A novel and effective method based deep learning model for detecting non-technical electricity losses

Israa Mohammed Ridha Baldawi[1], Timur İnan[2]
[1]Department of Information Technology, Altinbas University, Istanbul, Turkey
[2]Department of Software Engineering, Altinbas University, Istanbul, Turkey

Article Info

Article history:

Received Feb 12, 2023
Revised May 10, 2023
Accepted May 25, 2023

Keywords:

Classification
Deep learning
Machine learning
Non-technical electricity
Prediction

ABSTRACT

This study focused on non-technical electricity loss detection. As mentioned, non-technical losses (NTLs) affect utilities and economies financially. Electricity theft, fraud, and metering issues can create NTLs. NTL generate most distribution losses in electrical power networks, costing utilities a lot. NTL detection approaches are data-focused, network-oriented, or hybrid. Data-oriented writing dominated this analysis. After data collection and cleaning and labeling the unlabeled dataset with a target, a methodology was supplied that used four machine learning techniques random forest, decision tree, KNN, and logistic regression and four neural network models-DNN, CNN, CNN-LSTM, and CNN-GRU. The CNN and DNN model have the best accuracy, stability, fast learning, and training time.

Corresponding Author:

Israa Mohammed Ridha Baldawi
Department of Information Technology, Ultinbas University
Mahmutbey Mahallesi, Dilmenler Caddesi Mahmutbey Yerleşkesi, 34218 Bağcılar/İstanbul, Turkey
Email: shiningsun452@yahoo.com

1. INTRODUCTION

Worldwide, power providers struggle with the significant issue of energy loss throughout the transmission and distribution of electricity. Technical losses (TL) and non-technical losses (NTL) are the two main categories used to describe energy loss [1]–[3]. Power system components like transmission lines and transformers experience TL as a result of internal processes [4]–[7].

The NTL is calculated by subtracting the total losses (TL) from the total production (TP) and is mostly attributable to power theft. In addition to posing a threat to the reliability of the power grid, the fraudulent use of electricity might cost utilities money. For instance, fires might break out owing to the overloaded electrical systems caused by electricity theft [8]. Utilities face a significant challenge from non-technical electricity losses (NTL), which can result from a wide variety of causes such as human error during installation, tampering with meter readings through unauthorized database access, incorrect calculations of technical losses, meter fraud, a faulty meter, electricity theft via distribution lines, nonpayment by customers, billing errors, and so on. Not only can they result in significant revenue losses, but also, since they introduce uncertainty into the real consumption, they might have an impact on the operation of the power system [9].

When calculating the cost of power in an electrical grid [10], non-technical losses (NTL) are subtracted from the total to account for the energy that is lost as heat in the cables, transformers, and other components of the grid. Electricity theft, fraud, or inadequate metering assets can all lead to NTLs, which have a major financial impact on utilities and economies. NTL account for the largest share of distribution losses in electrical power networks worldwide, leading to substantial revenue losses for utilities. Customers' fraudulent actions are a major contributor to these losses, which in turn weaken network infrastructure and threaten grid security [11].

For this reason, academics are becoming increasingly interested in electrical fraud detection models since the (NTL) problem has become of crucial importance and a worry for everyone. Throughout US $96 billion is lost annually by utilities around the world owing to NTLs, according to a recent study [7]–[10]. Figure 1 demonstrates the severity of the NTL problem in various regions of the world, providing a visual representation of the aforementioned dilemma. The result in [12], the loss from NTLs is not confined to developing nations alone; it is projected that rich nations like the United Kingdom and the United States lose an equivalent of US $232 million and US $6 billion yearly.

Electricity distribution providers face a significant challenge in addressing non-technical losses (NTL). Therefore, these and other recent research efforts in the power sector are motivated by the need to find a long-term solution to this threat. The issues with (NTL) have been addressed by several authors. Assuming a strong association between power theft and its consumption patterns, Hussain *et al.* in [13] implemented some early work for NTL detection algorithms with supervised learning, dubbed support vector machines (SVM). Non-technical loss detection is addressed in [9], where Buzau *et al.* offer an approach based on the utilization of smart meter data and auxiliary databases as raw material for a supervised machine learning algorithm (XGBoost). The system was trained using data from customers who had at least one inspection performed.

The clustering-based novelty detection technique proposed by Viegas *et al*. [10] uses the gustafson-kessel fuzzy clustering algorithm to identify non-technical losses; this method is applicable to high-resolution consumption data obtained from smart meters. The best results were achieved with fuzzy clustering using the gustafson-kessel technique. Li *et al.* in [8], offer a unique CNN-RF model for spotting power theft. When looking into smart meter data, the CNN acts as an automatic feature extractor and the RF as an output classifier in this model. Using real-world energy consumption data, we conduct tests that demonstrate the superiority of the proposed detection model over state-of-the-art approaches. Fraud detection using semi-supervised deep learning was pioneered by Hu *et al.* in [1]. Their model is called MFEFD. MFEFD's potent feature extraction capacity stems from the model's deep structure and significant nonlinearity. To train MFEFD, we use a semi-supervised approach. Nagi *et al.* in [14], gave a comprehensive and thorough assessment and classification of the approaches studied for NTL detection in recent literature, outlining their advantages and disadvantages and serving as a one-stop resource for both novice researchers and seasoned professionals in the field. After proposing a hybrid neural network model for non-technical losses identification, Saeed *et al.* [15] demonstrated how the model's performance may be significantly enhanced with the incorporation of supplementary data. The approach was created and validated using actual smart meter data from Endesa, the largest electrical provider in Spain. In order to extract temporal patterns from a time series dataset, Buzau *et al.* [16] created a hybrid deep learning model that combines the strengths of GoogLeNet and gated recurrent unit (GRU). Meanwhile, the GoogLeNet is used to extract hidden patterns from the stacked EC dataset that is updated every week. More so, the temporal least square generative adversarial network (TLSGAN) was developed to address the issue of class disproportion. Support vector machine (LibSVM) was used to classify users and predict user's activities in terms of energy usage in order to detect fraudulent users and ensure their prompt disconnection from the grid in [12], where Aniedu *et al.* presented a solution to non-technical losses using machine learning techniques in conjunction with AMI technology.

By utilizing the ensemble bagged tree (EBT) method, Aniedu *et al.* [17] proposed a novel strategy for NTL identification in PDCs. Results showed that the EBT algorithm had a 93.1% accuracy rate for detecting NTLs, which was significantly greater than that of more traditional methods. Based on machine learning and feature engineering, Saeed *et al.* [18] proposed a model for the detection of non-technical losses; the model was developed using four different classifiers (logistic regression, support vector machine, decision tree, and random forest); and comparative analysis and evaluation with existing models demonstrated the proposed model's effectiveness and usefulness [19].

The fuzzy gustafson kessel clustering methodology, presented by Viegas *et al.* [10], is a method for identifying NTL that makes use of fuzzy logic. As a means of establishing the prototype employed in grading the NTL, the writers seek to understand users' consumption habits. The area under the curve (AUC) for this method is 0.741. In this paper, we present an approach based on feature engineering and deep learning to discover non-technical losses and reduce them to an absolute minimum, allowing the power provider to concentrate its efforts where they will have the greatest impact. The data from the actual mechanical meter is combined with other sources to create a more complete picture of the customer's consumption patterns, as well as to reveal more specifics about the meter's location and extraction of other features that can be used in calculations involving consumption rates. Baghdad governorate customer data was obtained from the dataset provided by the power provider in Iraq (residential and commercial residential category).

Then, these characteristics are fed into various model selection and assessment methods, including supervised ML and deep learning algorithms. All of the models have been developed, checked, and verified. There are five main parts to this study. The second part of this paper presents an overview of the various NTL detection methods and explains the data transformation approach proposed for electrical NTL detection. It is made up of the workflow design and instrumentation phase. The fourth section focus on applying the algorithms to real-world consumption data and simulating it with existing statistics from Iraq's electrical distribution corporation. The fifth segment provides the impetus for the conclusion.

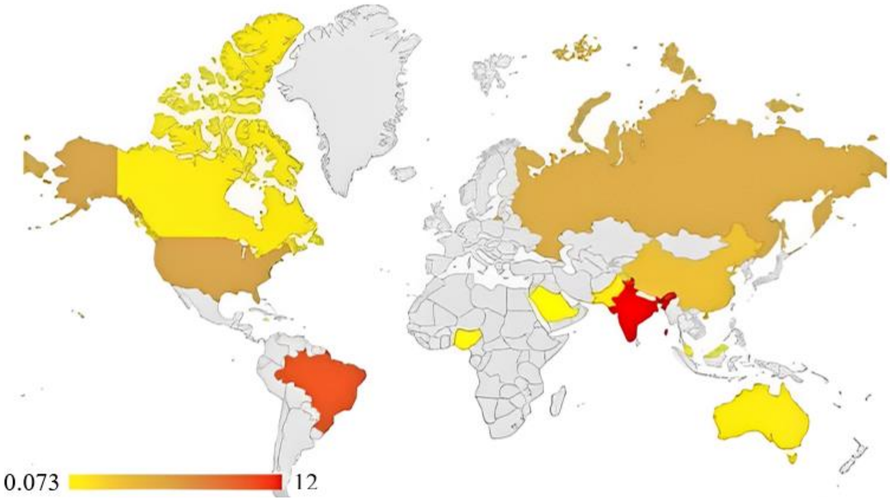


Figure 1. Variations in the severity of the NTL problem around the globe

2. PROPOSED METHOD: NON-TECHNICHAL LOSSES DETECTION

There are a few different ways to classify NTL detection techniques, but the three main types are data focused, network oriented, and hybrid. There are two types of data-oriented approaches: supervised and unsupervised. Supervised approaches leverage both labels (positive/fraud and negative/non-fraud classes) whereas unsupervised methods do not. network-centric are founded on the study of network structures and the physical laws that govern how they operate. State estimation, load flow, and specialized sensors are some examples of these approaches. Hybrid approaches take ideas from each of the aforementioned types [4]. Figure 2 displays the most common groupings. The impact that NTL have on the dependability of the electrical system and the substantial financial losses they cause for utilities is a big worry [20]. Recently published research [15] estimates that NTL causes yearly income losses of $96 billion throughout the world. NTL can have an impact on the functioning of a power system at the grid level by increasing the likelihood of transformer overload, voltage imbalances, and an absence of reliable information about actual power usage.

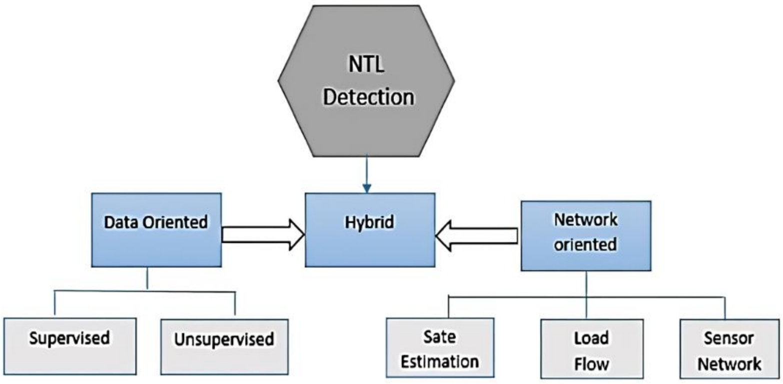


Figure 2. NTL detection methods categories

Higher NTL rates will be reflected in the price of electricity and the dependability of the power grid, therefore NTL impacts both dishonest and honest consumers. In addition, the dangers of fire and electrocution are exacerbated when people intentionally connect to the grid illegally [11]. Because of the technological and financial consequences, energy providers are devoting greater resources to minimizing NTL losses. All consumers, the general public, and society as a whole are negatively impacted by electricity theft and other forms of energy fraud [16], [21]–[23]. The research community has been working hard over the past decade to reduce the number of NTL incidents in the electricity sector. For instance, NTL detection techniques draw heavily from the usage of physical devices in addition to the data analytics on consumption patterns. The application of machine learning classifiers and deep learning algorithms on a dataset of hourly, daily, or monthly consumption records is one such commonly used method for NTL identification. These save time and money in the detection of probable anomalies [24]–[26].

3. METHOD IMPLEMENTATION

3.1. Procedure

There are a few things that need to be done in order to put this technique that was suggested into action. The next paragraphs will provide further explanation of these actions. The workflow that was intended is depicted in Figure 3. The model's input consists of the reader's interpretation of the consumption statistics supplied by the service provider. When looking at the statistics, several different types of customers and their individual consumption patterns are taken into consideration. During the pre-processing of the data, tasks such as feature engineering for data purification, missing value imputation, and data transformation are carried out. The data that was retrieved contains a great deal of irrelevant particulars and information. By removing these superfluous qualities, we are cleaning up the data we have collected.

The proposed work will use four model machine learning algorithms and four model deep learning techniques to analyze historical consumer data from an Iraqi electricity distribution company between the years 2015 and 2021 in order to identify non-technical losses (NTL) caused by any anomaly and resulting in a decrease in revenue for the ministry of electricity. This will be done in order to identify non-technical losses (NTL) caused by any anomaly and resulting in a decrease in revenue for the ministry of electricity (MOE). The data preparation stage is the most important part of this project since it requires using the necessary libraries to manage a big dataset that contains the information of (1,056,856) customers over a period of seven years. Research will be carried out in order to discover an algorithm that is capable of functioning on the proposed model; this algorithm has to be modern, efficient, and more accurate in order to enable the production of superior outcomes.

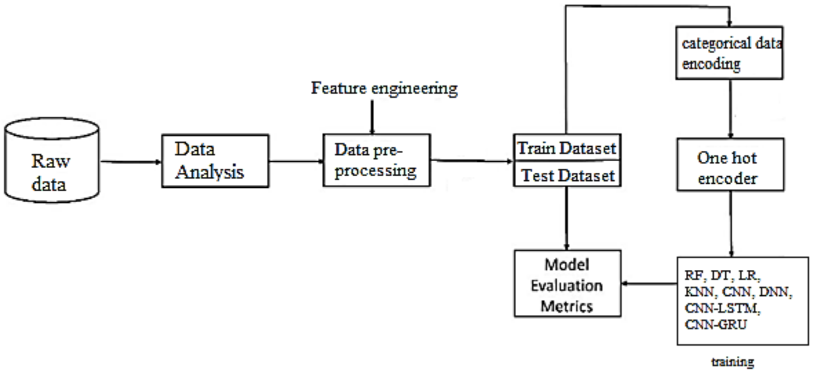


Figure 3. General flowchart of NTL detection

3.2. Data collection

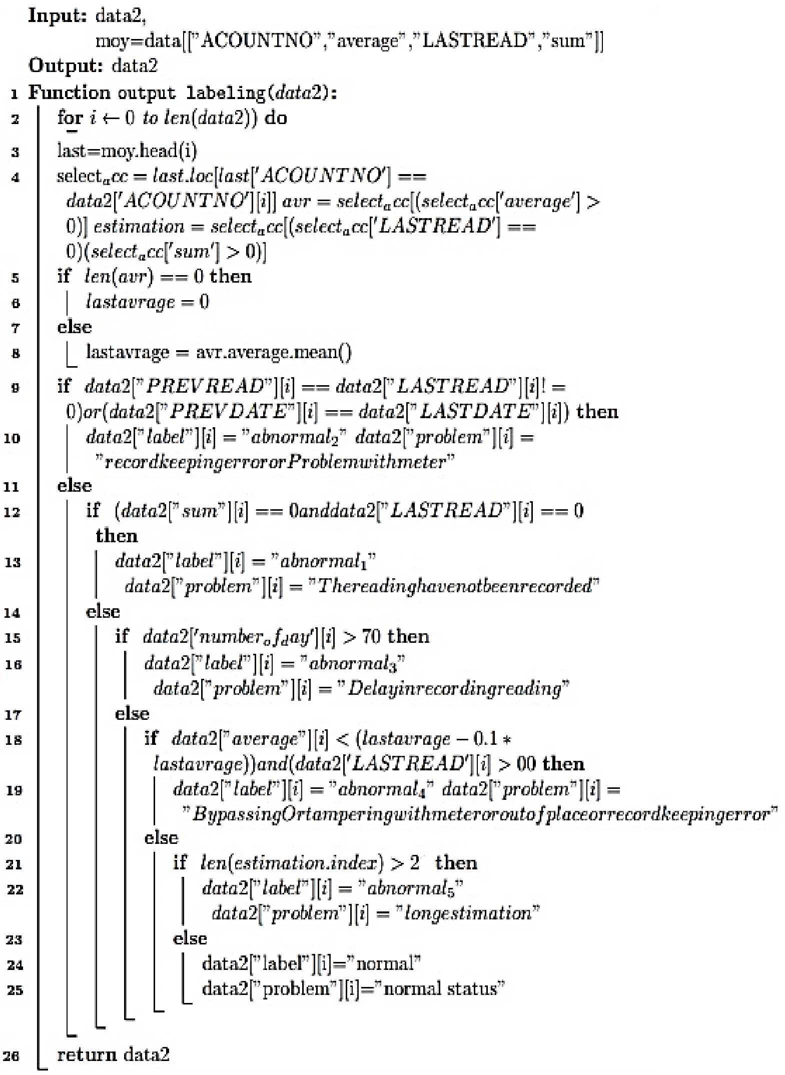
We need a real data collection to accomplish reliable NTL detection. Therefore, we contact a company that provides energy in Baghdad, Iraq, in order to get a real-world consumptions dataset from them. The figures cover the period from 2015 to 2021 and include consumption patterns from both households and businesses. The information was gathered from people who live in the Karkh, Rusafa, and Sader neighborhoods of Baghdad.

Number of users (as well as the size of the raw dataset): (25,613,835) (1,056,856). In order to determine the overall consumption of a household, mechanical meters were installed in the home's retail area as well as the living quarters. The data is segmented geographically and by category of consumption. The typical frequency at which customers get their power bills is once every two months.

In the dataset provided by the electric distribution utility, there is no tagged data target that identifies NTL or if it does not exist as shown in Table 1. Therefore, we manually transform an unlabeled dataset into a labeled dataset with a target by gaining knowledge from domain experts in the electrical facility about the characteristics included in the data set and the reasons for their presence, as well as the methodology used to derive consumption estimates. This is done by gaining knowledge about the characteristics included in the data set and the reasons for their presence, as well as the methodology used to derive consumption estimates. We are compiling information into a database to determine what constitutes a usual and an atypical intake, as well as to identify risk factors for NTL and classify drinking issues associated with NTL. Below is the required information that was learned from the experts, and the algorithm used is shown in Figure 4.

Table 1. Sample of collected data

	ACCOUNT NO	PREV READ	PREV DATE	LAST READ	…	CODE	CLOSE	PC
0	67021506	0	12/30/2017	0	…	3.0	NaN	0.0
1	67021506	0	12/30/2017	0	…	3.0	NaN	0.0
2	67021506	0	12/30/2017	0	…	3.0	NaN	0.0
3	67021506	0	12/30/2017	0	…	3.0	NaN	0.0
4	67021506	0	12/30/2017	0	…	3.0	NaN	0.0
…	…	…	…	…	…	…	…	…
26613832	536381518	68545	5/23/2018	0	…	2.0	NaN	0.0
26613833	536381518	68545	5/23/2018	0	…	2.0	NaN	0.0
26613834	536381518	68545	5/23/2018	70615	…	NaN	NaN	203961.0
26613835 rows × 40 columns								



```
Input: data2,
        moy=data[["ACOUNTNO","average","LASTREAD","sum"]]
Output: data2
Function output labeling(data2):
    for i ← 0 to len(data2)) do
    last=moy.head(i)
    $select_acc = last.loc[last['ACOUNTNO'] == data2['ACOUNTNO'][i]]\ avr = select_acc[(select_acc['average'] > 0)]\ estimation = select_acc[(select_acc['LASTREAD'] == 0)(select_acc['sum'] > 0)]$
    if $len(avr) == 0$ then
        $lastavrage = 0$
    else
        lastavrage = avr.average.mean()
    if $data2["PREVREAD"][i] == data2["LASTREAD"][i]! = 0)or(data2["PREVDATE"][i] == data2["LASTDATE"][i])$ then
        $data2["label"][i] = "abnormal_2"\ data2["problem"][i] = "recordkeepingerrororProblemwithmeter"$
    else
        if $(data2["sum"][i] == 0anddata2["LASTREAD"][i] == 0$ then
            $data2["label"][i] = "abnormal_1"$
            $data2["problem"][i] = "Thereadinghavenotbeenrecorded"$
        else
            if $data2['number_of_day'][i] > 70$ then
                $data2["label"][i] = "abnormal_3"$
                $data2["problem"][i] = "Delayinrecordingreading"$
            else
                if $data2["average"][i] < (lastavrage - 0.1 * lastavrage))and(data2['LASTREAD'][i] > 00$ then
                    $data2["label"][i] = "abnormal_4"\ data2["problem"][i] = "BypassingOrtamperingwithmeteroroutofplaceorrecordkeepingerror"$
                else
                    if $len(estimation.index) > 2$ then
                        $data2["label"][i] = "abnormal_5"$
                        $data2["problem"][i] = "longestimation"$
                    else
                        data2["label"][i]="normal"
                        data2["problem"][i]="normal status"
    return data2
```

Figure 4. Labeling algorithm

a) The description of some important fields in data set are:
   - ACCOUNTNO: represent consumer account number.
   - PREVREAD: represent previous read from meter.
   - PREVDATE: represent previous date for previous read.
   - LASTREAD: represent last read from meter.
   - LASTDATE: represent last date for last read.
   - CODE: represent consumer class (residential, commercial).
   - CONS1 CONS13: represent consumptions (unit is KWH).

b) To compute the consumption at given period (from previous date to last date):
   - Current consumption=LAST READ-PREV READ=Sum (CONS0, CONS1, CONS3, CONS13)
   - Note (LASTREAD) will be (PREVREAD) at the next reading (row)

c) To get zone location-based record code from ACCUNTNO column:
   - If ACCOUNTNO consist of (9 digit), the zone code will be (first 3 digit).
   - If ACCOUNTNO consist of (8 digit), the zone code will be (first 2 digit).
   - If ACCOUNTNO consist of (7 digit), the zone code will be (first 1 digit).

d) The three regions (Karkh, Rusafa, Sader) corresponding values are (1, 2, 3). Insert region column to data frame for each account ACCOUNTNO.

e) Classify target depending on the cases below:
   - Normal: there is a value in (CONS) column for consumer and not less than his estimated history consumption (no abnormal low between his consumptions).
   - Abnormal (NTL) cases as shown in Table 2.

Table 2. NTL abnormality

NTL Cases	
Case	Explanation
(LASTREAD) is null &(CONS) is Blank	The reading has not been recorded
(PREVREAD) & (LASTREAD) is equal, or (PREVDATE)=(LASTDATE)	record keeping error or Problem with meter
Difference between (LASTDATE & PREVDATE) more than (70) day	Delay in recording reading
Abnormal Little consumption the consumptions not regular lower than normal, there is abnormal low between the consumer's consumptions history lower than average of his previous consumptions with limit of tolerance ratio 0.1 to be fairer with consumer	Bypassing the meter Or tempering with meter Or record keeping error Or Unused (out of place)
((LASTREAD) is zero & sum (CONS)>0) & repeated more than 3 time for same consumer	Long estimation

3.3. Data pre-processing

According to the data obtained from the Iraqi distribution utility, there are a total of (40) characteristics, (26,613,835) raw, and (1,056,856) consumers. It has been demonstrated, however, that certain parts are not productive in any way. To give you an example, the features CONP1…CONP13, OUTS, TOTAL, EXCH, SPECIAL, CODE, and CLOSE, as well as PC, are all interrelated because of their usage in the billing system. Therefore, getting rid of these additions is not a problem at all.

- Check the number of empty spaces in each table, and if there are more than 0.99 of them, you can probably get rid of the columns labeled (CONS4…CONS13) and (CONP4…CONS13) because their values are all NaN.
- Alter the dates so that they are represented as a date-time object rather than as a string.

The raw information that was obtained has close to forty different features; however, only few of them are necessary for the analysis to be carried out. A new set of characteristics is developed as a direct consequence of doing research on the previously established characteristics. It is possible, according on this function, to categorize various kinds of clients as either normal or abnormal. Additionally, there should be a category for strange behaviour included in this list. First, in order to extract (number of days) for purposes, we first convert date columns to the date type, and then we apply an equation to retrieve the number of days. After that, we can proceed with our extraction (day count = last date - previous date).

Because our information covers a period of time spanning several years, the PREVDATE and LASTDATE columns need to have the year and month extracted first. The raw sum, which is computed by adding CONS1+CONS2+CONS3, contains a characteristic called extract (sum) that determines the entire current consumption as well as the daily average consumption. The weighted mean is calculated by dividing the total number of days by the weighted average. In the fourth step, a dataset is labelled with the help of an algorithm that has been trained using the expertise of domain experts.

3.4. Data processing

When using normalized (feature-scaled) data, estimators are able to learn features more effectively and efficiently. Standard scaling is utilized in order to acquire data with a mean value of zero and a standard deviation equal to one. Since machine learning algorithms can only handle numbers and not strings, the output feature, also known as the label, should be encoded using numbers rather than texts.

The final step is to select 12 features to use in the creation of a new dataset from the data frame. This new dataset should contain the columns that will be used as input features for deep learning models. Looking at the input and output features, we see that (12) features were employed.

Features={'prev_year', 'prev_month', 'PREVREAD', 'year', 'month', 'LASTREAD', 'CODE', 'Region', 'zone', 'number_of_day', 'sum', 'average'}. The "one-hot-encoding" of our target variable is an extremely important factor. A column will be created for each conceivable output category, and an entry will be produced for the binary variable that corresponds to that column. In the training and testing sets, the consumptions data will be stored in the variables x train and x test, while the output label class that those data represent will be stored in the variables y train and y test.

3.5. Build and fit the models

In addition to machine learning algorithms model, random forest, decision tree, KNN and logistic regression, in this study, four deep learning models was proposed and implemented to train data mentioned previously. A model of 5 layers of dense deep learning was proposed as shown in Figure 5(a). Another model build basing on 1-D convolutional neural network was proposed to be building as shown in Figure 5(b). Instead of separableConv and MaxPooling function, two layers of Long short-term memory LSTM was proposed after Average polling function as third model shown in Figure 5(c). Same architecture, replacing LSTM by gated recurrent unit (GRU), GRU was the fourth model tested is shown in Figure 5(d). To sum up the architecture of the proposed and build system for deep learning methods is show in Figure 6.

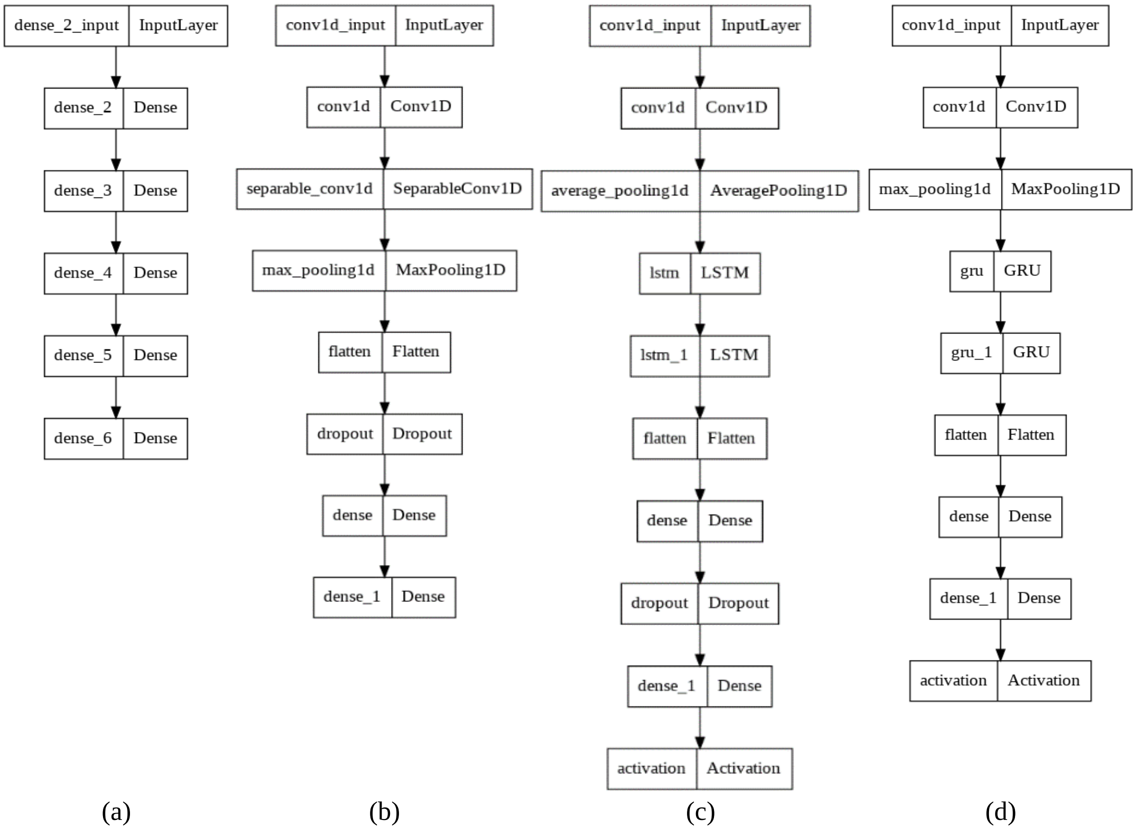


Figure 5. Layers of model: (a) DNN model, (b) CNN model, (c) CNN-LSTM model, and (d) CNN-GRU model

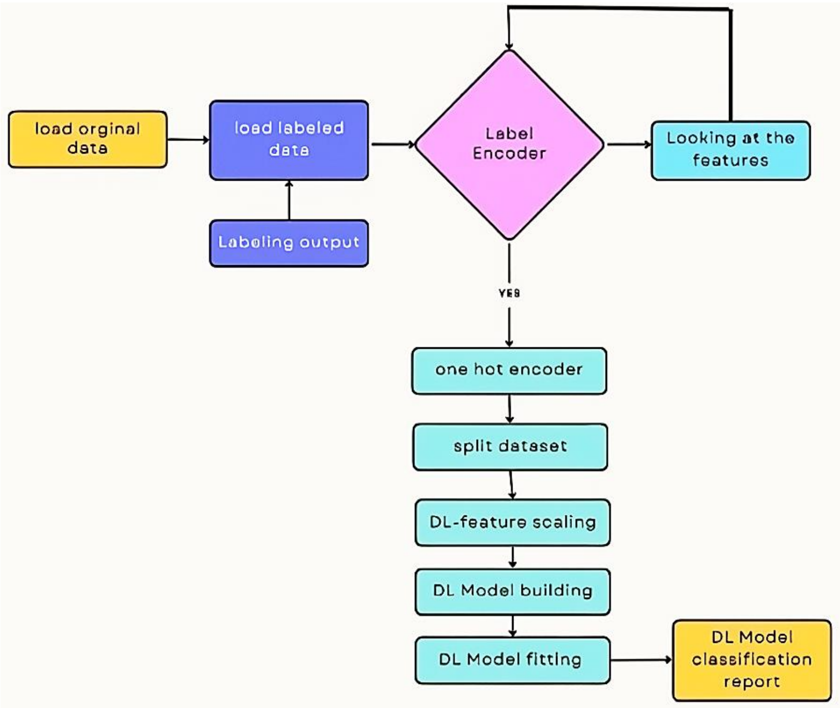


Figure 6. Methodology flowchart

4. RESULTS

Figure 7 illustrates the accuracy results for various machine learning methods including random forest, decision tree, KNN, and logistic regression. Random forest achieved the highest accuracy among these methods. Figure 8 presents the training times for these methods, with KNN having the shortest training time. Table 3 shows the recorded results for all models after fitting. Notably, DNN, and CNN achieved high accuracy rates exceeding 90% during the first 5 epochs, indicating that these proposed models are considered high-performance deep learning models.

While random forest outperformed the other models in terms of accuracy, it had the most significant training time. Deep learning models, except for CNN-GRU, performed well in terms of accuracy when compared to other models. Deep learning models have the advantage of dealing with large datasets and maintaining stable accuracy even with an increase in data size. While the cost for data stability may be higher than that of machine learning methods, it is still considered reasonable.

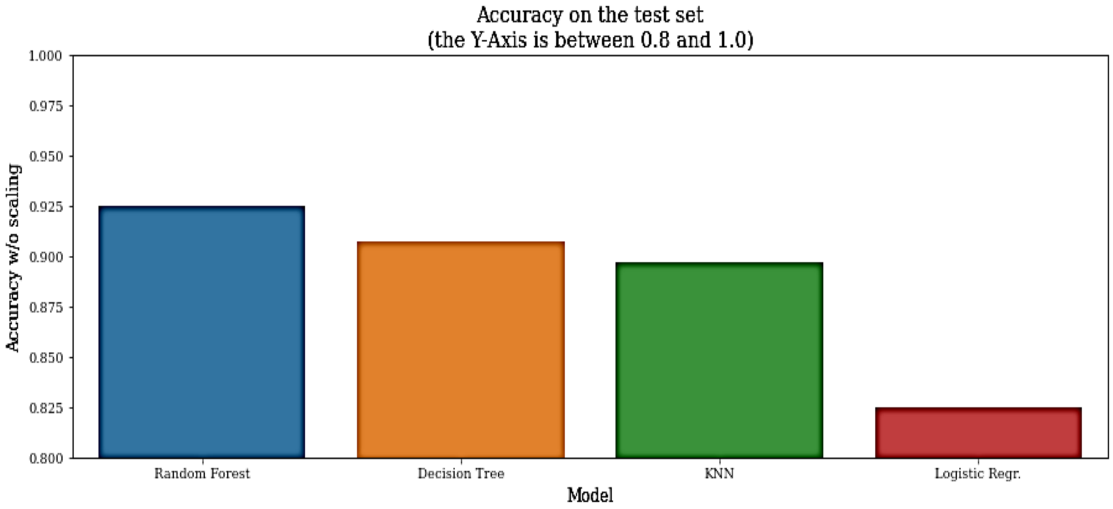


Figure 7. Machine learning accuracy results

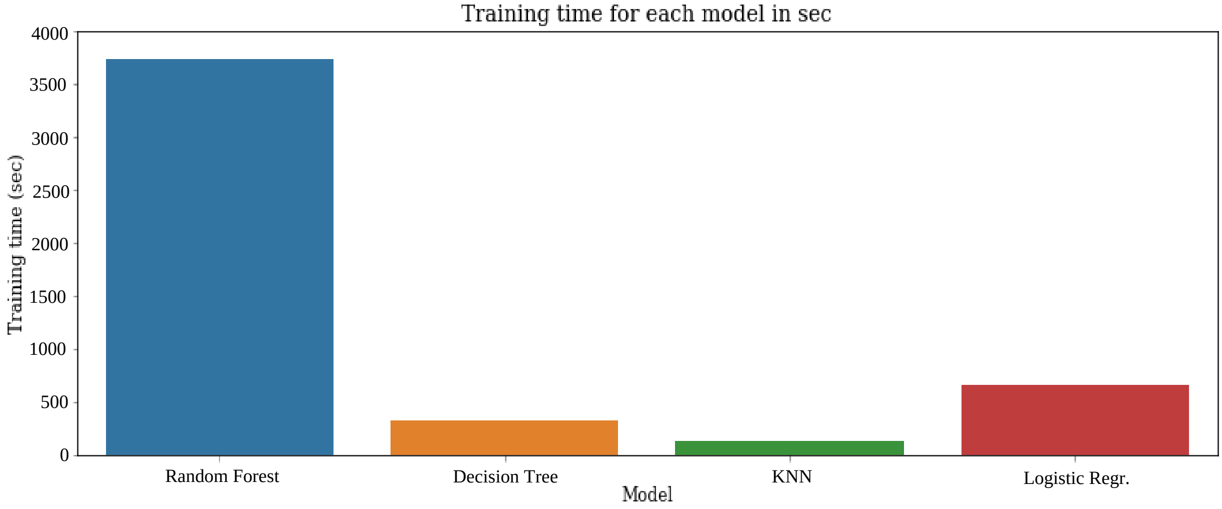


Figure 8. Machine learning methods training time

Table 3. Accuracy and time results of models

	Results				
Algorithm	Accuracy (%)	Time (s)	Algorithm	Accuracy (%)	Time (s)
Random forest	92.46	3733	DNN	90.74	593
Decision tree	90.71	327	CNN	90.73	566
KNN	89.71	129	CNN-LSTM	90.27	4058
Logistic regression	82.51	661	CNN-GRU	83.03	3146

5. CONCLUSION

This study primarily focused on investigating non-technical methods for monitoring power losses, which have a significant impact on the financial stability of utilities and economies. Non-technical losses (NTLs) can result from power theft, fraud, or inadequate metering assets, and are the primary cause of distribution losses in electrical power networks, imposing a significant financial burden on utilities. The study examined three basic types of NTL detection methods: data-focused, network-oriented, and hybrid approaches, with a particular emphasis on data-oriented writing to achieve its goals.

Following data collection and cleaning processes, the study presented a methodology that utilized machine learning methods including random forest, decision tree, KNN, and logistic regression, as well as neural network models such as DNN, CNN, CNN-LSTM, and CNN-GRU. The recommended CNN and DNN model offered maximum performance in terms of accuracy and stability, fast learning with efficient training time.

REFERENCES

[1] T. Hu, Q. Guo, X. Shen, H. Sun, R. Wu, and H. Xi, "Utilizing unlabeled data to detect electricity fraud in AMI: a semisupervised deep learning approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3287–3299, Nov. 2019, doi: 10.1109/TNNLS.2018.2890663.

[2] K. Fei, Q. Li, C. Zhu, M. Dong, and Y. Li, "Electricity frauds detection in low-voltage networks with contrastive predictive coding," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107715, May 2022, doi: 10.1016/j.ijepes.2021.107715.

[3] M. Badr, M. Ibrahem, H. Kholidy, M. Fouda, and M. Ismail, "Review of the data-driven methods for electricity fraud detection in smart metering systems," *Energies*, vol. 16, no. 6, p. 2852, Mar. 2023, doi: 10.3390/en16062852.

[4] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250–266, May 2018, doi: 10.1016/j.epsr.2018.01.005.

[5] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Science and Technology*, vol. 19, no. 2, pp. 105–120, Apr. 2014, doi: 10.1109/TST.2014.6787363.

[6] A. Chauhan and S. Rajvanshi, "Non-technical losses in power system: a review," in *2013 International Conference on Power, Energy and Control (ICPEC)*, Feb. 2013, pp. 558–561, doi: 10.1109/ICPEC.2013.6527720.

[7] A. Fragkioudaki, P. Cruz-Romero, A. Gómez-Expósito, J. Biscarri, M. J. de Tellechea, and Á. Arcos, "Detection of non-technical losses in smart distribution networks: a review," 2016, pp. 43–54.

[8] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," *Journal of Electrical and Computer Engineering*, vol. 2019, pp. 1–12, Oct. 2019, doi: 10.1155/2019/4136874.

[9] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, May 2019, doi:

10.1109/TSG.2018.2807925.

[10] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *International Journal of Electrical Power & Energy Systems*, vol. 101, pp. 301–310, Oct. 2018, doi: 10.1016/j.ijepes.2018.03.031.

[11] S. Osypova, "Consumption pattern detection through the use of machine learning: clustering techniques for non-technical losses detection RERORT," 2020.

[12] F. Shehzad, N. Javaid, A. Almogren, A. Ahmed, S. M. Gulfam, and A. Radwan, "A robust hybrid deep learning model for detection of non-technical losses to secure smart grids," *IEEE Access*, vol. 9, pp. 128663–128678, 2021, doi: 10.1109/ACCESS.2021.3113592.

[13] S. Hussain *et al.*, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Reports*, vol. 7, pp. 4425–4436, Nov. 2021, doi: 10.1016/j.egyr.2021.07.008.

[14] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010, doi: 10.1109/TPWRD.2009.2030890.

[15] M. S. Saeed *et al.*, "Detection of non-technical losses in power utilities—a comprehensive systematic review," *Energies*, vol. 13, no. 18, p. 4727, Sep. 2020, doi: 10.3390/en13184727.

[16] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1254–1263, Mar. 2020, doi: 10.1109/TPWRS.2019.2943115.

[17] A. N. Aniedu, H. C. Inyiama, A. C. O. Azubogu, and S. C. Nwokoye, "Pattern recognition using support vector machines as a solution for non-technical losses in electricity distribution industry," *International Journal of Innovative Science and Modern Engineering*, vol. 7, no. 2, pp. 1–8, Mar. 2021, doi: 10.35940/ijisme.B1280.037221.

[18] Saeed, Mustafa, Sheikh, Jumani, and Mirjat, "Ensemble bagged tree based classification for reducing non-technical losses in multan electric power company of Pakistan," *Electronics*, vol. 8, no. 8, p. 860, Aug. 2019, doi: 10.3390/electronics8080860.

[19] R. Yadav and Y. Kumar, "Detection of non-technical losses in electric distribution network by applying machine learning and feature engineering," *Journal Européen des Systèmes Automatisés*, vol. 54, no. 3, pp. 487–493, Jun. 2021, doi: 10.18280/jesa.540312.

[20] A. McIntyre *et al.*, *WP/16/53 caribbean energy: macro-related challenges*. Adrienne Cheasty, 2016.

[21] G. A. Ajenikoko and O. J. Ogunwuyi, "An energy fraud detection scheme for power utilities," *Journal of Engineering Research and Applications www.ijera.com ISSN*, vol. 5, no. 52, pp. 2248–962257, 2015, [Online]. Available: http://www.ijera.com/papers/Vol5_issue5/Part - 2/H505025760.pdf.

[22] N. Calamaro, Y. Beck, R. Ben Melech, and D. Shmilovitz, "An energy-fraud detection-system capable of distinguishing frauds from other energy flow anomalies in an urban environment," *Sustainability*, vol. 13, no. 19, p. 10696, Sep. 2021, doi: 10.3390/su131910696.

[23] S. Poudel and U. R. Dhungana, "Artificial intelligence for energy fraud detection: a review," *International Journal of Applied Power Engineering (IJAPE)*, vol. 11, no. 2, p. 109, Jun. 2022, doi: 10.11591/ijape.v11.i2.pp109-119.

[24] L. Vigoya, A. Pardal, D. Fernandez, and V. Carneiro, "Application of machine learning algorithms for the validation of a new CoAP-IoT anomaly detection dataset," *Applied Sciences*, vol. 13, no. 7, p. 4482, Apr. 2023, doi: 10.3390/app13074482.

[25] S. Brady, D. Magoni, J. Murphy, H. Assem, and A. O. Portillo-Dominguez, "Analysis of machine learning techniques for anomaly detection in the internet of things," in *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, Nov. 2018, pp. 1–6, doi: 10.1109/LA-CCI.2018.8625228.

[26] H. S. Mukhandi, "Developing machine learning methods for network anomaly detection," 2018.

BIOGRAPHIES OF AUTHORS

Israa Mohammed Ridha Baldawi received her B.Sc. in Computer Science from Baghdad University, Iraq, in 2004; and she working on Master degree in Information Technologies from Altinbas University, İstanbul, Turkey. Her research interests include the field of, artificial intelligence, machine learning, deep learning and processing metaheuristic optimization algorithms. She can be contacted at email: shiningsun452@yahoo.com.

Timur İnan is a lecturer Software Engineering Department at the Altınbaş University, İstanbul, Turkey. He received his B.Eng., M.Eng. and Ph.D. degrees in Electric-Electronics Engineering from Marmara University, Turkey, in 2005, 2013 and 2019, respectively. He has been an Associate Professor in Altınbaş University, İstanbul, Turkey since 2021. He is currently vice head of department in software engineering department at Altınbaş University. His research interests include the field of embedded system design, internet of things, artificial intelligence, intelligent control, java programming, image processing and metaheuristic optimization algorithms. He can be contacted at email: timur.inan@altinbas.edu.tr.