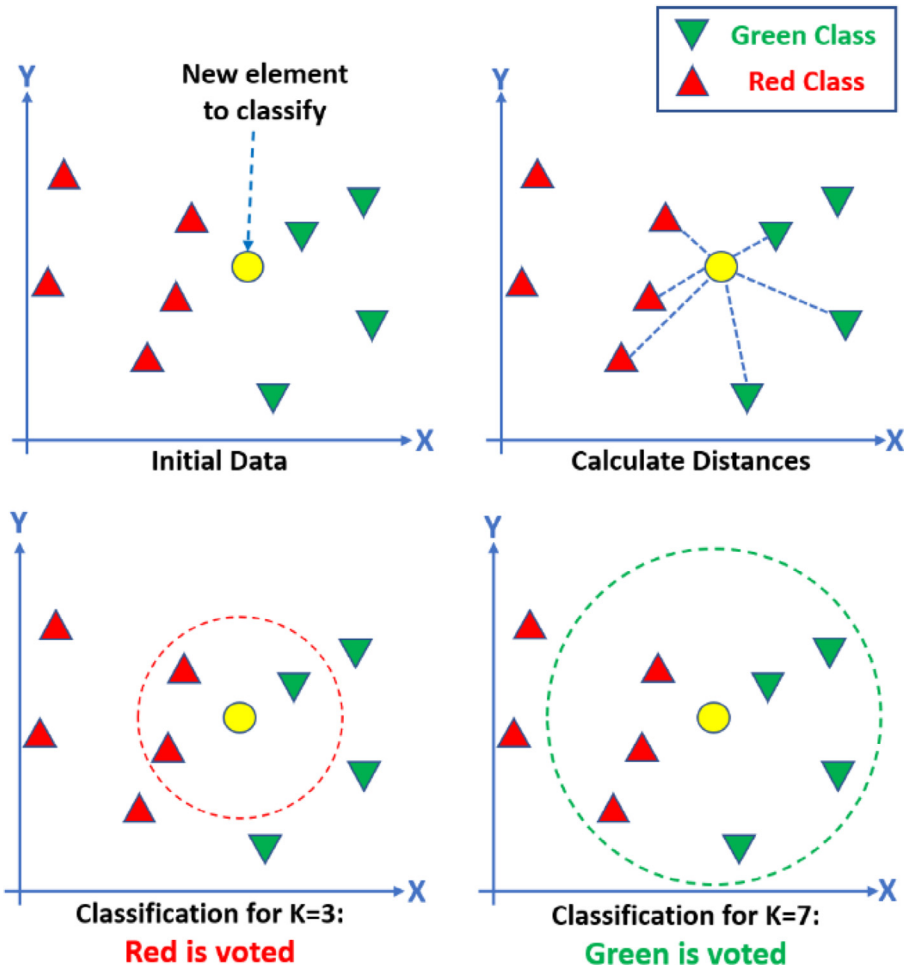

Fig. 4. An Illustration of the KNN algorithm.

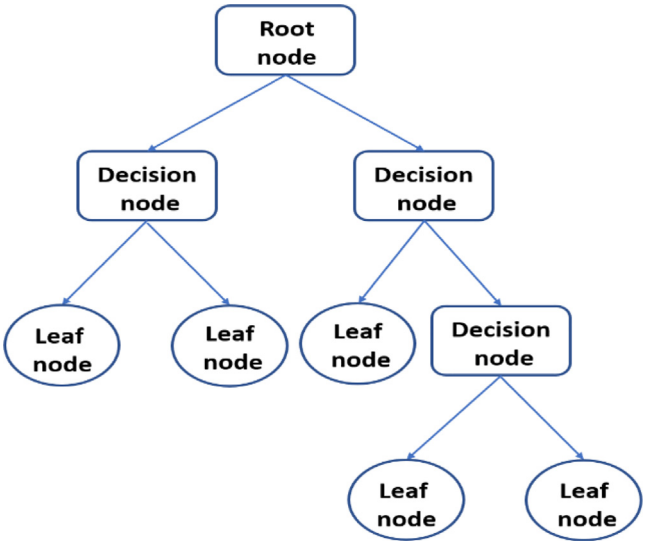

Fig. 5. The general scheme of a Decision Tree.

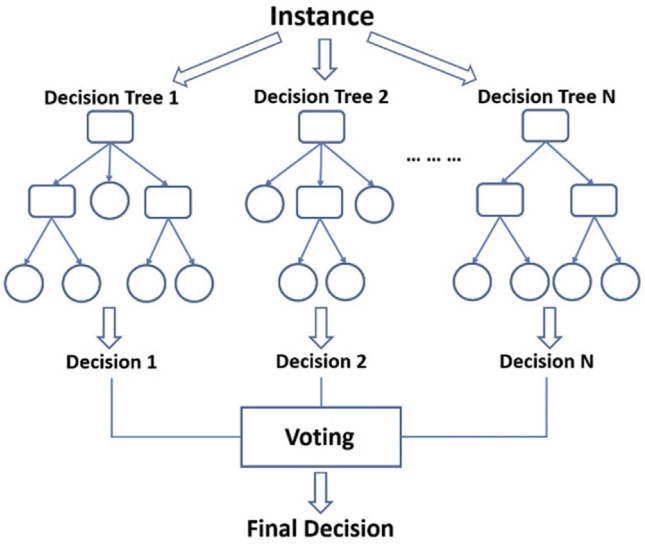

Fig. 6. An illustration of the Random Forest (RF) Algorithm.

D. Bagging

Bagging is an ensemble learning technique for reducing variance in a noisy dataset (González et al., 2020; Breiman, 1996). As illustrated in Fig. 7, this method involves picking a random sample of data from a training batch. These weak models are then trained independently after multiple data samples have been generated. The average or majority of the projections then produces a more

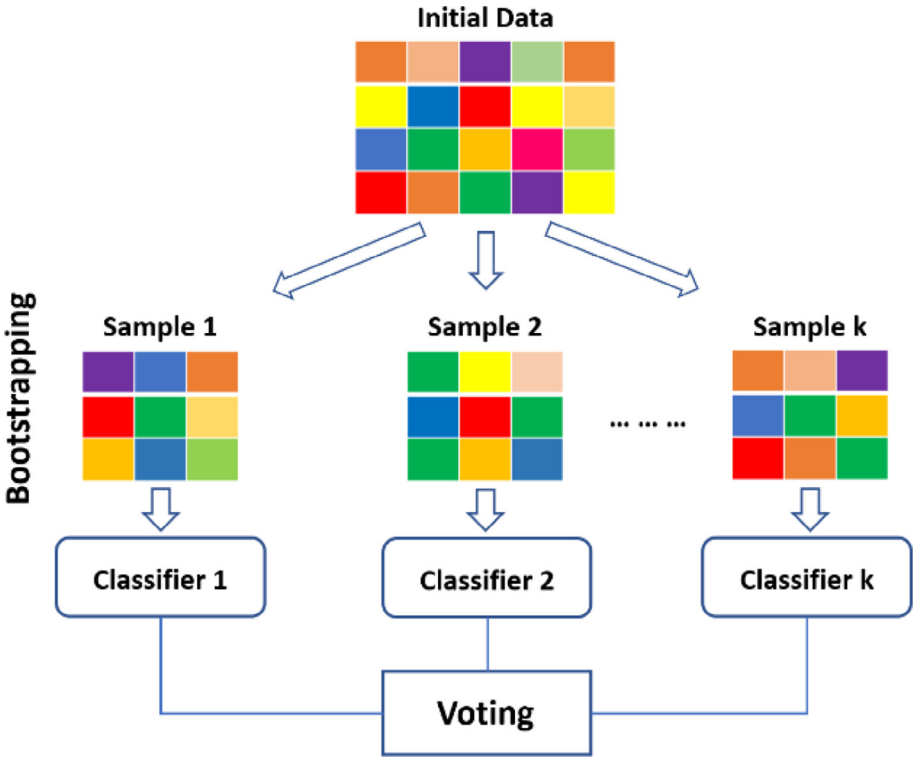

Fig. 7. An illustration of the Bagging Technique.

precise approximation. The random forest technique, which uses both bagging and feature randomness to produce an uncorrelated forest of decision trees, may be seen as an extension of the bagging method (Machová et al., xxxx).

E. Artificial Neural Network (ANN)

Artificial neural networks (ANN) are computer systems modeled after the biological neural networks that make up animal brains (Wang and Wang, 2003; Zhang and Zhang, 2018). Artificial neurons are a collection of connected units or nodes that make up an ANN. Each link has the capability of sending a signal to other neurons. An artificial neuron receives a signal, analyses it, and sends signals to connected neurons. Each neuron's output is generated by some non-linear function of the sum of its inputs, and the "signal" at a connection is a real number. Neurons are usually grouped into layers. As illustrated in Fig. 8, there are three types of layers, namely: one input layer, several hidden layers, and one output layer. On their inputs, separate layers may apply different transformations.

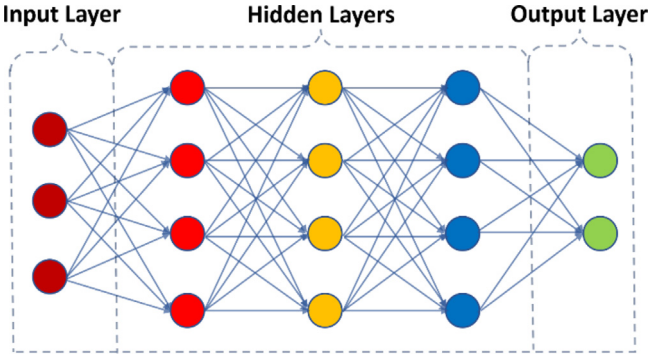

Fig. 8. The general structure of an Artificial Neural Network (ANN).

3.5. Performance evaluation measures

The main contribution of this paper is to present a new theft detection dataset. It can be used to benchmark the performance of different automatic classification approaches for categorizing different categories of theft. The findings, obtained after the first trials with considered machine learning algorithms, are presented in Section 4. This study considers the commonly used performance evaluation metrics such as Accuracy, F-measure, Kappa index (Kappa), and Area under the Receiver Operating Characteristic (ROC) Curve (AUC) (Mohammad and Sulaiman, 2015). The AUC is the graph's area under the curve, calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The resultant confusion matrices are used to calculate the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The Accuracy, F-measure, Kappa, TPR, and FPR values are computed using these parameters as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F - measure = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2}$$

$$Kappa = \frac{Accuracy - Pe}{1 - Pe} \tag{3}$$

Where *pe* is the theoretical probability of chance agreement, is given by:

$$Pe = \frac{(TP + FP) \times (TP + FN) + (TN + FP) \times (TN + FN)}{(TP + TN + FP + FN)^2} \tag{4}$$

$$TPR = \frac{TP}{(TP + FN)} \tag{5}$$

$$FPR = \frac{FP}{(FP + TN)} \tag{6}$$

The Accuracy is a metric for determining how successfully the developed models can categorize data. It is the proportion of labels that have been categorized properly. The percentage Accuracy might range from 0 to 100, with a greater number indicating better performance. The F-Measure encompasses both evaluations the precision and recall in a single measure. It complements the performance evaluation measure, provided by the simple Accuracy score. The Kappa is typically thought to be more reliable than basic Accuracy since it accounts for the potential of coincidental agreement. If the classification is flawless, the percentage value of Kappa is 100, and if it is only due to chance, Kappa = 0. The AUC displays the performance of a classifier graphically. It displays a tradeoff between the TPR and FPR and gives a more in-depth look at the categorization performance.

3.6. Application

Before the implementation of the aforementioned classifiers, the feature ranges were normalized (between 0 and 1) using a min–max scaler. This operation is suggested, and in some cases, required, for certain classification systems. Moreover, preprocessing operations, training, and testing steps were carried out using well-known Python-based data science libraries such as Pandas, Numpy, and Scikit-learn. The empirical results were obtained using a PC with 16 GB of RAM and an Intel® Core™ i7-8550U processor. We investigated hyper parameter tuning methods to determine the appropriate hyper parameters for each classification algorithm and to avoid overfitting problems.

In our application, and especially for the KNN approach, several k values were tested and the best results were obtained using k = 10. For the Random Forest approach, since it represents an ensemble method, it mixes the output of multiple estimators. The optimal number of estimators in our implementation was 100. For Decision Trees, the Scikit-learn package's default parameters were found to give excellent results compared to other modified values. Similar to Random Forests, the Bagging approach is part of the ensemble models since it aggregates the outputs of several classifiers. In our case, the KNN classifier is adopted as the base classifier and a total number of 10 classifiers yielded the best results. For the ANN model, the typical multilayer perceptron (MLP) technique is used for classification. Two hidden layers, each with about 50 nodes have been found to be the optimal topology for the MLP design. The results of all these implemented classifiers are presented and discussed in the next section.

4. Results and discussion

This section provides our extensive experimental results for the proposed electricity theft detection system using the aforementioned mechanisms and ML techniques. We remind that the evaluation metrics of the five chosen classifiers are described in subsection 3.5. In addition, two main mechanisms are tested: models with seven output classes and six output classes. For each main mechanism, two sub-mechanisms are also tested.

The first one, called "*known consumer*" has used for input features, consumption features, and consumer type. The second one, called "*unknown consumer*" uses only the consumption features; the type of the consumer here is dropped out from the features list. As a result, four mechanisms are tested and experimented as shown in Tables 2–5. These mechanisms are "*7 classes - known consumer*", "7 classes - unknown consumer", "6 classes - known consumer", and "6 classes - unknown consumer". We denote these four protocols, respectively P7C, P7U, P6C, and P6U.

The random forest (RF) classifier recorded the best overall results when comparing all classifiers. For instance, the accuracy rates scored for P7C, P7U, P6C, and P6U were respectively 85%, 84.89 %, 94.71%, and 94.64%. For the same classifier and mechanisms, the F-measure rates scored for P7C, P7U, P6C, and P6U were84.06%, 83.95%, 94.56%, and 94.50%. Respectively, the AUC results for the four protocols were 92.44%, 92.54%, 98.58%, and 98.55%. For the Kappa index results, they were slightly lower, for example, rates were90.74% and 90.63% for P6C and P6U.

In Table 6, we report on the results obtained for the best classifier RF model. The displayed results are averaged results to compare the overall performance of the 7-classes and the 6-classes mechanisms. According to the table, all metrics of the 6-classes mechanism are significantly larger than those for the 7-classes mechanism. Examples include the accuracy of the 6-classes mechanism (94.68%) which is higher than that of the 7-classes mechanism (84.95%) by nearly $\sim$ 10%.

Table 2
Accuracy results of all classifiers.

Accuracy				
Number of classes	7 classes		6 classes	
Classifier/Protocol	Known Consumer	Unknown Consumer	Known Consumer	Unknown Consumer
KNN	84.91 (±0.05)	84.69 (±0.07)	90.72 (±0.04)	90.50 (±0.04)
DT	82.67 (±0.10)	82.48 (±0.09)	93.36 (±0.08)	93.19 (±0.05)
RF	85.00 (±0.09)	84.89 (±0.10)	94.71 (±0.03)	94.64 (±0.05)
Bagging	84.85 (±0.05)	84.65 (±0.05)	90.76 (±0.03)	90.56 (±0.04)
ANN	80.49 (±0.69)	78.55 (±0.85)	86.41 (±0.11)	85.52 (±0.88)

Table 3
F-measure results of all classifiers.

F-measure				
Protocol	7 classes		6 classes	
Classifier/Dataset	Known Consumer	Unknown Consumer	Known Consumer	Unknown Consumer
KNN	81.80 (±0.05)	81.55 (±0.07)	90.20 (±0.04)	89.94 (±0.05)
DT	82.33 (±0.04)	82.13 (±0.05)	93.32 (±0.08)	93.15 (±0.05)
RF	84.06 (±0.07)	83.95 (±0.08)	94.56 (±0.03)	94.50 (±0.05)
Bagging	81.91 (±0.06)	81.68 (±0.06)	90.23 (±0.03)	89.99 (±0.04)
ANN	76.20 (±0.22)	73.24 (±0.90)	84.70 (±0.78)	83.62 (±0.94)

Table 4
Kappa results of all classifiers.

Kappa				
Protocol	7 classes		6 classes	
Classifier/Dataset	Known Consumer	Unknown Consumer	Known Consumer	Unknown Consumer
KNN	74.68 (±0.07)	74.26 (±0.11)	83.72 (±0.09)	83.29 (±0.09)
DT	71.92 (±0.16)	71.60 (±0.14)	88.38 (±0.16)	88.08 (±0.10)
RF	75.49 (±0.15)	75.31 (±0.16)	90.74 (±0.06)	90.63 (±0.09)
Bagging	74.60 (±0.08)	74.21 (±0.08)	83.80 (±0.07)	83.41 (±0.09)
ANN	66.68 (±1.44)	62.87 (±1.61)	75.82 (±0.17)	74.15 (±1.60)

Table 5
AUC results of all classifiers.

AUC				
Protocol	7 classes		6 classes	
Classifier/Dataset	Known Consumer	Unknown Consumer	Known Consumer	Unknown Consumer
KNN	91.12 (±0.05)	91.05 (±0.09)	96.40(±0.03)	96.28 (±0.04)
DT	84.26 (±0.08)	84.05 (±0.09)	92.48 (±0.10)	92.29 (±0.06)
RF	92.44 (±0.06)	92.54 (±0.06)	98.58 (±0.004)	98.55 (±0.01)
Bagging	91.74 (±0.05)	91.63 (±0.04)	96.66 (±0.02)	96.54 (±0.02)
ANN	89.34 (±0.90)	89.19 (±0.63)	94.80(±0.17)	94.62 (±0.48)

Table 6
RF average results within the P7 and P6 protocols.

RF classifier	Protocol with 7 classes protocol	Protocol with 6 classes
Average Accuracy	84.95	94.68
Average F-measure	84.01	94.53
Average Kappa	75.40	90.69
Average AUC	92.49	98.57

Table 7
RF average results depending on knowing or not the consumer type.

RF classifier	Protocol with Known Consumer	Protocol with Unknown Consumer
Average Accuracy	89.86	89.77
Average F-measure	89.31	89.26
Average Kappa	83.12	82.97
Average AUC	95.51	95.55

The same applies to the F-measure. This 10% difference of performance improvement for the 6-classes mechanism is since the seventh class (theft6) is difficult to detect. The confusion matrices provided in Table 8 and Table 9 for respectively P7C and P6C confirm the difficulty to detect the seventh class, which impacted the overall detection capability of P7C. The matrix shows a huge confusion between the "normal" class (the first class) and the "theft6" class (the seventh class). From a total number of 35,413 of the actual "theft6" type, only 1119 were correctly detected while 23,071 were detected as the "normal" class and 11,223 were detected as "theft1", "theft3", "theft4" and "theft5". The confusion between the "normal" class and "theft6" class may be explained by the intrinsic definition of that theft since the fraud in this particular type is just a consumption with a small time-swift, i.e., a fraudulent consumption from a previous timestamp that replaced the actual real consumption.

On the other hand, having the model not considering the "theft6" class, the performance results were significantly enhanced by a 10% accuracy difference over the P7C. Such performance improvement is exhibited in the second confusion matrix presented inTable 9. Between the two confusion matrices (Table 8 and Table 9), we remark that some errors persist, especially the confusion between the "theft1" class and the "theft3" class. Indeed, these two types are very similar since fraudulent consumption numbers in both types were computed by multiplying the real consumption by a coefficient between 0.1 and 0.8. The slight difference is that the "theft1" coefficient is constant while the "theft3" coefficient is hour-dependent.

Furthermore, another experimentation was conducted with the inclusion or not of the consumer type using the RF technique. The average results of this experimentation are shown in Table 7.

Table 8
Confusion matrix for RF classifier for the P7C protocol.

RF Classifier: 7 Classes – Known Consumer – 5CV								
Accuracy = 85.00%		Predicted						
		Normal	Theft1	Theft2	Theft3	Theft4	Theft5	Theft6
Actual	Normal	317,480	1	0	3	1	45	14,294
	Theft1	86	39,461	0	7403	1107	103	2923
	Theft2	0	0	22,958	0	0	0	0
	Theft3	76	14,091	0	25,473	2082	169	2458
	Theft4	4	354	0	454	38,429	7	2212
	Theft5	30	7	0	3	1	31,642	1870
	Theft6	23,071	3427	0	2712	2822	2262	1119

Table 9
Confusion matrix for RF classifier for the P6C protocol.

RF Classifier: 6 Classes – Known Consumer – 5CV							
Accuracy = 94.71%		Predicted					
		Normal	Theft1	Theft2	Theft3	Theft4	Theft5
Actual	Normal	331,770	0	0	1	0	53
	Theft1	91	41,832	0	7835	1214	111
	Theft2	0	0	22,958	0	0	0
	Theft3	82	15,130	0	26,712	2232	193
	Theft4	3	347	1	454	40,647	8
	Theft5	35	8	0	6	0	33,504

Accordingly, the experimental evaluation metrics for the two mentioned mechanisms (with and without consumer type, respectively), were as follows: 89.86% vs. 89.77% for the average Accuracy, 89.31% vs. 89.26% for the average F-measure, 83.12% vs. 82.97% for the average Kappa and, 95.51% vs. 95.55% for the average AUC. As demonstrated, the results of the two mechanisms are very close, suggesting the robustness of the general model (without consumer type) and its capacity to give high classification rates regardless of the consumer type. This independence (i.e.autonomous aspect) from the type of consumer represents an advantage for the authorities, especially when a consumer type changes or when new types are added. The results presented in Tables 2–7 are shown as histograms in the appendix available at the end of this article (see Figs. 9–14).

Moreover, the relevance of the general model is confirmed by another experimentation which consists of computing several individual models. The individual models are calculated differently

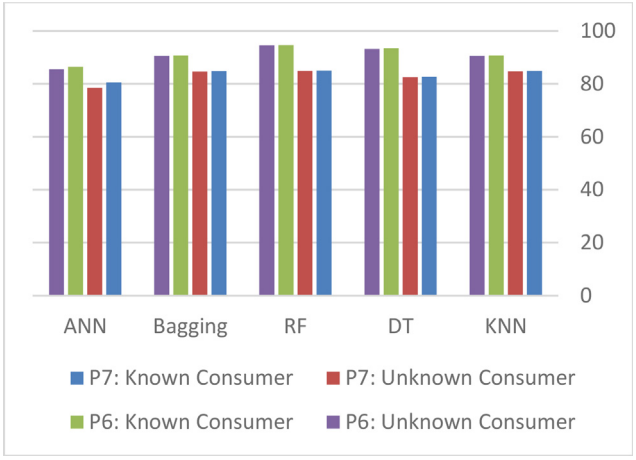

Fig. 9. Accuracy results of all classifiers.

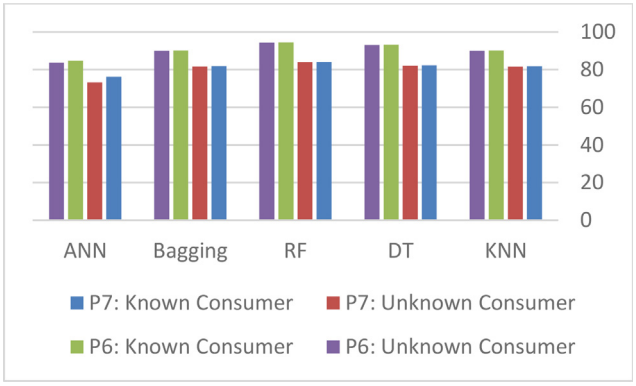

Fig. 10. F-measure results of all classifiers.

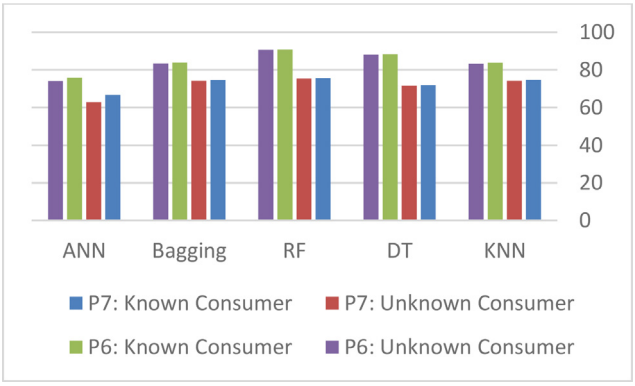

Fig. 11. Kappa results of all classifiers.

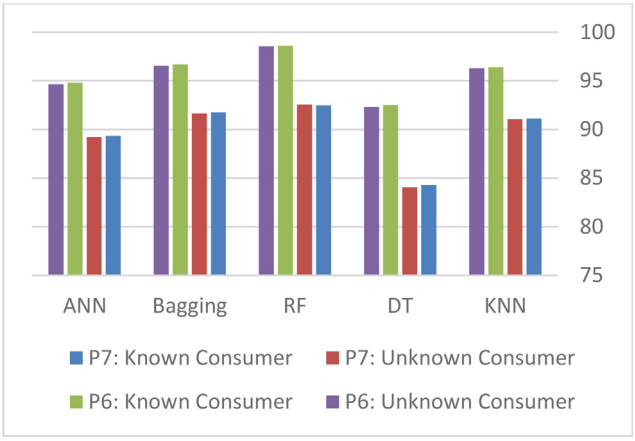

Fig. 12. AUC results of all classifiers.

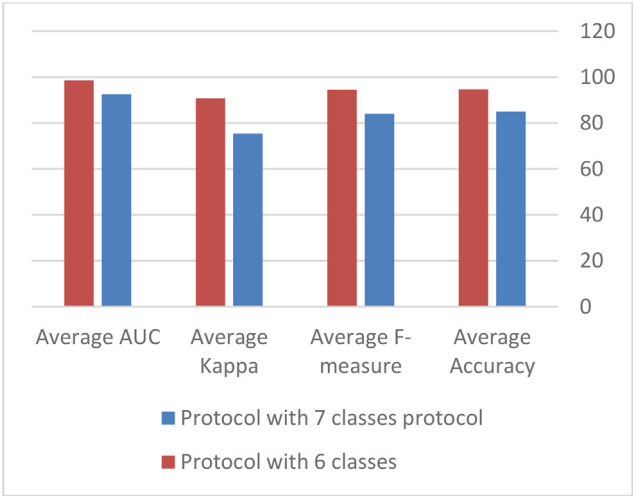

Fig. 13. RF average results within the P7 and P6 protocols.

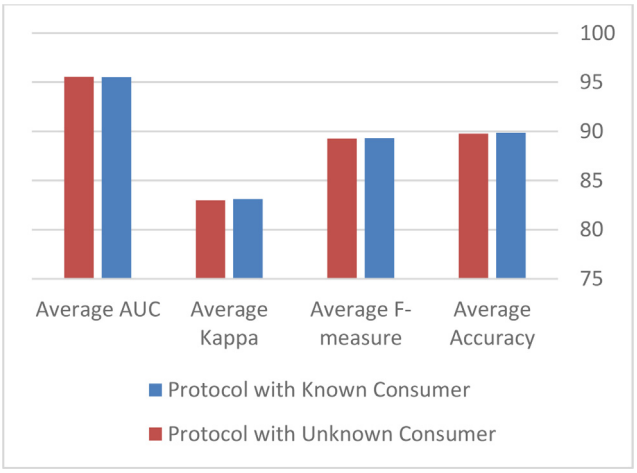

Fig. 14. RF average results depending on knowing or not the consumer type.

from all the previous models. In fact, for each consumer type, a new model is trained with exclusive data related to that type. This experiment has led to the training and testing of 16 individual models corresponding to the 16 consumer types (refer to Table 1). Due to the considerable performance of the RF model, which recorded the highest performance results in all previous experiments, it was selected and used as the base-classifier for these individual models. The same approach is also tested when dealing with the output classes (i.e.7-classes and 6-classes mechanisms are applied). The results of the individual models are shown for all the consumer types in Table 10. For instance, considering the "Full-service-restaurant" type, the related individual model has scored 84.71%for Accuracy and 84.02% for F-measure in the P7 mechanism. In addition, this example model scored 95.39% and 95.29% for the Accuracy and the F-measure in the P6 protocol.

To sum up, Table 11 provides the results of the three main models: the general model without the consumer type, the general model with the consumer type, and the individual models (weighted average). Based on the information stated in the table, the results of all models are very close. For the three approaches, the Accuracy is respectively equal to 84.89%, 85%, 85% for the 7-classes protocol and 94.64%, 94.71%, 94.87% for the 6-classes protocol. This finding confirms the previous results that demonstrate the general approach's robustness and relevance.

5. Conclusion

In this paper, a comprehensive and novel dataset for electricity theft detection (ETD2022), classification, and benchmarking in a smart grid environment has been developed, generated, and evaluated. ETD2022 is a multi-class dataset composed of 560,640 instances, each with 11-numerical features and one categorical feature used for the target class. ETD2022 is generated for the energy consumption of 16 different types of consumers, and it can be used to perform binary classification (normal vs. anomaly) or multi-classification (Normal, Theft1, Theft2, Theft3, Theft4, Theft5, and Theft6). To evaluate the proposed ETD2022 dataset, intelligent autonomous electricity theft detection has been developed on ETD2020 using five machine learning techniques: KNN, RF, DT, Bagging, and ANN. To gain more insights into the new dataset and the solution approach, the proposed theft detection-based ML models have been applied to four different classification mechanisms (scenarios), including P7C (7-classes with known consumers), P6C (6-classes with the known consumer), P7U (7-classes with unknown consumers), and P6U (6-classes with the unknown consumer). Also, four standard performance

Table 10
Individual models results.

Consumers' individual models based on the RF classifier and 5CV								
Number of classes	7 classes				6 classes			
Consumer/Metric	Accuracy	F-measure	Kappa	AUC	Accuracy	F-measure	Kappa	AUC
Full-service-restaurant	84.71	84.02	75.04	92.60	95.39	95.29	91.88	98.76
Hospital	84.91	84.21	75.95	93.64	95.27	95.13	91.93	98.81
Large-hotel	84.06	83.33	74.77	91.18	93.79	93.64	89.51	98.18
Large-office	86.08	85.20	76.97	92.69	94.50	94.41	90.32	98.54
Medium-office	84.67	83.43	74.01	94.22	94.82	94.72	90.53	98.64
Mid-rise-apartment	80.48	79.48	68.30	89.82	91.19	90.73	84.55	96.76
Outpatient	84.48	83.82	75.19	92.88	95.34	95.29	92.00	98.94
Primary-school	85.80	85.02	76.27	93.42	96.02	95.94	92.83	99.12
Quick-service-restaurant	85.11	84.37	75.57	92.24	94.73	94.57	90.73	98.43
Secondary-school	84.81	84.16	75.57	92.88	95.93	95.80	92.92	98.98
Small-hotel	84.88	84.15	74.94	92.18	94.08	94.01	89.52	98.31
Small-office	87.01	85.92	79.14	93.80	96.30	96.22	93.71	99.10
Standalone-retail	85.89	84.85	77.03	92.65	95.44	95.32	92.10	98.77
Strip-mall	86.34	85.42	77.48	92.35	95.56	95.43	92.17	98.75
Supermarket	85.74	84.93	76.59	92.72	95.35	95.16	91.84	98.80
Warehouse	85.04	83.75	75.23	93.11	94.23	94.11	89.79	98.49
Weighted Average	85.00	84.13	75.50	92.65	94.87	94.74	91.02	98.59

Table 11
Comparison between all the trained models: general models and individual models.

Number of classes	7 classes				6 classes			
Models/Metric	Accuracy	F-measure	Kappa	AUC	Accuracy	F-measure	Kappa	AUC
Average of the General Model (Unknown Consumer)	84.89	83.95	75.31	92.54	94.64	94.50	90.63	98.55
Average of the General Model (Known Consumer)	85.00	84.06	75.49	92.44	94.71	94.56	90.74	98.58
Weighted Average of Individual Models	85.00	84.13	75.50	92.65	94.87	94.74	91.02	98.59

indicators have been used to measure the effectiveness of the developed models, including the detection Accuracy, the harmonic mean (F-measure), the Kappa index, and the area under the curve (AUC). We noticed that the classification systems involving 6-classes classifiers (normal, theft1, theft2, theft3, theft4, and theft5) have recorded a significant improvement in performance trajectories over the classification systems involving 7-classes classifiers (normal, theft1, theft2, theft3, theft4, theft5, and theft6). This is due to the large similarity between the 'normal' class and the 'theft6' class which exhibits difficulty in detecting the 'theft6' class and in turn affects the overall system efficiency. Therefore, our proposed system is highly recommended for deployment in the application requiring our 6-classes model. Eventually, the empirical results demonstrated the superiority of the theft detection-based RF model, which recorded the best performance indicators by more than 10% improvement over the other ML-based models. This is because our RF model represents an ensemble learning which can make better predictions and achieve better performance than any single contributing model.

Therefore, RF has been selected and used as the base classifier for individual electricity theft detection models. Hence, the proposed dataset can efficiently build an automated machine learning-based electricity theft identification in a smart grid environment. In the future, we will work more on developing more distinguishing instances of the 'theft6' class that provide better discrimination from the 'normal' class to strengthen the efficiency of our 7-class classification system. Also, more efforts can be devoted to the data engineering and feature selection processes in order to improve the recognition results and improve the inferencing overhead. Moreover, as we expect to obtain more computational resources in the near future, we will attempt to employ deep learning techniques to obtain more accurate and precise models, where, indeed, this was not valid at this time due to the limitation of the computational resources available in our commodity computers.

Ethical approval

This article does not contain any studies with human participants or animals performed by the authors.

Data Availability

Dataset related to this article can be found at link1: https://data.mendeley.com/datasets/c3c7329tjj/1 or link2: http://dx.doi.org/10.17632/c3c7329tjj.1 an open-source online data repository hosted at Mendeley Data (Salah et al., 2022).

Conflicts of interest

Authors declare no conflict of interest.

Acknowledgment

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

References

"A random forest guided tour | SpringerLink." https://link.springer.com/article/10.1007/s11749-016-0481-7 (accessed Feb. 18, 2022)..

Adil, M., Javaid, N., Qasim, U., Ullah, I., Shafiq, M., Choi, J.-G., 2020. LSTM and bat-based RUSBoost approach for electricity theft detection. Appl. Sci. 10 (12), 4378.

Adil, M., Javaid, N., Ullah, Z., Maqsood, M., Ali, S., Daud, M.A., 2021. Electricity Theft Detection Using Machine Learning Techniques to Secure Smart Grid. Complex, Intelligent and Software Intensive SystemsCham, 233–243. https://doi.org/10.1007/978-3-030-50454-0_22.

Ahmad, T., Chen, H., Wang, J., Guo, Y., 2018. Review of various modeling techniques for the detection of electricity theft in smart grid environment. Renew. Sustain. Energy Rev. 82, 2916–2933.

Ahuja, R., Chug, A., Gupta, S., Ahuja, P., Kohli, S., 2020. Classification and clustering algorithms of machine learning with their applications. In: Yang, X.-S., He, X.-S. (Eds.), *Nature-Inspired Computation in Data Mining and Machine Learning*. Springer International Publishing, Cham, pp. 225–248. https://doi.org/10.1007/978-3-030-28553-1_11.

Amin, S., Schwartz, G.A., Cardenas, A.A., Sastry, S.S., 2015. Game-theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure. IEEE Control Syst. Mag. 35 (1), 66–81.

Arango, L.G., Deccache, E., Bonatto, B.D., Arango, H., Pamplona, E.O., 2017. Study of electricity theft impact on the economy of a regulated electricity company. J. Control Autom. Electr. Syst. 28 (4), 567–575. https://doi.org/10.1007/s40313-017-0325-z.

Bohani, F.A., Suliman, A., Saripuddin, M., Sameon, S.S., Md Salleh, N.S., Nazeri, S., Mandeep, J.S., 2021. A comprehensive analysis of supervised learning techniques for electricity theft detection. J. Electr. Comput. Eng. 2021, 1–10.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Cao, Y., Li, Q., Tan, Y.i., Li, Y., Chen, Y., Shao, X., Zou, Y., 2018. A comprehensive review of Energy Internet: basic concept, operation and planning methods, and research prospects. J. Mod. Power Syst. Clean Energy 6 (3), 399–411.

Cárdenas, A.A., Amin, S., Schwartz, G., Dong, R., Sastry, S., 2012, A game theory model for electricity theft detection and privacy-aware control in AMI systems, pp. 1830–1837..

Chen, Y.-C., Gieseking, T., Campbell, D., Mooney, V., Grijalva, S., 2019. A hybrid attack model for cyber-physical security assessment in electricity grid. IEEE Texas Power and Energy Conference (TPEC) 2019, 1–6. https://doi.org/10.1109/TPEC.2019.8662138.

Chuwa, M.G., Wang, F., 2021. A review of non-technical loss attack models and detection methods in the smart grid. Electr. Power Syst. Res. 199,. https://doi.org/10.1016/j.epsr.2021.107415 107415.

Daily, F., 2013. The first high-tech smart meter electricity theft case in China reported solved..

Gholinejad, H.R., Loni, A., Adabi, J., Marzband, M., 2020. A hierarchical energy management system for multiple home energy hubs in neighborhood grids. J. Build. Eng. 28, 101028.

González, S., García, S., Del Ser, J., Rokach, L., Herrera, F., 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Inf. Fusion 64, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007.

Gul, H., Javaid, N., Ullah, I., Qamar, A.M., Afzal, M.K., Joshi, G.P., 2020. Detection of non-technical losses using SOSTLink and bidirectional gated recurrent unit to secure smart meters. Appl. Sci. 10 (9), 3151.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. "KNN Model-Based Approach in Classification", in On The Move to Meaningful Internet Systems,CoopIS. *DOA, and ODBASE*, Berlin, Heidelberg 2003, 986–996. https://doi.org/10.1007/978-3-540-39964-3_62.

Hasan, M.N., Toma, R.N., Nahid, A.-A., Islam, M.M.M., Kim, J.-M., 2019. Electricity theft detection in smart grid systems: A CNN-LSTM based approach. Energies 12 (17), 3310.

Henriques, H.O., Barbero, A.P.L., Ribeiro, R.M., Fortes, M.Z., Zanco, W., Xavier, O.S., Amorim, R.M., 2014. Development of adapted ammeter for fraud detection in low-voltage installations. Measurement 56, 1–7.

Henriques, H.O., Corrêa, R.L.S., Fortes, M.Z., Borba, B.S.M.C., Ferreira, V.H., 2020. Monitoring technical losses to improve non-technical losses estimation and detection in LV distribution systems. Measurement 161,. https://doi.org/10.1016/j.measurement.2020.107840 107840.

Hussain, S. et al., 2021. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Rep. 7, 4425–4436.

Jadidbonab, M., Mohammadi-Ivatloo, B., Marzband, M., Siano, P., 2020. Short-term self-scheduling of virtual energy hub plant within thermal energy market. IEEE Trans. Ind. Electron. 68 (4), 3124–3136.

Jamil, F., Ahmad, E., XXXX. An Economic Investigation of Corruption and Electricity Theft, p. 19..

Jiang, R., Lu, R., Wang, Y., Luo, J., Shen, C., Shen, X., 2014. Energy-theft detection issues for advanced metering infrastructure in smart grid. Tsinghua Sci. Technol. 19 (2), 105–120. https://doi.org/10.1109/TST.2014.6787363.

Jokar, P., Arianpoo, N., Leung, V.C., 2015. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans. Smart Grid 7 (1), 216–226.

Jokar, P., Arianpoo, N., Leung, V.C.M., 2016. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans. Smart Grid 7 (1), 216–226. https://doi.org/10.1109/TSG.2015.2425222.

Karnouskos, S., Terzidis, O., Karnouskos, P., 2007. An advanced metering infrastructure for future energy networks. In: Labiod, H., Badra, M. (Eds.), New Technologies, Mobility and Security. Springer Netherlands, Dordrecht, pp. 597–606.

Kataria, A., Singh, M.D., XXXX. A Review of Data Classification Using K-Nearest Neighbour Algorithm..

Khan, Z.A., Adil, M., Javaid, N., Saqib, M.N., Shafiq, M., Choi, J.-G., 2020. Electricity theft detection using supervised learning techniques on smart meter data. Sustainability 12 (19), 8023.

Kotsiantis, S.B., 2013. Decision trees: a recent overview. Artif. Intell. Rev. 39 (4), 261–283. https://doi.org/10.1007/s10462-011-9272-4.
Leite, J.B., Mantovani, J.R.S., 2016. Detecting and locating non-technical losses in modern distribution networks. IEEE Trans. Smart Grid 9 (2), 1023–1032.
Maamar, A., Benahmed, K., 2018. Machine learning Techniques for Energy Theft Detection in AMI, in Proceedings of the 2018 International Conference on Software Engineering and Information Management, New York, NY, USA, 2018, pp. 57–62. doi: 10.1145/3178461.3178484..
Machová, K., Barčák, F., Bednár, P., XXXX. A Bagging Method using Decision Trees in the Role of Base Classifiers..
Marzband, M., Azarinejadian, F., Savaghebi, M., Pouresmaeil, E., Guerrero, J.M., Lightbody, G., 2018. Smart transactive energy framework in grid-connected multiple home microgrids under independent and coalition operations. Renew. Energy 126, 95–106.
Messinis, G.M., Rigas, A.E., Hatziargyriou, N.D., 2019. A hybrid method for non-technical loss detection in smart distribution grids. IEEE Trans. Smart Grid 10 (6), 6080–6091. https://doi.org/10.1109/TSG.2019.2896381.
Mian Qaisar, S., 2020. Event-driven coulomb counting for effective online approximation of Li-ion battery state of charge. Energies 13 (21), 5600.
Mohammad, Hossin, Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. IJDKP 5 (2), 01–11.
Mujeeb, S., Javaid, N., 2019. ESAENARX and DE-RELM: Novel schemes for big data predictive analytics of electricity load and price. Sustain. Cities Soc. 51, 101642.
Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. 18 (6), 275–285. https://doi.org/10.1002/cem.873.
Nabil, M., Is, M., Mahmoud, M., Shahin, M., Qaraqe, K., Serpedin, E., 2018. Deep recurrent electricity theft detection in AMI networks with random tuning of hyper-parameters. In: *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 740–745. https://doi.org/10.1109/ICPR.2018.8545748.
Nazari-Heris, M., Mirzaei, M.A., Mohammadi-Ivatloo, B., Marzband, M., Asadi, S., 2020. Economic-environmental effect of power to gas technology in coupled electricity and gas systems with price-responsive shiftable loads. J. Clean. Prod. 244, 118769.
Neto, E.A.A., Coelho, J., 2013. Probabilistic methodology for Technical and Non-Technical Losses estimation in distribution system. Electr. Power Syst. Res. 97, 93–99.
Nizar, A.H., Dong, Z.Y., 2009. Identification and detection of electricity customer behaviour irregularities, in *2009 IEEE/PES Power Systems Conference and Exposition*, pp. 1–10. doi: 10.1109/PSCE.2009.4840253..
Northeast Group LLC, 207. Electricity Theft and Non-Technical Losses: Global Markets, Solutions, and Vendors..
Pereira, J., Saraiva, F., 2020. A Comparative Analysis of Unbalanced Data Handling Techniques for Machine Learning Algorithms to Electricity Theft Detection, in *2020 IEEE Congress on Evolutionary Computation (CEC)*, Jul. 2020, pp. 1–8. doi: 10.1109/CEC48606.2020.9185822..
Punmiya, R., Choe, S., 2019. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. IEEE Trans. Smart Grid 10 (2), 2326–2329. https://doi.org/10.1109/TSG.2019.2892595.
Razavi, R., Gharipour, A., Fleury, M., Akpan, I.J., 2019. A practical feature-engineering framework for electricity theft detection in smart grids. Appl. Energy 238, 481–494. https://doi.org/10.1016/j.apenergy.2019.01.076.
Salah, Z., Alaeddine, M., Qaisar, M., Krichen, S., Abu, M., Qasem, A., 2022. Theft detection in smart grid environment. Mendeley Data V1. https://doi.org/10.17632/c3c7329tjj.1.
Savian, F.d.S., Siluk, J.C.M., Garlet, T.B., do Nascimento, F.M., Pinheiro, J.R., Vale, Z., 2021. Non-technical losses: A systematic contemporary article review. Renew. Sustain. Energy Rev. 147. https://doi.org/10.1016/j.rser.2021.111205.
Shaik, A.B., Srinivasan, S., 2019. A Brief Survey on Random Forest Ensembles in Classification Model, in International Conference on Innovative Computing and Communications, Singapore, pp. 253–260. doi: 10.1007/978-981-13-2354-6_27..
Singh, S.K., Bose, R., Joshi, A., 2018. Minimizing Energy Theft by Statistical Distance based Theft Detector in AMI. Twenty Fourth National Conference on Communications (NCC) 2018, 1–5. https://doi.org/10.1109/NCC.2018.8600016.
Sun, M., Konstantelos, I., Strbac, G., 2016. C-vine copula mixture model for clustering of residential electrical load pattern data. IEEE Trans. Power Syst. 32 (3), 2382–2393.
Sun, M., Wang, Y., Strbac, G., Kang, C., 2018. Probabilistic peak load estimation in smart cities using smart meter data. IEEE Trans. Ind. Electron. 66 (2), 1608–1618.
Tao, J., Michailidis, G., 2020. A statistical framework for detecting electricity theft activities in smart grid distribution networks. IEEE J. Sel. Areas Commun. 38 (1), 205–216. https://doi.org/10.1109/JSAC.2019.2952181.
Viegas, J.L., Esteves, P.R., Melício, R., Mendes, V., Vieira, S.M., 2017. Solutions for detection of non-technical losses in the electricity grid: A review. Renew. Sustain. Energy Rev. 80, 1256–1268.
Wang, S.-C., 2003. Artificial neural network. In: Wang, S.-C. (Ed.), Interdisciplinary Computing in Java Programming. Springer US, Boston, MA, pp. 81–100.
Wang, Y., Chen, Q., Kang, C., Xia, Q., 2016. Clustering of electricity consumption behavior dynamics toward big data applications. IEEE Trans. Smart Grid 7 (5), 2437–2447.
Wang, K., Yu, J., Yu, Y., Qian, Y., Zeng, D., Guo, S., Xiang, Y., Wu, J., 2018. A survey on energy internet: architecture, approach, and emerging technologies. IEEE Syst. J. 12 (3), 2403–2416.
Yip, S.-C., Wong, K., Hew, W.-P., Gan, M.-T., Phan, R.-C.-W., Tan, S.-W., 2017. Detection of energy theft and defective smart meters in smart grids using linear regression. Int. J. Electr. Power Energy Syst. 91, 230–240. https://doi.org/10.1016/j.ijepes.2017.04.005.
Yurtseven, Ç., 2015. The causes of electricity theft: An econometric analysis of the case of Turkey. Util. Policy 37, 70–78.
Zanetti, M., Jamhour, E., Pellenz, M., Penna, M., Zambenedetti, V., Chueiri, I., 2017. A tunable fraud detection system for advanced metering infrastructure using short-lived patterns. IEEE Trans. Smart Grid 10 (1), 830–840.
Zhang, Z., 2018. Artificial neural network. In: Zhang, Z. (Ed.), Multivariate Time Series Analysis in Climate and Environmental Research. Springer International Publishing, Cham, pp. 1–35. https://doi.org/10.1007/978-3-319-67340-0_1.
Zhang, W., Dong, X., Li, H., Xu, J., Wang, D., 2020. Unsupervised detection of abnormal electricity consumption behavior based on feature engineering. IEEE Access 8, 55483–55500. https://doi.org/10.1109/ACCESS.2020.2980079.
Zheng, K., Chen, Q., Wang, Y., Kang, C., Xia, Q., 2018. A novel combined data-driven approach for electricity theft detection. IEEE Trans. Ind. Inform. 15 (3), 1809–1819.
Zheng, Z., Yang, Y., Niu, X., Dai, H.-N., Zhou, Y., 2017. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans. Ind. Inform. 14 (4), 1606–1615.