Archives of Computational Methods in Engineering (2024) 31:4879–4892
https://doi.org/10.1007/s11831-024-10137-z

REVIEW ARTICLE

AI Techniques in Detection of NTLs: A Comprehensive Review

Rakhi Yadav[1] · Mainejar Yadav[2] · Ranvijay[3] · Yashwant Sawle[4] · Wattana Viriyasitavat[5] · Achyut Shankar[6,7,8,9]

Received: 27 February 2023 / Accepted: 22 April 2024 / Published online: 13 June 2024

Abstract

In the operation of power grid, worldwide, non-technical losses (NTLs) occur in a massive amount of proportion which is observed up to 40% of the total electric transmission and distribution losses. These dominant losses severely affect to adverse the performance of all the private and public distribution sectors. By rectifying these NTLs, the necessity of establishing new power plants will automatically be cut down. Hence, NTLs have become a critical challenge to do research in this emerging area for researchers of power systems due to the limitations of the current methodologies to detect and fix up these prominent type of losses. The existing survey so for basically contains the detail of identification of non-technical losses by machine and deep learning methods while this paper is a complete trouble shooting to resolve this issue by systematic approach. To address this, causes of NTLs along with its impact on economies and types of NTLs are elaborated in various countries. In addition, we have also prepared a comparative analysis based on several essential parameters. Further, implementation process of detection of NTLs or electricity theft based on Machine Learning or Deep Learning has also been demonstrated. Moreover, major challenges of detection of NTLs or electricity theft based on ML and Deep Learning, and its possible solutions are also described. Hence, definitely this comprehensive survey will help to the leading researchers to reach a new height in this thrust area.

1 Introduction

The overall evolution of any nation depends on the electricity sector. Eventually, both private and public sectors are deteriorating due to non-technical losses (NTLs) and becomes a challenging issue in front of developing countries. Electricity theft by the fraud consumers is the prime root of the NTLs. However, there are many other causes to occur NTLs [1] i.e. errors in meter reading, record keeping, accounting, broken or faulty infrastructure. But, electricity theft contributes a large proportion to count NTLs [2]. Due to electricity theft, power distribution utilities not only have the financial losses in terms of revenue [3, 4] but also these utilities pay a very high cost of electricity theft by failing in

✉ Achyut Shankar
ashankar2711@gmail.com

Rakhi Yadav
yadavrakhi.87@gmail.com

Mainejar Yadav
rahulit1210@gmail.com

Ranvijay
ranvijay@mnnit.ac.in

Yashwant Sawle
yashsawle@gmail.com

Wattana Viriyasitavat
wattana@cbs.chula.ac.th

[1] Electrical Engineering Department, MANIT Bhopal, Bhopal, MP, India

[2] Computer Science & Engineering Department, REC Sonbhadra, Sonbhadra, India

[3] Computer Science & Engineering Department, MNNIT Allahabad, Prayagraj, India

[4] Electrical Engineering Department, MITS Gwalior, Gwalior, India

[5] Chulalongkorn University, Pathum Wan, Thailand

[6] Department of Cyber Systems Engineering, WMG, University of Warwick, Coventry, CV74AL, UK

[7] Department of CSE, University Centre for Research & Development, Chandigarh University, Punjab 140413 Mohali, India

[8] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India

[9] School of Computer Science Engineering, Lovely Professional University, 144411 Phagwara, Punjab, India

safe and reliable operation of the power system. Detection of NTLs and electricity theft can be done by using hardware device or by generated data or through the random inspection visit in the suspected consumers' premise. Identification of NTLs and electricity theft by developing hardware is very costly and time consuming while data based detection in done by mining and analysis the data through the artificial intelligence techniques such as ML and deep learning which is very economic and having fast processing. After the evolution of advanced metering infrastructure (AMI), attacks on smart meter to modify its data by data storage device and communication technology between consumers and control center is a challenging issue to detect NTLs and electricity theft. Meanwhile, smart meter generates a massive amount of data of various consumers at real time of intervals, this generated massive data has accelerated the research in the field of NTLs by AI and motivates us to write this paper for identification and minimization of these losses.

The major contributions of this paper are summarized as follows:

- This paper has detailed information about various ML and deep learning models which are currently employed to detect NTLs and electricity theft.
- Discussion about the causes of NTLs along with its impact on economies in different countries.
- A comparative analysis of existed solutions for detection of NTLs and electricity theft after extensive survey of literature's.
- Implementation process of detection of NTLs or electricity theft based on Machine Learning or Deep Learning.
- Discussion about major challenges of detection of NTLs based on ML and Deep Learning along with its possible solutions.
- Possible future research directions.

Organization of this paper is as follows: Sect. 2 explains the state of the art. Implementation process of detection of NTL or electricity theft based on Machine Learning or Deep Learning is described in Sect. 3. Sections 4 and 5 describe the challenges for applying the Artificial Intelligence technique to detect and to reduce the non-technical losses followed by possible solutions. Finally, the conclusion and future research directions are described in Sect. 6.

2 Literature Review

Authors in Refs. [5, 6] have provided the general survey of the NTLs in which such type of losses are the main factors of cheating identification. In addition, the expert system and machine learning technique have been discussed for the detection of NTLs. However, the novelty of the proposed survey is not only limited to the review along with comparison of the reported works of literature but also discusses some significant challenges and its possible solutions for applying the Artificial Intelligence techniques to detect the NTLs.

2.1 Causes of NTLs

For NTLs occurrence, both contractual and irregular consumers are responsible. There are many noticeable practices which are the root causes of NTLs i.e. bypassing the meters for fraud purposes, any fault or damage in infrastructure, unpaid bills of public and private electricity sectors [3]. In [7, 8] authors have described that some poor consumers cannot pay the electricity bill. While, sometimes, consumers do not pay electricity bill intentionally.

2.2 Economic Effects of NTL

NTLs in electric utilities, report the total loss in amount of income. Consequently, extra charges are implied on honest customers. One example of indirect economic effect of NTls is inspection cost as more losses lead to more inspections in the customer's premises which further causes unreliability in the network. In order to maintain the system's reliability, running cost is invested to furnish the indirect effects. In [9], the authors have described that the expenditures for the detection and prevention NTLs is a very massive amount than the actual return rate.

2.3 Variation of NTLs

The proportion of the NTLs in various countries is shown in Table 1.

2.4 Features

The machine learning techniques work on the dataset's features such as electricity consumption records of consumers by smart meters. The collected dataset from the electric distribution utility has many features, but we have taken some of the essential features primarily used in literature-related works.

Table 1 Proportion of the NTLs country-wise

References	NTL Proportion (%)	Country
Golden [2]	1.6–37.9	UP, India
Yurtseven [7]	4–73	Turkey
Mwaura [10]	18	Rwanda
Ramos [11]	3–40	Brazil
Katiyar [12]	Up to 70	India

2.4.1 Monthly Consumption

Some of the researchers have worked on NTL detection by using the traditional meter. Based on the conventional meter, we have calculated the average consumption by reading data which is used as a feature in the previous studies [13–17]. Based on six months' meter reading, the average consumption, standard deviation, maximum consumption and number of inspections have been calculated [18].

2.4.2 Smart Meter Consumption

In [19, 20], authors have used the consumption features of 15 min' intervals, whereas intervals of 30 min have been used in [21]. In [22, 23] have used the maximum consumption in any 15 min. In [24], the shape factor is used as a feature obtained from the consumption data, including the effect of weekends and different intervals of time during the whole day.

2.4.3 Master Data

Authors in [15] have used the features such as business class (e.g., residential, or industrial), location (city and neighborhood), number of phases (1, 2, or 3), voltage (110 V or 200 V), and meter type from the master data. In [23], the consumer's total demand in kw of installed equipment and the demand contracted are used as the features. In [17], the following features are used, i.e., the type of voltage, the contracted power, the electricity tariff, the number of phases (1 or 3), and location. In [22], information of the power transformer of a connected customer has been used as a feature.

2.4.4 Creditworthiness

In [13], authors have used each user's creditworthiness (CWR) as a feature. The range of this feature is 1 to 5, depending on the customer's ignorance or delaying bill payments. In addition, this feature provides detailed information on the customer's income, payment performance, and financial conditions. Apart from the above-discussed features, which are used in the existing works of literature, some additional features can be used to improve the accuracy of the electricity theft or NTL detection model. These features are discussed in the following subsections.

2.4.5 Healthy Consumer Flag (HCF)

HCF feature is also considered to differentiate between ordinary and defaulter consumers. The value of this flag dynamically changes after every three months.

2.4.6 Meter Read Remark (MRR)

It shows the four different types of status such as Identified Defective (IDF), Appear Defective (ADF), Reading Defective (RDF) and Ceiling Defective (CDF) of the installed meter on the consumer side. For example, the CDF status of meter reading indicates the overloading consumption (consumer's use is more than the sanctioned load).

2.4.7 Late Payment Surcharge (LPSC)

This feature is used to recognize the behavior of the consumers. For example, the amount of LPSC shows the deadlines or timelines of paying the electricity bill. The delay in paying the electricity bill consequently affects the NTLs, so this feature is also remarkable. To analyze the consumer's load pattern, the previous month's consumption, the current month's consumption, and the next month's consumption are used as a feature. Consumers who have crossed the maximum limit of load i.e., overload and MDI (maximum demand indicator) are enlisted and can also be used as feature. If the value of MDI is more than the sanctioned load, it is considered as overload. The overload indicates the excess use of electricity than the sanctioned load, which comes under electricity theft. In addition, excess electricity' consumption also shows how much electricity is lost due to electricity theft. Consumer's behavior is also observed from the last paid electricity bill date and the payment patterns. Some other features such as multiply factors, are also used. All the above features are used to train the model to detect electricity theft or NTLs.

Above discussed features are applicable when the dataset is taken from traditional meter. Moreover, in the case of dataset taken from SMART METER, the following additional features are useful to remark:

- Types of consumers: Residential, Commercial (Hospital, Industry, Office, Super Market etc.) [25]
- Types of electrical appliances: Heating, Cooling, Fan, etc. [25]
- Load: Sanctioned load for consumers [25]

2.5 Expert Systems and Fuzzy Systems

Nagi et al. have proposed a fuzzy inference system based approach for detecting the NTLs [14]. This approach has used around data of 100k consumers for training and testing the model. The recall value achieved by this model is 0.72. Werley et al. have introduced a fuzzy c-means clustering based approach to detect the abnormalities in consumer's consumption [18]. This approach has basically used the consumer's consumption pattern along with inspection data. All the consumers are grouped into different classes based on the features using fuzzy c-means clustering method. After that, consumers are classified into two catagories i.e theft and normal using euclidean method. The precision value achieved by this approach is 0.745. Glaune has proposed a NTLs detection approach based on ML and expert system [26]. The flowchart of this approach is shown in Fig. 1. In this method, initially, dataset is collected from the distribution utilities. This collected dataset is pre-processed and the selection of useful features is done. After that, model is trained by using training dataset and is tested by using trained dataset. After prediction of the fraud consumers, the targeted consumers are selected for inspection after holographic spatial visualization. In [27], first of all, authors have categorized the dataset on the basis of irregular and regular customers. Thereafter, it is used for training, in which, a neuro-fuzzy hierarchical system is used for classification. This approach uses neural network for optimizing the fuzzy membership parameters and achieves.682 accuracy.

2.6 Support Vector Machines

Support vector machines (SVM) are supervised learning models which are utilized to analyze data for classification and regression. Thereafter, the data is splitted into classes on account of features and creates a maximum marginal hyperplane (MMH). The perpendicular distance between the closest data points of two categories is known as the margin. The high value of the margin is good for the classification. SVM classifiers offer excellent performance with high dimensional space. A SVM requires high training time. Hence, it is not suitable for large datasets. Moreover, SVM is not more efficient with overlapping classes.

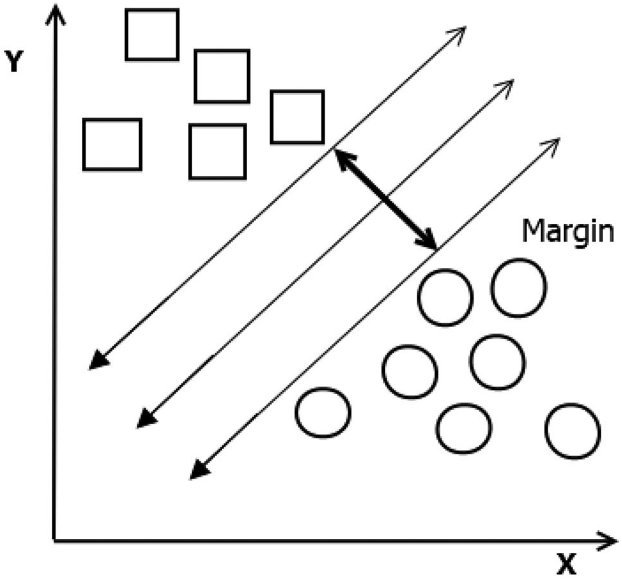
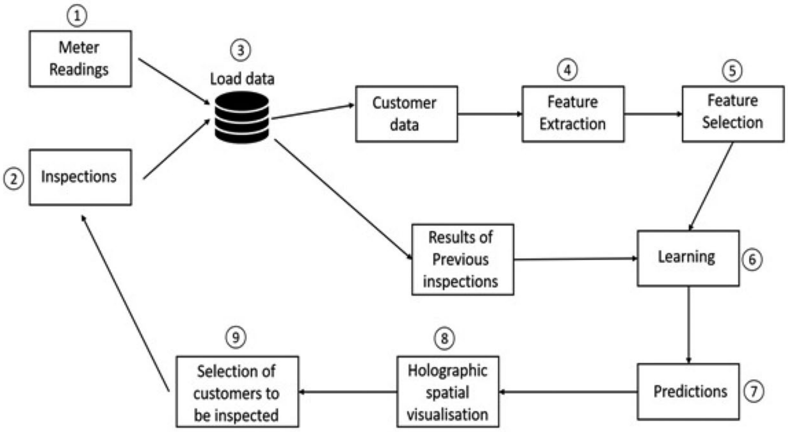


Fig. 2 Marginal hyperplane of SVM

The selection of parameters, including kernel parameters, is a difficult task, which is very useful to analyse the performance of SVM. Face recognition, text categorization, biometrics, image classification, detection of NTL, and electricity theft are significant SVM applications. Figure 2 shows the hyperplane of SVM corresponding to a dataset which has two different categories of objects such as square bracket and circle. This figure shows that the SVM classifier separates these two objects with a maximum margin. Whenever, any new object or test dataset is applied for prediction, the learn model quickly identifies the new objects' category. Negi et al. [13] have introduced an SVM-based approach to overcome the above issue. This approach has achieved an accuracy 0.86. In [20], the authors have used Indian data set of size 1350 customers and applied SVM. This data set is split into 135 different patterns with a size of 10k customers where each customer has reading time for 15 min. This

Fig. 1 NTLs detection system based on ML and expert knowledge



work achieves the test accuracy of 0.984. The improved and updated version of this work is presented in reference [28], which has accuracy 0.92. In [29], authors have used high-performance computing to improve previous work performance [28]. The test accuracy of this model is 0.89.

2.7 Deep Learning

2.7.1 Deep Learning in the Context of ML

Machine Learning (ML) algorithms are the result of various deep learning algorithms. ML algorithms convert complex concepts into simpler ones. various deep learning technologies are discussed below.

Supervised Learning There are different training sets of inputs and outputs [30]. Supervised Learning algorithms club input and output together. This methodology uses several models, i.e., Feedforward Neural Networks, Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM). These models work as function approximators which is made by stacking various hidden layers together. Hochreiter and Schmidhuber [31] have proposed to resolve the issue of diminishing gradients by maintaining a more constant error by using gated cells, due to which continuous learning is possible for large time steps. Nazmul Hasan et al. have proposed a CNN and LSTM based approach to detect electricity theft. In this approach, 10k consumer's data were used for training and testing the model. The accuracy of this approach was 89%.

Unsupervised Learning Unsupervised learning develops the models for collecting essential facts from high-dimensional sensory unlabeled data. The visual cortex motivates to do research in this direction and requires a minimal amount of labeled data. The deep Belief Networks (DBNs) [32, 33] have helped in learning several layers of nonlinear features in an unsupervised manner. DBNs are built by stacking several Restricted Boltzmann Machines (RBMs) [34, 35].

2.7.2 Benefits and Challenges in Deep Learning

In CNN models, the convolution layers extract the features from the raw data. Input is convolved with learned filters, while pooling layers reduce the dimension over previous convolution layers. An object is identified by feature extraction in a robotic system [27] and scene classification [36]. The latest research is mainly focused on resolving the scene classification problem. Deep learning methods can generalize particular sets of labeled input data. From raw inputs, i.e., images, LIDAR (Light Detection and Ranging) sensor data and deep learning help us to know about the actual pattern. However, deep learning suffers from various challenges, i.e., the lack of geometrical information of the objective function in deep neural networks. That's why performance of certain architecture is still unanswered.

2.8 Other Methods

Costa et al. have introduced ANN based electricity fraud detection method [15]. In this method, data sets has size 22K customers for training a neural network. It has used customer's features such as types of customers, location, voltage, meter reading period, and the average consumption of the last 12 months. The accuracy of this approach is 0.8717. Shehzad et al. [37] have provided an electricity theft detection method based on artificial bee colony and genetic algorithm. In this technique, auto-encoder is used for feature selection. This method reduces ML classifiers' overfitting, storage, and computational overhead. By this method, we achieve accuracy 0.87 and 0.80 on two different datasets. Ullah et al. [38] have introduced a hybrid model based on CNN, particle swarm optimization (PSO), and gated recurrent unit (GRU). This method has focused on two major issues i.e. data unbalancing and overfitting. The parameters are tuned by using a metaheuristic model with GRU and CNN. This metaheuristic model helps to resolve the overfiting problem while the SMOTE algorithm is used to overcome the data-unbalancing problem. Li et al. [39] have introduced a hybrid model to detect electricity theft in 2019. This model was used to cash the benefit of CNN and the random forest algorithm. Moreover, this study has also covered the unbalancing data problem through SMOTE algorithm. In addition, the overfitting problem can also be resolved by adding a dropout layer.

Xia et al. [40] have provided a hardware-based solution for the identification of malicious users in the smart grid. This approach was focused on designing and characterizing the types of equipment that detect and estimate fraudulent activity. In [24], some searching techniques have used for customer's master data like greedy, genetic, best-first search (BFS), and complete search. Some other features are also included, such as shape factors which are derived from the consumption time series. A decision tree is used for the classification, which predicts whether the customer causes the NTL or not. The test accuracy of 0.9997 is achieved. Ahir and Chakraborty have introduced an electricity theft detection approach for residential consumers [41]. This method is based on the usage patterns of the consumers. The F1 score of this approach is 94%. This research aims to mitigate the issues of theft-based NTL. The description about neural network (NN) is available in [42]. Nizar et al. have provided a Extreme learning machines (ELM) based NTLs detection approach [43]. The accuracy of this approach is 0.546. Muniz et al. have introduced a NN based another method for detection of irregularity of electricity [44]. This approach has used around 20k consumers data for training and testing of the model. The accuracy achieved by this approach is 0.686. Saddam Hussain et al. have introduced a CatBoost-based model to detect electricity theft. In this approach, data unbalancing problem is resolved by using SMOTETomek algorithm. The accuracy of this model is 0.93. Huang and Xustacked have provided an electricity theft detection model based on sparse denoising autoencoder [45]. Kong et al. have proposed a decision tree, KNN and SVM based model to detect electricity theft [46]. This model had achieved more than 90% accuracy.

Alaeddine et al. [47] have provided a dataset for electricity theft. In this dataset, six different types of theft categories have been discussed. This research also shows the numerous outcomes of different machine learning classifiers on this dataset. This dataset is available publically; it can be used for electricity theft, Further, it can also be compared with other techniques for the same on this dataset. Shehzad et al. have proposed a meta-heuristic technique to detect electricity theft [48]. This approach has focused on over-fitting and computational overhead of used classifiers. This approach is tested on two different datasets i.e. PRECON and SGCC and has achieved accuracy is 0.86 and 0.80, respectively. Kocaman and Tumen [49] have proposed an electricity theft detection method based on LSTM. In this approach, data preprocessing has been used for making the useful dataset. The accuracy of this approach is 93.60 ± 1.22. de Souza et al. have introduced a Multi-layer Perceptron Artificial Neural Network (MP-ANN) [50] based approach to detect electricity theft. The accuracy of this approach is 93%. In [51], a feature-engineering framework has been used, which is a combination of Finite Mixture Model clustering (FMMC) and Genetic Programming (GP). Viegas have proposed a fuzzy Gustafson-Kessel clustering based model to detect NTLs [52]. The AUC score of this approach is 0.741.

Zidi et al. [25] have introduced an ML-based electricity theft detection technique. They have also provided a smart meter dataset that can be used for bench-marking. The dataset is a collection of 1-year data of the 16 types of users of 24 h. In this data, there is normal data and six types of thefts: theft 1, theft 2, theft 3, theft 4, theft 5, and theft 6. Along with all the theft or normal, there is 7 classes are formed. In this paper, KNN (K-Nearest Neighbour), Random Forest (RF), Decision Trees (DT), Bagging technique, and Artificial Neural Networks (ANN) classification methods are used. The Random Forest gives the best result out of all the techniques, 94.71% for known classes and 94.64% for unknown classes. Xia et al. [53] have proposed another electricity theft detection method. This technique uses the hybrid model by using wide CNN and deep CNN. The dataset is an imbalanced dataset, taken from SGCC (China power supply company). The dataset shows the daily consumption of 42,372 consumer's records from January 2014 to October 2016. Lagrange polynomials are used for data pre-processing. The pre-processed data is divided into 1-D data in days and 2-D data in weeks and provide the 1-D data to the wide component through a fully connected layer and 2D data to the CNN with dilated convolution. It uses the focal loss as the loss function and sigmoid to classify the users. This approach has an AUC score of 0.8361 and F1-Score of 0.5372. Shi et al. [54] have proposed an improved Transformer Neural Network (TNN) based approach. In this method, authors have focused on data imbalancing and overfitting problems along with its solutions by using weighted cross entropy and focal loss. For the feature visualization, the author has used tDistributed Stochastic Neighbour Embedding (t-SNE) algorithm. It achieves the highest accuracy of 98.23%. Using various ML techniques, Fei et al. [55] have proposed an electricity fraud detection in low-voltage networks. The dataset has been taken from the SGCC. This dataset contains three types of theft and three types of normal, which are as follows (Theft I, Normal I, Theft II, Normal II, Theft III, Normal III). This approach has used various machine learning techniques such as SVM; Extreme Gradient boosted trees

(XGB), Random Forests (RF), Logistics Regression (LR), Multi-layer PerceptronNetworks (MLP), and Wide & deep CNN. The authors have found the result using the ND-CP model with an accuracy of 77%.

2.9 Summary

A comparison of different models based on data sets and different performance measurement parameters has been shown in Table 2. The popularity of other models is also analyzed, which concludes that the support vector machine and neural network model are the most popular techniques. A wide range of users are used in the data sets, like 30 to 700K. In addition, expert systems, genetic methods, regression methods, and OPF have been used. The most famous performance measurement parameters are recall and accuracy, which range from 0.29 to 1 and 0.45 to 0.99, respectively. In some works, author's have used precision as a performance measure ranging from 0.51 to 0.85. Glauner [26] has studied and analyzed the most frequently used machine learning techniques. The popularity's of the ML techniques used to detect NTL from 2007 to 2016 are support vector machines and neural networks. At the same time, clustering and genetic-based algorithms are seldomly used while SVM and neural networks are more popular. The identification of challenges and its solutions have been discussed in Sects. 4 and 5.

2.10 Simulation Environments

In the literature, different simulation environments have been used, which are listed here-

Table 2 Comparative analysis of existed approach

References	Model	Accuracy	Recall	Precision	AUC	Number of Customers	Dataset source
[13]	SVM	0.86	0.77	–	–	400	–
[14]	SVM+ Fuzzy	–	0.72	–	–	100k	TNBD
[15]	NN	0.87	0.29	0.65	–	22k	–
[18]	Fuzzy	0.745	–	–	–	–	Brazil
[19]	SOM	0.93	0.98	0.85	–	2k	–
[20]	SVM(Gauss)	0.98	–	–	–	1350	–
[21]	Fuzzy logic	0.55	–	–	–	–	TNBD
[44]	NN	0.835	–	–	–	–	Brazil
[56]	CatBoost	0.93	–	–	–	–	SGCC
[57]	CNN, LSTM	0.89	0.87	0.90	-	17,120	–
[49]	LSTM	0.93	–	–	–	-	–
[50]	SOM MP-ANN	0.934	–	–	–	5000	–
[52]	Fuzzy clustering	–	–	–	0.741	2k	–
[58]	Wide and Deep CNN	0.9404	–	–	–	42,372	SGCC
[59]	DT coupled SVM	0.925	–	–	–	N/A	N/A
[60]	Neuro-fuzzy	0.68	–	0.51	–	20k	–
[61]	K-mean-SVM	–	0.94	–	–	5000	CER in Ireland
[25]	RF	0.9471			0.924	560K	Open Energy Data initiative
[53]	Wide and Deep CNN	–	–	–	0.836	42K	SGCC
[54]	CA-TNN	0.9823	–	–	–	5k	IRISH CER SMART
[55]	ND-CP	0.77	–	–	0.832	20K	SGCC

- LIBSVM MATLAB and WEKA
- ORACLE, C++

Moreover, the electricity theft detection model and detection of NTL model can be simulated on Anaconda (Jupyter notebook) software using the python programming language. Different libraries, such as Scikit-learn, Numpy, Pandas, etc., are also used in this simulation process. The efficiency and effectiveness of the models are checked by using different evaluation metrics which are discussed in Sect. 5.1.

3 Implementation Process of Detection of NTL or Electricity Theft Based on Machine Learning or Deep Learning

Here, we have discussed the detailed procedure involved in electricity theft or NTLs detection model's approach. The dataset is collected from the service provider and it is used as an input in the NTL or electricity theft detection model. This raw dataset is analyzed based on consumer types and weather types. This dataset has many features, and the models select appropriate features based on performance reflection value. The collected dataset is not directly used because it could have many anomalies. So, before using this dataset, data preprocessing is required. Data preprocessing involves several jobs such as removing outliers and data transformation. The number of faulty consumers are significantly less than the number of normal consumers in the dataset, which generates the unbalancing. Therefore, the performance of supervised ML approaches can be biased. Hence, some algorithms, such as Synthetic Minority Oversampling Technique (SMOTE) algorithm, is applied to put these consumers in an equilibrium condition. The dataset is split into train and test using some algorithms such as cross-validation algorithm. After that, the model is imported through Scikit-learn libraries and is trained on the training dataset. By using the testing dataset, we test the trained models. We give the data to the model, then the model predicts the output. Thereafter, the accuracy is calculated based on the model's predicted value and its true value.

The flowchart for implementation process of detection of NTL or electricity theft based on Machine Learning or Deep Learning is shown in Fig. 3.

- Pre-Requisite: For implementing the ML and deep learning-based methods for predicting electricity theft or NTL, one should have knowledge of the python programming language along with the python build libraries such as Pandas, NumPy, Seaborn, Matplotlib, etc. Moreover,

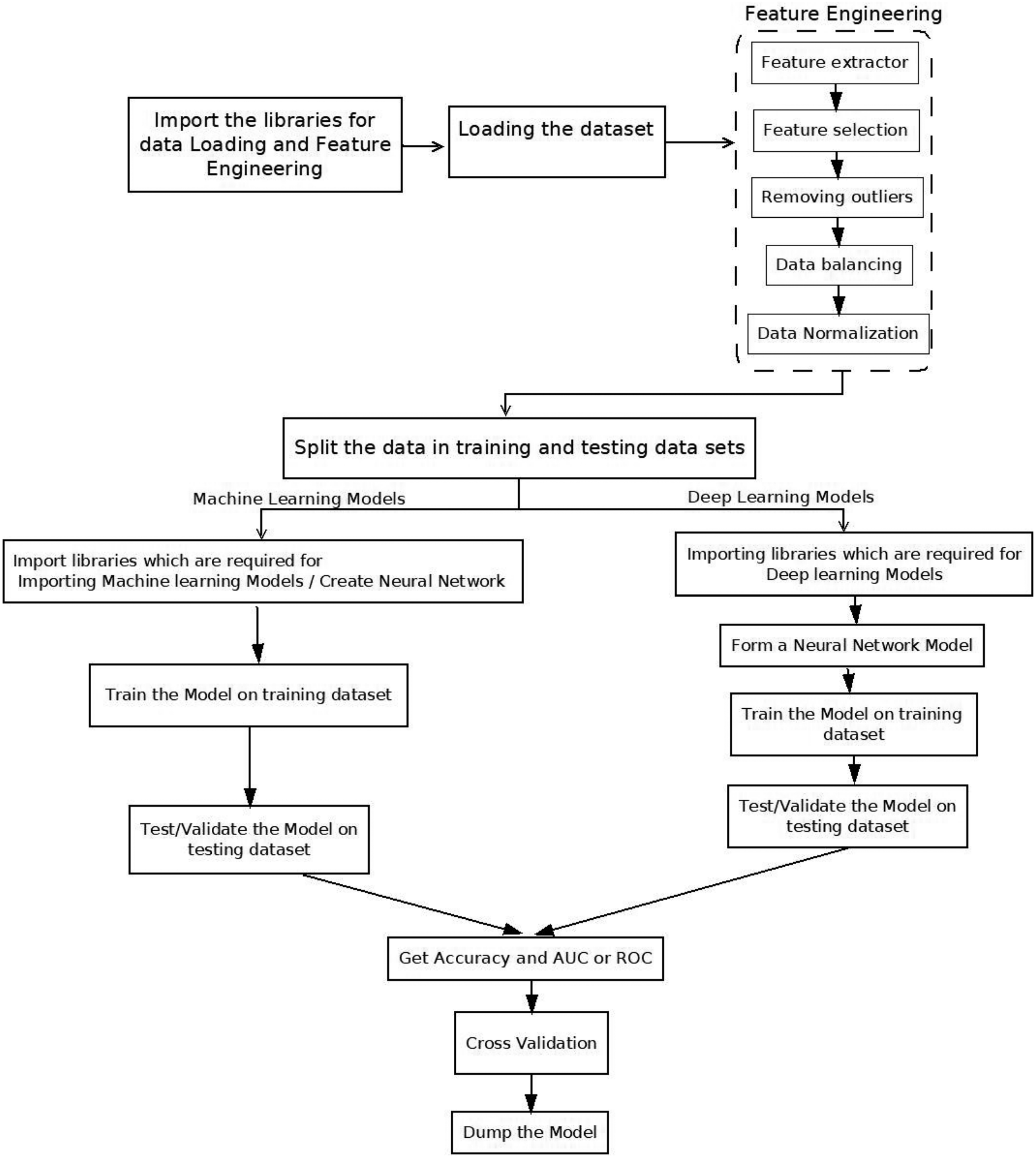

Feature Engineering
Import the libraries for data Loading and Feature Engineering
Loading the dataset
Feature extractor
Feature selection
Removing outliers
Data balancing
Data Normalization
Split the data in training and testing data sets
Machine Learning Models
Deep Learning Models
Import libraries which are required for Importing Machine learning Models / Create Neural Network
Importing libraries which are required for Deep learning Models
Form a Neural Network Model
Train the Model on training dataset
Train the Model on training dataset
Test/Validate the Model on testing dataset
Test/Validate the Model on testing dataset
Get Accuracy and AUC or ROC
Cross Validation
Dump the Model

Fig. 3 Flowchart: implementation process of detection of NTL or electricity theft based on Machine Learning or Deep Learning

knowledge of Tensorflow, Keras API, and the machine learning models such as Decision Tree, Random Forest, etc., are also required. Further, Jupyter notebook or Google Collab can be used as per requirement.

- Dataset loading: Some essential libraries which are required for importing the dataset in their jupyter notebook i.e. pandas.

  For Example: df = pd.read_csv (“file loacation”);
- Feature Engineering: Feature extraction is done by drawing the correlation graph. This graph helps to check the dependency of feature with each other. After that, we decide dependent features and independent features. Based on the dependence, It is determined the necessity of features which are required to predict the result.

  Outliers Removing: First of all, we detect the outliers from the data set using ‘Z-SCORE’ and ‘IQR’ methods. The rows which having outliers, we remove those rows completely. The ‘Z-SCORE’ is defined as-

$$Z_{score} = \frac{(X - x)}{\sigma} \tag{1}$$

  where X, x and σ are standardized random variable, mean and standard deviation.

  Data balancing: For data balancing, it is checked that the classes on which we have done classification are balanced or not because there is a possibility of a biased model due to an imbalanced dataset. For example: in electricity theft detection, we check the count of normal and theft data is equal or not; if these numbers of data are not equal, then we apply data balancing. Data balancing is categorized into two parts, i.e., Under Sampling and Over Sampling. If we have enough data, we do undersampling; if we have less data, we use over-sampling. SMOTE and ADASYN etc. are used for data balancing.

Data Normalization Data normalization is done when the same data is not required for every model, but it is beneficial when we apply a deep learning model for training. Data normalization methods are min-max scaling and Standardization scaling, which is computed as:

$$X_{scaled} = \frac{X - min(X)}{max(X) - min(X)} \tag{2}$$

- Split the dataset: In this step, the dataset is split into two parts i.e., training data and testing data, by using a scikit-learn library. After that, the K-fold method is used for selecting the dataset rows from different locations in the split dataset. After applying this method, we have a dataset with different values. The training dataset contains all the different variations in the data, and the model is trained on all those variations.
- Machine learning models: Machine learning is of basically three types: supervised, unsupervised, and reinforcement. Based on the requirement, we select supervised or unsupervised and reinforcement learning. The data type decides the prediction used for the classification or regression in supervised learning. For example: In electricity theft detection, we use supervised learning classification algorithms such as K-nearest neighborhood, Decision Trees, Naïve Bayes, etc.

  Training Models: We import the model through Scikit-learn libraries and train the model on the training dataset and get the trained model.

  Testing models: By using the testing dataset, we test the trained models. We give the data to the model, then the model predicts the output. Thereafter, we check the correct answer of that specific data and calculate the accuracy.
- Deep learning models: The TensorFlow and Keras libraries are very essential to make their Neural Networks. Some of the Neural Networks are ANN, CNN and ResNet-50.

  Training neural networks: The Neural Network is trained on the training dataset and validate the model based on the validation dataset.

  Testing the neural networks: The accuracy of the NN is tested by using the testing dataset. The accuracy is calculated based on the model’s predicted value and its true value.
- Calculating the AUC and ROC: It is mainly used to measure the accuracy rate of the model which is trained. High value of the AUC indicates that model is better and it is more efficiently distinguishing between normal and theft types of consumers.
- Cross validation: Cross-validation is used to detect overfitting, i.e., failing to generalize a pattern.
- Dump the model: For dumping the model, the pickle and PyYMAL can be used in Machine Learning and Deep Learning, respectively.

4 Challenging Issues

The major challenges in detection of NTLs or electricity theft based on ML and Deep Learning are listed below:

- Data unbalancing
- Data quality
- Availability of the standard data-set
- Feature description and selection
- Non-malicious factors

The collected data-set has a significantly higher number of ordinary consumers than the defaulter or fraud consumers;

this data-set shows the class imbalance. The performance of the machine learning (ML)-based models may be affected due to class imbalance. The feature selection process plays a crucial role in ML-based approaches because it directly reflects the performance results of the models. Moreover, another challenging task is to handle the data-set because the size of the data-set is significant. Hence, it requires the high processing power of hardware and software. The data-set's quality also affects the model's performance, thus, maintaining the data-set's quality is also mandatory. The collection of the standard data-set and inspection data is a crucial challenge because of the unavailability of the data-set publicly. This data-set has consumers' private data such as an address, phone number, etc., due to which there is a possibility of breach of the user's privacy. Therefore, the service provider avoids the disclosure in providing the data-set to anyone.

5 Possible Solutions

After reviewing state-of-the-art research thoroughly, we have identified the challenges related to NTL detection. The methodology which can be applied to rectify these issues, is described as follows.

5.1 Handling Data Unbalancing and Evaluation Parameters

The accuracy of various models deteriorates due to the problem of class imbalance data. The data imbalance problem can be resolved by using oversampling or under-sampling techniques. Synthetic Minority Oversampling Technique (SMOTE) is one of the oversampling technique. The working of SMOTE can easily be understood by Fig. 4. In this figure, we can see that given data has two classes, i.e. minority and majority classes. Minority and majority classes are shown by red and black colors, respectively. Therefore, SMOTE first utilizes the KNN algorithm to produce synthetic data. After that, it chooses random data from the minority class, then k-nearest neighbors from the data are determined. Then, synthetic data is developed between the random data and the randomly selected k-nearest neighbor. The procedure is repeated several times until the minority class has the same proportion as the majority class. Hence, the bias-ness is removed, and the problem of data unbalancing is resolved. In the existing research, authors have used various evaluation parameters such as accuracy, recall, precision, and area under curve (AUC). The regions with high and low NTLs ratios show that NTLs have the class imbalance type problem. So, there is a need for more investigation for NTLs to report reliable and imbalance-independent results. In [26], the authors have suggested some more evaluation parameters like the rank of the customer, inspection cost, and possible revenue increments. In the NTLs detection process, the main objective is to increase the true positive rate (TPR) because the possible occurrences of NTL is directly proportional to the TPR. An increment in TPR consequently reduces the false positive rate (FPR), and less FPR ultimately reduces the costly number of inspections that are done for the surveillance. Therefore, we need a Receiver operating characteristic (ROC) curve, which shows the relationship between TPR against FPR and is used to detect NTLs. Another performance measure is the classification rate or accuracy. This shows how close the measured value is to the real value. The accuracy is defined in Eq. 3, where TP, FP, TN and FN are true positive, false positive, true negative and false negative, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

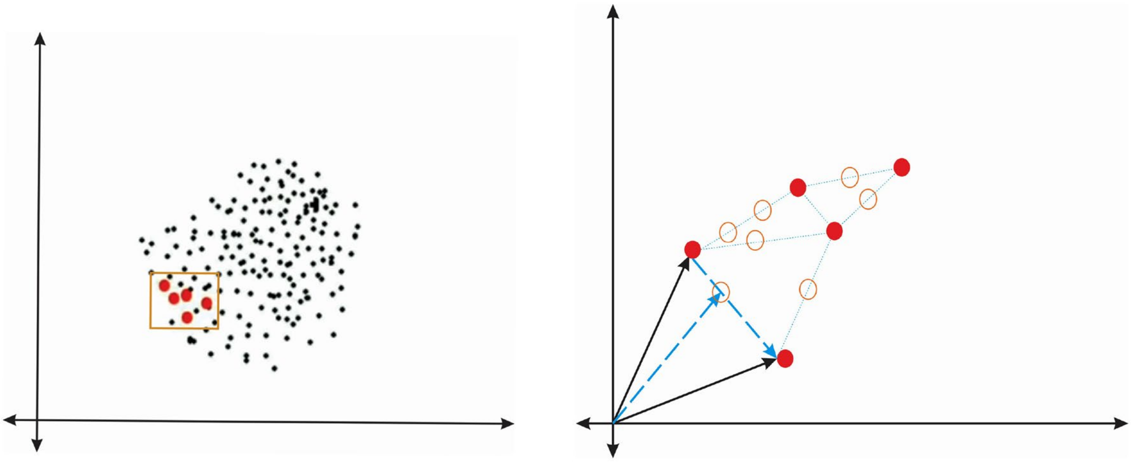


Fig. 4 Description of SMOTE algorithms

The recall is a metric that shows the fraction of retrieved relevant instances. The recall is also known as the True positive rate or sensitivity. It is defined by Eq. (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The specificity shows how well a classifier can identify true negatives. Specificity is also called as true negative rate (TNR). It is defined in Eq. (5).

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

The false-positive rate (FPR) is defined as

$$FPR = 1 - TNR \tag{6}$$

The area under the curve (AUC) is widely used as a performance metric defined in Eq. (7). The value of AUC more than 0.5 shows that the performance of that particular classifier is better than the random guessing.

$$AUC = \frac{Recall + Specificity}{2} \tag{7}$$

FPR must be low for NTL detection so that expenditure of onsite inspection in consumer's premises will be decreased, whereas an increment in TPR value shows all possible NTLs. Precision or positive predictive value (PPV) is a vital evaluation parameter in Eq. (8). The outstanding value of the PPV, with a perfect test, is 1.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

F1 score and Dominance parameters are also used for the performance assessment of models, which are calculated as:

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{9}$$

$$Dominance = TPR - TNR \tag{10}$$

ROC curve is another essential parameter for the assessment of the classification model. It shows the relationship between TPR and FPR. The structure of ROC curve is drawn in Fig. 5, where the dotted blue line indicates the performance of the random classifier with 0.5 probabilities, and the red line shows that the classifier is perfect. It shows the trade-off between the TPR and FPR for a predictive model using different probability thresholds.

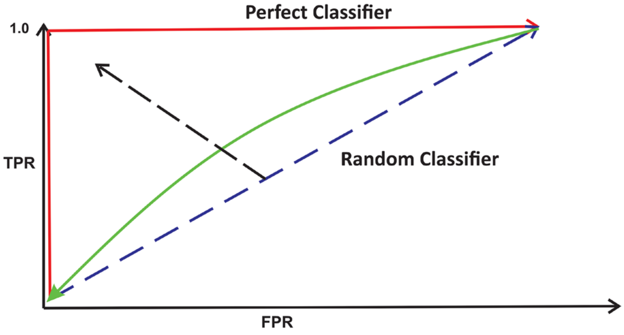


Fig. 5 ROC curve

5.2 Feature Description

The literature has observed that NTLs mostly use information primarily from the consumption time series. In [13, 16, 44, 60], authors have used the traditional meter data, whereas in [20, 21, 24, 29] the smart meter data has been used. Both meter types will co-exist in the next decade, and the results of those researches are not easily replaceable. Therefore, in [26], authors have suggested self-learning of features from the consumption time series.

5.3 Characteristics of Smart Meter Profiles for Millions of Consumers

In the works of literature, authors of [13] have used the few hundred of customers' data, whereas authors of [15, 60] have used ten thousand users' data. So, variations in the data-set require a high computation environment for computation efficiency. For this purpose, GPU is required which is introduced in [62], and map-reduce architecture is introduced in [63]. As time complexity of an algorithm defines the amount of time taken by an algorithm to run as a function of the length of the input. Hence for the detection purpose of NTLs, analysis of the data set of many customers over a long period is done. Consequently, the time complexity of learning algorithm becomes inefficient for large size data set; for such analysis, some scalable computing technology is Google Tensor Flow [62] or Apache Spark [15] must be applied which are capable of handling Big Data sets with efficient time complexity for NTLs detection.

5.4 Construction of Standard Data-Set

We have observed that different types of data-sets are being used. Then the question arises, how to compare different models? Reference [64], states "no free lunch theorem," which means there is no standard benchmark available, so we cannot compare various learning algorithms and are not capable to recognize which algorithm is superior to others. The solution to the above-stated problem is creating a common standard data set. All the researchers easily and freely get this data-set and can compare the proposed model with works of literature. This data set should fulfill the following properties:

- Different types of customers: Based on energy consumption, consumers have been categorized as residential and industrial.
- Number of customers and inspections
- Spread of customers across the geographical area
- Sufficiently long period of meter readings: consumption profile of the customer depends on the climate. The data-set's period should be available for at least one year for good results.

6 Conclusion and Future Research Directions

As non-technical losses (NTLs) are the dominating losses in the operation of the distributed power system. In this paper, we have enlisted complete information about the NTLs and their impact on the utility's total income return per annum. Furthermore, this paper provides a detailed illustration of Artificial Intelligence techniques, i.e., Machine Learning and Deep Learning techniques which are implemented to detect these losses. The data sets used in the literature range from 100 to more than one million consumers. To address this, in this survey, major challenges which the engineers are still facing during the operation of the power system are described. In addition, all possible solutions to detect NTLs and electricity theft are also described. After an extensive survey, we have also provided a tabulated comparative analysis based on several essential parameters. Moreover, this paper also includes the implementation process of detection of NTL or electricity theft based on Machine Learning or Deep Learning.

Hence, this comprehensive survey with myriad information will be very useful for researchers to do research in this thrust area. In the coming days, several possible research related to NTLs and electricity theft detection, i.e., Spatio-temporal behavior, can be modeled by interested researchers. The need for real-world datasets to provide it as a publicly available and expert system can be modeled to make decisions automatically. As earlier, visiting the consumer's premises, the service provider wants to know the causes of NTLs. As on-site physical inspections are not economical without particular consumer's history.

Data Availability Data sharing is not applicable to this paper as no datasets were generated during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Antmann P (2009) Reducing technical and non-technical losses in the power sector. World Bank, Washington, pp 1–35
2. Golden M, Min B (2012) Theft and loss of electricity in an Indian state. International Growth Centre, Tech. Rep. pp 1–38
3. Smith TB (2004) Electricity theft: a comparative analysis. Energy Policy 32(18):2067–2076
4. Dasgupta K, Padmanaban M, Hazra J (2017) Power theft localization using voltage measurements from distribution feeder nodes. IET Generation, Transmission & Distribution, pp 2831–2839
5. Saeed MS, Mustafa Mohd W, Hamadneh NN, Alshammari NA, Sheikh UU, Ahmed Jumani T, Bin KS, Abd KI (2020) Detection of non-technical losses in power utilities-a comprehensive systematic review. Energies 13(4727):1–25
6. Bolton Richard J, Hand David J (2002) Statistical fraud detection: a review. Statistical Science, pp 235–249
7. Yurtseven C (2015) The causes of electricity theft: an econometric analysis of the case of Turkey. Utilities Policy 37:70–78
8. Lewis FB (2015) Costly 'throw-ups': electricity theft and power disruptions. Electric J 28:118–135
9. Linares P, Rey L (2013) The costs of electricity interruptions in Spain. Are we sending the right signals? Energy Policy 61:751–760
10. Mwaura FM (2012) Adopting electricity prepayment billing system to reduce non-technical energy losses in Uganda: lesson from Rwanda. Utilities Policy 23:72–79
11. Ramos CC, Rodrigues D, de Souza AN, Papa JP (2018) On the study of commercial losses in Brazil: a binary black hole algorithm for theft characterization. IEEE Trans Smart Grid, pp 676–683
12. Katiyar SK (2005) Political economy of electricity theft in rural areas: a case study from Rajasthan. Econ Polit Wkly. pp 644–648
13. Nagi J, Yap KS, Tiong SK, Ahmed SK, Mohamad M (2010) Non-technical loss detection for metered customers in power utility using support vector machines. IEEE Trans Power Deliv. pp 1162–1171
14. Nagi J, Yap KS, Tiong SK, Ahmed SK, Nagi F (2011) Improving SVM-based non-technical loss detection in power utility using the fuzzy inference system. IEEE Trans Power Deliv. pp 1284–1285
15. Costa BC, Alberto BLA, Portela AM, Maduro W, Eler EO (2013) Fraud detection in electric power distribution networks using an ANN based knowledge-discovery process. Int J Artif Intell Appl 4(6):17–23
16. Glauner P, Boechat A, Dolberg L, State R, Bettinger F, Rangoni Y, Duarte D (2016) Large-scale detection of non-technical losses in imbalanced data sets. In IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp 1–5
17. Spirić JV, Stanković SS, Dočić MB, Popović TD (2014) Using the rough set theory to detect fraud committed by electricity customers. Int J Electr Power Energy Syst 62:727–734
18. Werley E, Angelos S, Saavedra OR, OA Carmona C, Souza Andr'e Nunes de (2011) Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Trans Power Deliv. pp 2436–2442
19. Cabral Jose E, Pinto Joao OP, Pinto Alexandra MAC (2009) Fraud detection system for high and low voltage electricity consumers based on data mining. In: IEEE Power & Energy Society General Meeting, pp 1–5
20. Shekara S, Reddy Depuru S, Wang L, Devabhaktuni V (2011) Support vector machine-based data classification for detection of electricity theft. In: Power systems conference and exposition (PSCE), pp 1–8

21. Nagi J, Yap KS, Nagi F, Tiong SK, Koh SP, Ahmed SK (2010) NTL detection of electricity theft and abnormalities for large power consumers in Malaysia. Student Conference, pp 202–206
22. Oba Ramos CC, Souza Andre Nunes De, Gastaldello DS, Papa JP (2012) Identification and feature selection of non-technical losses for industrial consumers using the software weka. In: Industry applications (INDUSCON), IEEE, pp 1–6
23. Ramos CCO, Sousa AN, Papa JP, Falcao Alexandre X (2011) A new approach for non-technical losses detection based on optimum-path forest. IEEE Trans Power Syst. pp 181–189
24. Nizar AH, Zhao JH, Dong ZY (2006) Customer information system data preprocessing with feature selection techniques for non-technical losses prediction in an electricity market. In: International conference on power system technology, IEEE, pp 1–7
25. Zidi S, Mihoub A, Qaisar SM, Krichen M, Al-Haija QA (2022) Theft detection dataset for benchmarking and machine learning-based classification in a smart grid environment. J Comput Inf Sci 35(1):13–25
26. Glauner P (2019) Artificial Intelligence for the detection of electricity theft and irregular power usage in emerging markets. PhD-FSTC-2019-07 The Faculty of Sciences, Tech. and Comm., Université du Luxembourg
27. Liu W, Anguelov D, Erhan D (2016) Ssd: Single shot multibox detector. In: Proceedings of the European conference on computer vision, Springer, pp 21–37
28. Reddy Depuru SSS, Wang L, Devabhaktuni V (2012) Enhanced encoding technique for identifying abnormal energy usage pattern. In North American Power Symposium (NAPS), IEEE. pp 1–6
29. Shekara S, Depuru SR, Wang L, Devabhaktuni V, Green RC (2013) High-performance computing for detection of electricity theft. Int J Electr Power Energy Syst 47:21–30
30. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
31. Hochreiter S, Schmidhuber J (1996) LSTM can solve hard long time lag problems. In Proceedings of the 10th annual conference on neural information processing systems. pp 473–479
32. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput. pp 1527–1554
33. Bengio Y, Lamblin P, Popovici D, Larochelle H (2007) Greedy layer-wise training of deep networks. Adv Neural Inf Process Syst 19:153–160
34. Smolensky P (1986) Information processing in dynamical systems: foundations of harmony theory. Tech. Rep, DTIC Document, pp 194–281
35. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800
36. Gangopadhyay A, Tripathi SM, Jindal I, Raman S (2015) SACNN: dynamic scene classification using convolutional neural networks. pp 1–18. arxiv:1502.05243
37. Shehzad F, Javaid N, Aslam S, Javed MU (2022) Electricity theft detection using big data and genetic algorithm in electric power systems. Electric Power Syst Res 209:107975
38. Ullah A, Javaid N, Sani Yahaya A, Sultana T, Ahmad Al-Zahrani F, Zaman F (2021) A hybrid deep neural network for electricity theft detection using intelligent antenna-based smart meters. Wirel Commun Mob Comput 9933111:1–19. https://doi.org/10.1155/2021/9933111
39. Shuan Li, Han Y, Yao Xu, Song Yingchen, Jinkuan Wang, Zhao Q (2019) Electricity theft detection in power grids with deep learning and random forests. J Electr Comput Eng. https://doi.org/10.1155/2019/4136874
40. Xia X, Xiao Y, Liang W (2020) SAI: a suspicion assessment-based inspection algorithm to detect malicious users in smart grid. IEEE Trans Inf Forensics Secur 15:361–374
41. Ahir RK, Chakraborty B (2022) Pattern-based and context-aware electricity theft detection in smart grid. Sustain Energy Grids Netw 32:100833
42. Bishop CM (1996) Neural networks: a pattern recognition perspective
43. Nizar AH, Dong ZY, Wang Y (2008) Power utility non-technical loss analysis with extreme learning machine method. IEEE Trans Power Syst. pp 946–955
44. Muniz C, Figueiredo K, Vellasco M, Chavez G, Pacheco M (2009) Irregularity detection on low tension electric installations by neural network ensembles. In International Joint Conference on Neural Networks. pp 2176–2182
45. Huang Y, Xu Q (2021) Electricity theft detection based on stacked sparse denoising autoencoder. Int J Electr Power Energy Syst 125:106448
46. Kong X, Zhao X, Liu C, Li Q, Dong D, Li Y (2021) Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM. Int J Electr Power Energy Syst 125:106544
47. Al-Haija SZAMSMQMKQA (2022) Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2022.05.007
48. Shehzad F, Javaid N, Aslam S, Javed MU (2022) Electricity theft detection using big data and genetic algorithm in electric power systems. Electric Power Syst Res 209:107975
49. Kocaman B, Tümen V (2020) Detection of electricity theft using data processing and LSTM method in distribution systems. Indian Acad Sci. https://doi.org/10.1007/s12046-020-01512-0
50. de Souza MA, Pereira JLR, de Guilherme O, de Alves BC, Oliveira Igor D, Melo Paulo AN (2020) Detection and identification of energy theft in advanced metering infrastructures. Electric Power Syst Res. https://doi.org/10.1016/j.epsr.2020.106258
51. Razavi R, Gharipour A, Fleury M, Akpan IJ (2019) A practical feature-engineering framework for electricity theft detection in smart grids. Appl Energy 238:481–494
52. Viegas JL, Esteves PR, Vieira SM (2018) Clustering-based novelty detection for identification of non-technical losses. Electr Power Energy Syst 101:301–310
53. Xia R, Gao Y, Zhu Y, Gu D, Wang J (2023) An attention-based wide and deep CNN with dilated convolutions for detecting electricity theft considering imbalanced data. Electric Power Syst Res 214:108886
54. Shi J, Gao Y, Dexi Gu, Li Y, Chen K (2023) A novel approach to detect electricity theft based on conv-attentional transformer neural network. Int J Electr Power Energy Syst 145:108642
55. Fei K, Li Q, Zhu C, Dong M, Li Y (2022) Electricity fraud detection in low-voltage networks with contrastive predictive coding. Int J Electr Power Energy Syst 137:107715
56. Hussain S, Mohd W, Mustafa TA, Jumani SK, Baloch HA, Khan I, Khan A (2021) A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Rep 7:4425–4436
57. Nazmul Hasan Md, Toma RN, Abdullah-Al Nahid MM, Islam M, Kim J-M (2019) Electricity theft detection in smart grid systems: a CNN-LSTM based approach. Energies 12(3310):1–18
58. Zheng Z, Yang Y, Niu X, Dai H-N, Zhou Y (2018) Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans Ind Inform 14:1606–1615
59. Jindal A, Dua A, Kaur K, Singh M, Kumar N, Mishra S (2016) Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Trans Ind Inform 12:1005–1016
60. Muniz C, Bernardes MM, Vellasco R, Tanscheit R, Figueiredo K (2009) A neuro-fuzzy system for fraud detection in electricity distribution. In IFSA/EUSFLAT Conf. pp 1096–1101

61. Jokar P, Arianpoo N, Leung VCM (2016) Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans Smart Grid 7:216–226
62. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado Greg S, Davis A, Dean J, Devin M (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. pp 1–19
63. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I, Spark (2010) cluster computing with working sets. Hot Cloud 10:1–10
64. Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Comput 8:1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.