Sādhanā (2020) 45:286
https://doi.org/10.1007/s12046-020-01512-0

Detection of electricity theft using data processing and LSTM method in distribution systems

BEHÇET KOCAMAN[1,*] and VEDAT TÜMEN[2]

[1]Department of Electrical and Electronics Engineering, Bitlis Eren University, Bitlis, Turkey
[2]Department of Computer Engineering, Bitlis Eren University, Bitlis, Turkey
e-mail: bkocaman@beu.edu.tr; vtumen@beu.edu.tr

MS received 16 April 2020; revised 6 October 2020; accepted 9 October 2020

Abstract. Electricity theft is a big problem faced by all energy distribution services and continues to rising. Therefore, studies on electricity theft detection techniques have increased in recent years. Unsuitable calibration and illegal calibration of energy meters during production may cause non-technical losses. Non-technical losses have been a major concern for the resulting security risks and the immeasurable loss of income. In most of the meter tampered locations, damaged meter terminals and/or illegal applications cannot be distinguishable during checking. In fact, electric distribution companies will never be able to eliminate electricity theft. But it is possible to take measure to detect, prevent and reduce it. In this paper, we developed by using deep learning methods on real daily electricity consumption data (Electricity consumption dataset of State Grid Corporation of China). Data reduction has been made by developing a new method to make the dataset more usable and to extract meaningful results. A Long Short-Term Memory (LSTM) based deep learning method has been developed for the dataset to be able to recognize the actual daily electricity consumption data of 2016. In order to evaluate the performance of the proposed method, the accuracy, prediction and recall metric was used by considering the five cross-fold technique. Performance of the proposed methods were found to be better than previously reported results.

Keywords. Electricity theft; non-technical loss; long short term memory.

1. Introduction

In electricity consumption, very low technical losses occur due to the system and these losses are considered as maximum 5% of total consumption. Malicious consumers consuming electricity with different methods are called non-technical losses (NTL) or electric theft. Electric theft can lead to electricity unit prices, heavy load of electrical systems, huge loss of energy company and the dangers of public safety (such as fires and electric shocks). NTL behavior usually involves skipping the electricity meter, tampering the meter reading, or hacking the meter [1]. With the increase in energy needs on global based, the amount of consumption increased proportionally. An increase in energy prices affects the economy and increases the search for illegal ways for those responsible. Electricity theft rates differ in developed and developing countries. The average of this rate in OECD countries is about 7%.

While the rate of electricity theft is 1–2% in developed countries, this rate is quite high in developing countries. Especially in developing countries such as in India, Bangladesh, Turkey and Malaysia illegal use of electricity by up to 30% are available [2]. It has been observed that the high use of illegal electricity in developing countries generally depends on the financial situation and poverty of the customers. On the other hand, the rate of electricity theft has been quite low in China, where the income level of the people is quite low. The Chinese government has studies and serious sanctions for the detection of electrical theft detection (ETD).

Electricity thefts cause serious financial losses, serious injuries and even fatal accidents. In India, the country with the highest rate of electricity theft, about 4.5 billion dollars of damage occurs every year [3]. Electricity theft rate in Turkey about 11.8% [4] and annual material loss was determined to be around $ 1 billion [5]. Public services in the US suffer around $ 1–6 billion annual [6, 7]. In Canada, $ 100 million was damaged annually [8].

While electricity theft causes serious financial losses in public institutions, it also causes long term failures in the electricity distribution networks with the overloads in power consumption. This, shortens the life of power elements and affects the performance of consumers' electrical equipment. In addition, these overloads increase fossil fuel based electricity production and thus carbon dioxide emissions. Another problem caused by the theft of

*For correspondence
Published online: 18 November 2020

electricity is to share the lost costs in electricity to honest customers who regularly pay the electricity bill. This situation causes unfair earnings and injustice and also increases the unit price of energy.

Many new ideas and technologies are being developed to monitor the electricity consumption of residential and non-residential subscribers, and these technologies do not completely solve the problem but cause serious material losses. Researches are conducted in different fields to produce more economical solutions. Low-cost detection systems are developed to reduce losses in the NTL area. The most appropriate determination is the development of a recognition-prediction system for tracking and determining the daily electricity consumption data of customers.

In this study, a deep learning-based LSTM model, which has been developed using real daily electricity consumption data, has been proposed for the detection of electrical theft. In this proposed method, the system has been trained using a part of the dataset. The method has been tested with the rest of the dataset.

In section 2, a literature review on the detection of electricity theft is given and the methods developed are evaluated. In section 3, the model and materials used for the study have been explained. In section 4, the results of the method proposed and its comparisons and contributions with other studies in the literature are presented. Conclusions are provided in section 5.

2. Related work

Machine learning based classifiers are used because the electricity consumption data is generally in one dimensional and time series form. There are many new studies [9–12] using Support Vector Machine (SVM) based classifiers in particular. Apart from this method, there are studies [13, 14] which are artificial neural networks are used to detect of electricity theft. The success rates of these studies in detecting electrical theft are low. In addition, artificial feature extraction is required.

Depuru, et al. used some of the energy consumption data of 20.000 customers with different time interval values for the training of SVM and Rule Engine algorithms. They achieved high successes of 85.5% and 92% with the different models they proposed in their study [12].

Zheng *et al* [15] proposed wide and deep Convolutional Neural Network (CNN) model to detect electrical theft in smart grids. The method they developed was applied on the State Grid Corporation of China Dataset (SGCCD) that we used in our study and achieved high performance [15].

Hasan *et al* [16] proposed a CNN-LSTM model to detect electrical theft in smart grids. In the study he proposed, he used only 9956 total customer data over only 2015 year on the SGCCD database Since there was an unbalanced data distribution, he used synthetic data generation technique on the NTL class. It increased the number of NTL data to 8562, the number of normal customers. In this study, he achieved 89% success despite using synthetic data [16].

Souza *et al* [17] Multilayer Perceptron has developed a new method for the detection and identification of energy theft in distribution systems using the Multilayer Perceptron Artificial Neural Network (MP-ANN) algorithm. In his study, they used real data obtained from 5000 consumer values with different qualifications and successfully classified malicious users and normal users with an average of 93.4% [17].

Because of the theft of electricity rate is 30% in India, many studies have been done on the data collected in this country. Gaur *et al* [18] analyzed the annual data collected in 28 states of India between 2005–2009. In these analyzes, they determined that the socio-economic situation affected electricity theft at a very high rate [18].

Yip *et al* [19] has tried to identify customers who are doing electrical theft using smart meters with two different algorithms developed based on linear regression method [19].

Ghasemi *et al* [20] proposed a unified method to detect two different states of illegal electricity consumption. To determine the type of electrical theft, the customer energy consumption pattern classification method based on the probabilistic neural network and the mathematical model based on the Levenberg–Marquardt method have used. It has detected the consumers using electricity theft with low success with the method proposed [20].

Viegas *et al* [21] Using the cluster-based methods which are Fuzzy C-means and Fuzzy Gustafson–Kessel algorithms, 74.1% successfully distinguished between normal consumers and malicious consumers from the electricity consumption data collected between 2009–2010 [21].

Razavi *et al* [22] designed a new model for the detection of electricity theft in smart grids. They proposed a new Genetic Programming algorithm combination to define new properties suitable for prediction with Finite Mixture Model aggregation. Gradient Boosting Machine algorithm has been applied together with the k-fold method to evaluate demand data from more than 4000 households, and previous machine learning has achieved significantly better performance than the results of theft detection methods [22].

Different artificial intelligence based methods have been developed for NTL detection. It was developed with the study of fuzzy logic and evaluation of daily consumption data using SVM [23]. Adil *et al* have detected electrical theft on SGCCD datasets with the proposed RUSBoot module LSTM model. It also achieved very low results using SVM algorithm. Hence, he claimed that SVM algorithm is not suitable for class imbalance problem in high dimensional data [24]. Ullah *et al* used the Gated Recurrent Unit (GRU) method, which is an advanced version of LSTM, also a type of Recurrent Neural Networks (RNN). This method consists of updating and reset gates. The

updated gate determines how much of the past information needs to be transferred to the future [25]. In this study, using the geographic information of the customers, using SVM, k-Nearest Neighbors (k-NN), Random Forest RF algorithms [26].

In the literature review for NTL detection, it is seen that real or synthetic 1B electricity consumption data are used and all recognition-prediction-determination study is done on the processing and analysis of these data. For this reason, the aim of this study we have proposed is to overcome high performance electrical theft by using the limited dataset that contains one-dimensional real customer electricity consumption data.

3. Material and Method

The methods developed by processing data using smart meter readings and using electricity consumption data from smart grids for the detection of electricity theft have attracted great attention recently. In this study, a method has been developed using real daily electricity consumption data. This method, we have proposed for the detection of electricity theft, basically consists of two stages. In the first stage, data needs passed improved preprocessing and data editing stages for data processed and produce meaningful inferences. In the second stage, an LSTM based framework is proposed for a recognition system by analyzing one-dimensional data and using this data.

3.1 *Database*

Many countries or electricity distribution companies record real electricity consumption values on a daily and periodic basis to investigate consumers' electricity consumption behavior. Due to electricity theft is high in China, consumption values are constantly recorded and necessary values are tried to be taken according to these values. Electricity consumption data was collected from 1035 days and 42372 customers by SGCCD. Randomly selected a monthly (October, 2016) daily electricity usage change of a three electricity theft usage (a) and three normal consumers (b) is shown in figure 1.

When figure 1 is examined, we can see that there is a different fluctuation for each day in the electricity consumption data. When the database is examined, it is detected that the consumption of almost every month has similar characteristics. Daily energy consumption of consumers who are electricity theft and normal consumers is closely similar some days, however, there are differences when looking at the whole month consumption information. It is not similar to detect electrical theft with these observations. When the data in two-dimensions produced using the values of one-dimensional consumption data is observed weekly, it is seen that there is a regular distribution. In fact, we can have similar findings for the entire dataset (i.e., 1035 days of electricity consumption data). Without much repetition, we only show a quote data from the entire dataset. If we put together the electricity consumption data of all 35 months, we can observe that there is a periodicity for most normal customers. Different data preprocessing methods have been developed by analyzing these pattern values found in the dataset. In fact, we can see similar pattern values when the entire dataset is examined. If we put together the 35-month electricity consumption data in the database, there are quite high differences in the daily energy consumption of normal customers, which are similar but normal customers. Different data pretreatment methods have been developed by analyzing these pattern values in the database.

3.2 *Data Preprocess*

Since electricity consumption data are situations where smart meters fail, cannot be read, record incorrect values, erroneous data is generated. When the dataset is examined, there is a missing or incorrect value in the data of about 3 years of electricity consumers. The main reason for this is due to various reasons such as malfunction of smart meters, unreliable transmission of measurement data, unscheduled system maintenance and storage problems. The numbers of situations such as not reading the meter or getting a 0 value due to problems or system-related errors are shown in table 1.

Normal and NTL data numbers of the database are given in table 1. For 2014, there was no data from the system where at least one data was NaN in all days, and the number of zero readings is 22.421. The number of customers with more than 200 NaN data was 13.316, while the number of customers with 0 was 4.284. Considering the consumption data for 2015, the number of containing more than 200 NaN values was determined as 13.316. NaN data was found in 29,735 customers in total. When the number of zero is considered, although there is less than the data of 2014, there is 10.718 zero data in total. When the data of 2014 and 2015 are analyzed, both NaN raw and Zero data numbers are quite high. Since the data that cannot be read by the system for any reason is in a number that will directly affect the success of the recognition system to be developed, the preprocessing stage of the dataset is of great importance for the system.

3.2a. *Data selection*: In this study, it is necessary to have real and consistent data as the proposed system is based on the examination of daily electricity consumption data of electricity consumers. Due to most of the customer data for 2014 and 2015 contain NaN and zero value, the system to be developed by taking these into account is not expected to produce accurate and consistent results. At the same time, it will negatively affect the training stage of the system to be developed. Despite this situation, data for 2016 because of

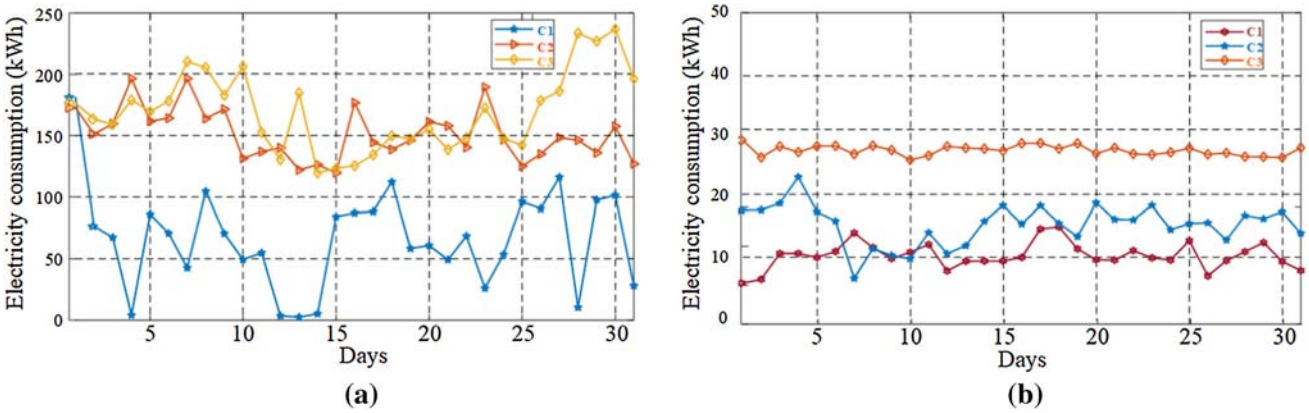

Figure 1. (a) Example of electricity consumption (kWh) of electricity theft usage by month. (b) Example of electricity consumption (kWh) of normal usage by month.

Table 1. SGCCD electricity theft detection database.

	Number of NaN data			Number of Zero data			Number of Normal data		
Number	2014	2015	2016	2014	2015	2016	2014	2015	2016
<10	15.718	9.637	13.693	9.128	2.124	2.441	17.525	30.610	26.237
11–50	7.407	2.471	3.113	5.661	1.821	2.341	29.303	38.079	36.917
51–100	2.206	926	387	1.434	1.067	1.410	38.731	40.378	40.574
101–200	3.724	2.106	169	1.914	1.493	2.117	36.733	38.772	40.085
>200	13.316	14.595	132	4.284	4.213	6.023	24.771	23.563	36.216
Total	42.371	29.735	17.494	22.421	10.718	14.332	24.846	28.903	30.341

more consistent and contains less NaN data and it can be edited by preprocessing these data has been approved to use. For these reasons, total 304 daily electricity consumption data of 2016 have been selected in the study. It has been observed that consecutive daily consumption data is NaN or zero in this data. For this reason, it must be re-preprocessing. The numbers and rates of dataset according to the characteristics of the data for 2016 are shown in table 2.

As you can see in table 2, while the number of customers containing between 101 and 200 NaN data has been 169, the number of customers containing more than 200 NaN data has been 132. When the dataset is examined, the number of data without any NaN value and zero value is 30.341. However, since the number of data with electrical theft label is very low, a high amount of data reduction should not be made. The number of data labeled NTL must be high to create a better trained model. In addition, since NaN data originating from the system and zero data can always occur, these data must be preprocessing and included in the study according to a proposed equation. Since the number of zero and NaN data in the dataset with the same consumer is proportionally higher than all data is not accurate measurement, this data should be eliminated. Therefore, according to Equation (1), daily electricity consumption values of customers containing NaN and zero data are eliminated.

Table 2. Consumer data SGCCD electricity theft detection dataset in 2016.

Num. of data	NaN	Zero	Normal	Rate (%)
<10	13.693	2.441	26.237	61.92
11–50	3.113	2.341	36.917	87.12
51–100	387	1.410	40.574	95.75
101–200	169	2.117	40.085	94.60
>200	132	6.023	36.216	85.47
Total	17.494	14.332	30.341	71,60

$$df = \begin{cases} crop\ data, & if\ s(NaN) > \dfrac{s(df_{feature})}{3}\ or\ s(Zero) > s(df) \\ crop\ data, & if\ (s(NaN) + s(Zero)) > \dfrac{s(df_{feature})}{2} \end{cases} \quad (1)$$

According to Equation (1), the number of data has been decreased and more qualified data has been obtained with the preprocessing made on the dataset. With the application of this equation, a total of 33979 customer data has been obtained, which is less than 1/3 (101) of the feature of dataset (304) containing NaN and Zero data, or 1/3 of the

total of NaN and zero data numbers. The NaN and zero data contained in this dataset should be preprocessing since the data will not produce accurate measurements. For this purpose, NaN data has been recalculated according to the method specified in Equation (2).

$$f(x) = \begin{cases} avg(x_{i-5} + \cdots + x_{i+5}, & if\ x_i \in NaN, x_{i-5}, x_{i-4} \ldots x_{i+4}, x_{i+5} \notin NaN \\ 0, & if\ x_i \in NaN\ and\ Other \end{cases} \tag{2}$$

According to the calculation specified in Equation (2), NaN data has been calculated for the new value and zero value by looking at the 10 data on the left and right. The calculation has been made by looking at the 5 values to the left and right of the NaN value. If the 5 values to the left and right are not NaN, the arithmetic average of these values is taken. If NaN cell has more than 5 NaN values on the left and right, this is taken as zero. With this method, the dataset of 33979 electricity consumers has been obtained, which the dates of between 01.01.2016 and 31.10.2016, has been done in a total of 304 days, with regular and various recognition processes. The use of this method can be considered as another unique value of the proposed study.

3.2b. *Data normalization*: Maximum and minimum value of consumption data which are in dataset are in the range of 0–39 kWh. The daily values of these data are approximately similar to each other. Due to the patterns that can extraction from these similarities cannot be detection, a range of value for the weights should be determined. The value ranges of the data have been examined to normalize the dataset and all data have been normalized in the range from -5 to 5. With this normalization process, the value ranges of the data became more evident.

3.2c. *Update of weights in loss function*: When the label of the data in the dataset is examined, 9.18% of the data is NTL data and the others are available as normal electricity consumption data. This can be considered as an unstable dataset. The bias value should be adjusted in the deep learning model to be created. The initial bias value of the proposed model is calculated as $b_0 = -\log_e(\text{pos}/\text{neg})$. Where, *pos* is the number of NTL customers, *neg* is the number of normal customers. The value of the weights of the classes has been calculated as $\text{weight} = 1/\text{class} \cdot \text{total}/2$ where $\text{class} = \{\text{pos}, \text{neg}\}$ and $\text{total} = \text{size}(\text{data})$. With this calculation, weights of data which is include NTL data have been detected as 5.45 (*pos*), and weights of normal data as 0.55 (*neg*). To update the b_0 value, the method will be able to rearrange the network, select new weights, and select cells more accurately.

3.2d. *Prepared datasets*: In this study, the dataset has been divided according to the cross fold method with determined value at different rates, to reach acceptable results. Since the study has been done according to the cross fold validation method, the dataset is randomly divided into 5 different parts. In this way, testing is possible on all data. The total number of data formed according to fold-1 in the dataset is shown in table 3.

As seen in table 3, three different datasets were prepared. Each dataset was divided into 5 different groups with equal data numbers according to the cross-fold method. There are NTL data between 8.09 and9.71% in each of these folds.

3.3 *Method*

For early detection of electricity theft, a method consisting of data processing and a LSTM-based deep learning model has been proposed. In this method, data selection, normalization of the data and updating of the weights are evaluated as data processing. In addition, the LSTM-based deep learning model proposed using the dataset created was trained and tested. Block diagram of the proposed method is shown in figure 2.

As seen in figure 2, data selection was made according to the calculation in the proposed Equation (1) in SGCCD database. Then, this database has been preprocessed and some of the data containing NaN and zero has been eliminated or arranged according to Equation (2). many datasets were obtained by dividing the preprocessed database into different folds randomly for both more and acceptable results. The update of the weight parameters of the unbalanced data in the processed of obtained datasets was made during the data normalization phase. With this method, NTL data, which is only 9% of all data, was tried to be balanced. LSTM model has been developed for classification of electrical data at the last stage of the proposed method. By using this model, which has a working logic based on recalling and processing the previous situation, it is aimed to recognize and classify the consumption data that will occur later.

3.3a. *Long-short term memory*: LSTM based model was developed to evaluate the performances of this study. LSTM model basically consists of RNN with cellular memory which perform better than deep neural network systems for classification of voice and signal data [27, 28]. The input memory cell must be checked to control the input. The output memory cell is also checked to control the output flow to the LSTM blocks. The LSTM memory block and its internal structure are shown in figure 3.

LSTM architectures have some of their hidden layers in memory and may be decisive in the next node of the previous result. The recall gate on the LSTM memory block is controlled by a single layer neural network. The activation of this important gate is done by Equation (3).

$$f_t = \sigma\left(W[x_t, h_{t-1}, c_{t-1}] + b_f\right) \tag{3}$$

where x_t is the input data, h_{t-1} is the previous cell output, $c_{t-1}$is the previous LSTM cell memory, and the bias vector value is b_f. σ represents basic sigmoid function and W is weight vectors of inputs. The output of the forget gate (σ) is

Table 3. The total number of data formed according to fold-1 in the dataset.

	Dataset-1			Dataset-2			Dataset-3		
Label	Num. of training data (60%)	Num. of validation data (20%)	Num. of testing data (20%)	Num. of training data (70%)	Num. of validation data (15%)	Num. of testing data (15%)	Num. of training data (80%)	Num. of validation data (10%)	Num. of testing data (10%)
Normal	3.700	1.234	1.249	4.320	933	930	4.952	680	680
Theft	377	126	110	437	87	89	484	63	66
Total	4.077	1.360	1.359	4.757	1.020	1.019	5.436	617	614
Rate	9.25%	9.26%	8.09%	9.19%	8.53%	8.73%	8.90%	9.26%	9.71%

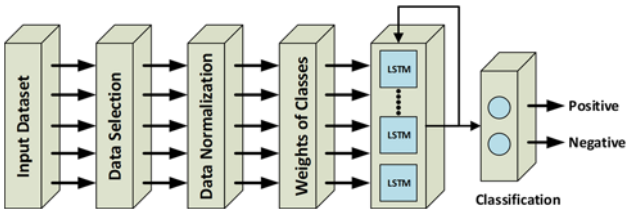

Figure 2. Block diagram of the proposed method.

applied to the previous memory block by element-based multiplication. The sigmoid activation function is applied by making an element-based multiplication on the previous memory cell. Thus, the effect of the previous memory block is determined on the current LSTM cell block. The evaluation of the previous memory depends on the value of the output vector. When this value is near to zero, the previous memory cell is forgotten. The other gate is a section where the new memory is created by a simple neural network with the input gate, the tanh activation function and the effect of the previous memory block [29]. These processes are calculated with the Equations (4) and (5).

$$i_t = \sigma(W[x_t, h_{t-1}, c_{t-1}] + b_i) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \; tanh \; (W[x_t, h_{t-1}, c_{t-1}] + b_c) \tag{5}$$

The output gate is section where the output of the current LSTM block is produced. These outputs are calculated as in Equations (6) and (7).

The output gate produced by the output of the LSTM block cell is referred to as the output gate. It is shown in Equation (6) to calculate this output.

$$O_t = \sigma\big(W[x_t, h_{t-1}, c_{t-1}] + b_f\big) \tag{6}$$

$$h_t = \tanh(c_t) O_t \tag{7}$$

3.3b. *Proposed LSTM networks*: In this study, a new deep learning model named Electric Theft Detection- Long Short-Term Memory (ETD-LSTM) was developed during the classification phase of the proposed method. In addition to LSTM cells, Dropout layers, ReLU activation function between layers Softmax classifier were used. In addition to input parameters were determined according to different measures. The block diagram of the proposed model is shown in figure 4.

The prepared datasets are presented to the input layer of the first ETD-LSTM model. There are two LSTM cells after the input layer. A 20% dropout layer was used to reduce the number of data generated after calculating the data from LSTM cells with 64-unit weight values. The neurons obtained in the last LSTM cell were converted into one dimension with the Flatten layer and overfeeding was tried to be prevented by using the dropout layer again. Finally, all the parameters formed are given to the Softmax classifier with the Dense layer. With this layer, it is aimed to determine the class information of the data in the dataset. The parameter details of the proposed ETD-LSTM model is shown in table 4.

The proposed model has a total of 6 layers. In the dataset with 304 feature, two consecutive LSTM cell have been used to remember and process of the previous value. Especially for the unbalanced dataset, the values created by the weight update method developed in the data processing phase should be calculated with more neurons. The first LSTM cell gets more neuron as 64 Units. The subsequent LSTM layer are created more layers with 32 units. It is aimed to reduce by 20% the data generated by the dropout layer of the LSTM layers. In this way, it is possible to detect more prominent features by dealing with less parameters. With the Flatten Layer, all formed neurons are flattened. All these formed neurons are crossed with Dense layer with 256 new neurons and the best values are detected with ReLU layer. The data obtained with the Dropout layer have been reduced by 20%, and the most obvious neurons have been obtained. With the last layer, Dense layer, it is aimed to obtain class information of data with Softmax.

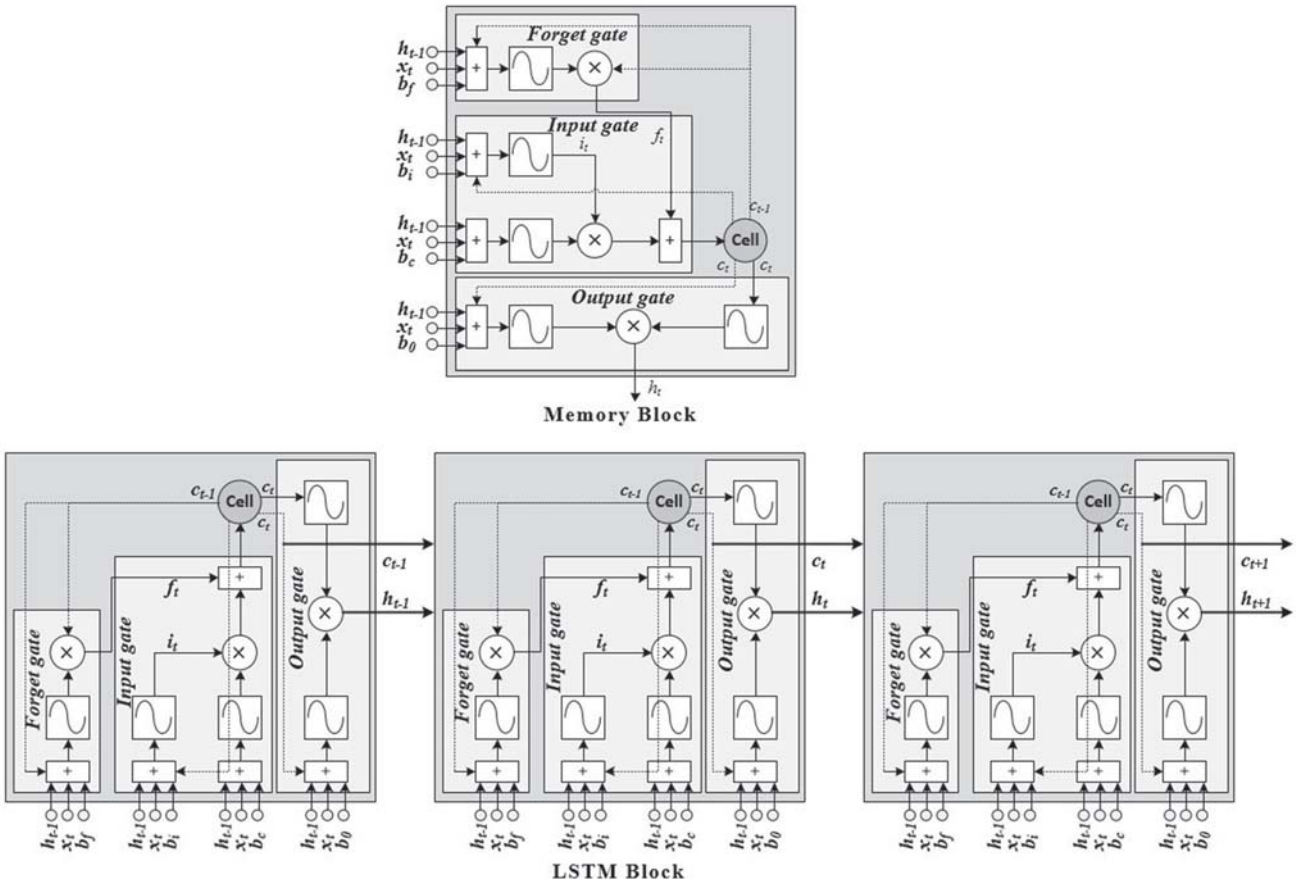

Figure 3. LSTM memory cell units and blocks.

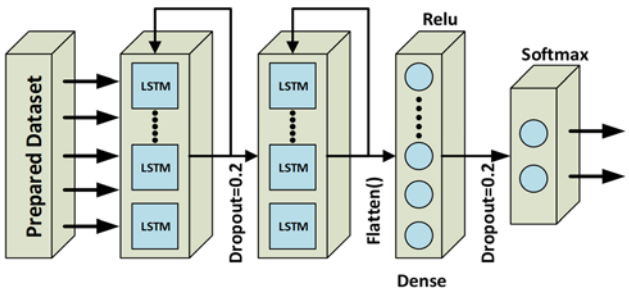

Figure 4. Block diagram of proposed ETD-LSTM model.

Table 4. Detailed properties of proposed LSTM network models.

Layers name	Main parameters	Other parameters
LSTM cell	64 unit	Dropout=0.2
LSTM cell	32 unit	Dropout=0.2
Flatten	–	–
Dense	256 unit	Activation=ReLU
Dropout	–	Rate=0.2
Dense	1 unit	Activation=Softmax

4. Experimental Detail

4.1 *Dataset*

Experimental studies: SGCCD database with two different label data including normal and NTL electricity consumer information was used. With the preprocessing, 33979 original electricity customer data with 304 feature were obtained for the first ten months of 2016 (304 days). The data numbers were divided to evaluate the performance of the proposed models during the training, validation and testing stages. Experimental data were divided into three different data in order of 60%, 70% and 80% in the training phase, 20%, 15% and 10% in the validation phase and 20%, 15% and 10% remaining in the test. The data numbers of all datasets are given in table 5.

When table 5 is examined, the dataset is divided into 3 different ratios to monitor the training and test success of the model. The model is trained with some of the data and the trained model is created. Unused test data were used in the same dataset to evaluate the performance of the trained model obtained. Test, validation and training were used to

Table 5. Number of training, validation and testing electric consumption samples.

	Dataset-1 (60, 20, 20)%			Dataset-2 (70, 15, 15)%			Dataset-3 (80,10,10)%		
Class	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
C1	4077	1360	1359	4757	1020	1019	5436	680	680
C2	4077	1360	1359	4757	1020	1019	5436	680	680
C3	4077	1360	1359	4757	1020	1019	5436	680	680
C4	4077	1360	1359	4757	1020	1019	5436	680	680
C5	4077	1359	1359	4756	1020	1019	5436	680	679
Total	20385	6799	6795	23784	5100	5095	27180	3400	3399

compare the experimental results and determine the best performance using different rates. The database was randomly dived into five folds. The distribution of the data in the data set can be seen in table 5. All data has 304 feature and label information. There are approximately 6797 data in each fold.

4.2 *Experiments on proposed ETD-LSTM*

Experimental studies were performed on a graphic card Tesla K80 with a 3.6 GHz Turbo GPU and 12 GB RAM. Various parameters have been determined for the designed ETD-LSTM models. Batch size value 64, loss function and Adam optimizer was selected as the categorical cross-entropy. Taken with 100 epochs to get consistent results and make comparisons. Stochastic gradient origin was used for the training of the model and initial learning rate was set to 0.0001 for less change. Accuracy, sensitivity and recall scores were used for performance evaluation of the developed model. Table 6 shows the Performance results of the ETD-LSTM model.

According to the cross-fold method determined on 3 different databases, the ETD-LSTM model we proposed was trained and after the validation was made with some of the data, the remaining part was tested. In this study, it was determined that the highest accuracy was Dataset-1 (Fold-4) with 95.24%. When looking at the performance in other folds, they show similarity with each other, which shows that dataset has an equal distribution and the preprocessing process is stable. It is seen that the standard deviation in accuracy is between ±1.09 and1.22 and the model shows close performance in different datasets. Dataset-1 (Fold-5) success was the second highest accuracy with 94.12. Also, high success was achieved in other folds. The average success of Dataset-1 was determined to be 93.60% ± 1.22. In Dataset-2, the highest success was in fold-3 with 93.24%, while the lowest success was in fold-4 with 90.22%. This indicates that the electricity usage trends of the same group of consumers are different. The average success of Dataset-2 was found to be 91.76% ± 1.09. Dataset-3's high success is seen in fold-2 with 92.62%. It can be seen that the number of data tested in Dataset-3, which has the lowest average with 90.98% ± 1.13, may be due to being low compared to other datasets. The highest precision success is seen in the Dataset-1(fold-4) with 92.82%, while the lowest is seen in fold-5 of dataset-3 with 88.08%. When we look at the recall value, Dataset-2 (Fold-4) are seen with the highest 93.20. The mean Recall scores were found to be 90.60% ± 1.38, 91.44% ± 1.78 and 88.23% ± 1.21. According to these results, it was determined that the highest accuracy was in fold-4 dataset and the highest average accuracy in Dataset-1with 93.60 ± 1.22. In addition, comparisons of the results we obtained in this study with similar previous studies [12, 15–17, 20, 22] are given in detail in table 7.

[12] has classified the electricity consumption data of 20,000 customers by using SVM and Role Engine

Table 6. Classification performance of the proposed ETD-LSTM model.

	Accuracy (%)			Precision (%)			Recall (%)		
Folds	Dataset-1	Dataset-2	Dataset-3	Dataset-1	Dataset-2	Dataset-3	Dataset-1	Dataset-2	Dataset-3
Fold-1	91.90	91.54	90.14	89.40	90.14	88.14	88.28	89.02	87.54
Fold-2	93.28	92.16	91.48	92.18	88.16	89.74	90.62	90.14	89.06
Fold-3	93.48	93.24	92.62	91.58	92.04	91.62	91.08	92.64	90.14
Fold-4	95.24	90.22	89.76	92.82	91.24	87.84	91.94	93.20	87.12
Fold-5	94.12	91.66	90.88	92.26	89.23	88.08	91.06	92.22	88.78
Mean and std.	93.60 ± 1.22	91.76 ± 1.09	90.98 ± 1.13	91.65 ± 1.33	90.16 ± 1.55	89.08 ± 1.61	90.60 ± 1.38	91.44 ± 1.78	88.53 ± 1.21

Table 7. Comparison of proposed method with previously studies results.

Method	Accuracy (%)	Other properties	Year	Dataset
SVM [12]	76.00–92.00%	20.000 customers	2013	–
Rule engine [12]	92.00%			
MP-ANN [17]	93.4%	5.000 customers	2020	Irish smart energy trial
PNN [20]	96.11%	660 customers	2018	PJM dataset
SVM [20]	94.39%			
SVM [22]	68.4–81.1%	4.000 customers	2019	CBT
CNN [15]	78.15–80.01%	42.372 customers	2018	SGCCD
CNN-LSTM [16]	89.00%	9.956 customers	2019	
CNN-LSTM [24]	87.90%	10.152 customers	2020	
GRU-LSTM [25]	89.00%	5.000 customers	2020	
ETD-LSTM (proposed)	93.60 ± 1.22 (Dataset-1)	33.979 customers		
	91.76 ± 1.09 (Dataset-2)	3.120 NTL data		
	90.98 ± 1.13 (Dataset-3)	5 cross-fold dataset		

algorithms in the method it has developed in order to encoder customer energy consumption data. It achieved SVM accuracy 76.00–92.00% for multiple iterations and Rule Engine accuracy 92.00%. On the Irish Smart Energy Trial database, [17] used artificial neural networks method (MP-ANN) and achieved a high success of 93.4%. Although it is different from the SGCCD database we use, it contains daily and hourly energy consumption values. [20] has classified on the PJM database using (Probabilistic Neural Network) PNN and SVM methods and achieved high success. [22] achieved 68.4–91.1% success on the CBT database using only SVM method. [15] and [16] used CNN and CNN-LSTM methods on the same dataset that we used in this study, and the highest success 89% (CNN-LSTM) was achieved. [24] used CNN-LSTM methods on SGCCD that we used in this study, and the highest success 87.9% (CNN-LSTM) was achieved. Using the GRU-LSTM model, using the data of 5000 customers in the same data, [25] achieved 89.00% success. With the ETD-LSTM model, which we have proposed, the data reduction has been achieved by making important pre-processing of the database and the separator has been developed, the customer types have been classified with high success rates. Our study has achieved high success (90.98–93.60%) compared to other studies, in addition, the study we have proposed in terms of the number of customers is more than all studies except the one examined [15]. It has been determined that data quality affects the success by obtaining results with different data rates.

5. Conclusions

In this paper, an LSTM model based on end-to-end learning for data preprocessing techniques and classification over the public SGCCD database containing normal and abnormal electricity consumption data is proposed. In the first stage of this new proposed method, qualified data in the database containing daily electricity consumption data was obtained. Because the database contains time series data, the deep learning based LSTM model was designed for high level feature extraction. The results obtained by the application of this method proved to be significantly higher. In addition, the effect and quality of the number of data were evaluated by testing all databases created at different rates with the cross-fold method. The results obtained exceed other results in the literature running in the same database. In future studies, we plan to create real-time applications and develop new methods for the detection of electrical theft.

Abbreviations

NTL	Non-technical losses
ETD	Electrical theft detection
LSTM	Long short-term memory
SVM	Support vector machine
SGCCD	State grid corporation of china dataset
CNN	Convolutional neural network
MP-ANN	Multilayer perceptron artificial neural network
RNN	Recurrent neural networks
GRU	Gated recurrent unit
k-NN	k-Nearest neighbors
ETD-LSTM	Electric theft detection- long short term memory

References

[1] McLaughlin S, Holbert B, Fawaz A, Berthier R and Zonouz S 2013 A multi-sensor energy theft detection framework for

advanced metering infrastructures. *IEEE Journal on Selected Areas in Communications.* 31(7): 1319–1330

[2] Bhattacharyya S C 2005 The Electricity Act 2003: will it transform the Indian power sector? *Utilities Policy*. 13(3): 260–272

[3] Ministry of power, G.O.I. Overview of power distribution. Available from: http://www.powermin.nic.in

[4] kayip kacak durumu. 2019; Available from: https://www.enerjiportali.com/yilmaz-elektrik-dagitim-sektorunde-kayip-orani-yuzde-5e-inecek/

[5] Yurtseven Ç 2015 The causes of electricity theft: An econometric analysis of the case of Turkey. *Utilities Policy*. 37: 70–78

[6] Electricity thefts on the rise. Available from: http://www.wtsp.com/news/local/story.aspx?storyid=109056

[7] Pulling the plug on energy theft, e.l.a.p. Pulling the plug on energy theft, electric light and power. Available from: http://www.elp.com/index/display/article-display/305071/articles/utilityautomation-engineering-td/volume-12/issue-9/features/pulling-the-plug-onenergy-theft.html

[8] Electricity theft by B.C. Grow-ops costs $100M a year. Available from: http://www.cbc.ca/news/canada/british-columbia/story/2010/10/08/bc-hydro-grow-optheftw.html

[9] Depuru S S, Wang L and Devabhaktuni V 2011 Support vector machine based data classification for detection of electricity theft. In: *2011 IEEE/PES Power Systems Conference and Exposition, pp. 1-8*

[10] Nagi J, Yap K S, Tiong S K, Ahmed S K and Mohamad M 2009 Nontechnical loss detection for metered customers in power utility using support vector machines. *IEEE Transactions on Power Delivery.* 25(2): 1162–1171

[11] Nagi J, Yap K S, Tiong S K, Ahmed S K and Nagi F 2011 Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. *IEEE Transactions on Power Delivery*, 26(2): 1284–1285

[12] Depuru S S S R, Wang L, Devabhaktuni V and Green R C 2013 High performance computing for detection of electricity theft. *International Journal of Electrical Power & Energy Systems*, 47: 21–30

[13] Costa B C, Alberto B L A, Portela A M, Maduro W and Eler E O 2013 Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process. *International Journal of Artificial Intelligence & Applications*. 4(6): 17–23

[14] Guerrero J I, León C, Monedero I, Biscarri F and Biscarri J 2014 Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection. *Knowledge-Based Systems*. 71: 376–388

[15] Zheng Z, Yang Y, Niu X, Dai H N and Zhou Y 2017 Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*. 14(4): 1606–1615

[16] Hasan M N, Toma R N, Nahid A A. Islam M M M and Kim J M 2019 Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. *Energies.* 12(17): 3310

[17] De Souza M A, Pereira J L R, Alves G de O, de Oliveira B C, Melo I D and Garcia P A N 2020 Detection and identification of energy theft in advanced metering infrastructures. *Electric Power Systems Research*. 182: 106258

[18] Gaur V and Gupta E 2016 The determinants of electricity theft: An empirical analysis of Indian states. *Energy Policy*. 93: 127–136

[19] Yip S C, Wong K, Hew W P, Gan M T, Phan R C W and Tan S W 2017 Detection of energy theft and defective smart meters in smart grids using linear regression. *International Journal of Electrical Power & Energy Systems*. 91: 230–240

[20] Ghasemi A A. and Gitizadeh M 2018 Detection of illegal consumers using pattern classification approach combined with Levenberg–Marquardt method in smart grid. *International Journal of Electrical Power & Energy Systems*. 99: 363–375

[21] Viegas J L, Esteves P R and Vieira S M 2018 Clustering-based novelty detection for identification of non-technical losses. *International Journal of Electrical Power & Energy Systems*. 101: 301–310

[22] Razavi R, Gharipour A, Fleury M and Akpan I J 2019 A practical feature-engineering framework for electricity theft detection in smart grids. *Applied Energy*. 238: 481–494

[23] Glauner P, Boechat A, Dolberg L, State R, Bettinger F, Rangoni Y and 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

[24] Adil M, Javaid N, Qasim U, Ullah I, Shafiq M and Choi J G 2020 LSTM and Bat-Based RUSBoost Approach for Electricity Theft Detection. *Applied Sciences*, 10(12): 4378

[25] Ullah A, Javaid N, Samuel O, Imran M and Shoaib M 2020 CNN and GRU based Deep Neural Network for Electricity Theft Detection to Secure Smart Grid. In: *International Wireless Communications and Mobile Computing (IWCMC),* pp. 1598-1602

[26] Glauner P, Meira J A, Dolberg L and State R 2016 Neighborhood features help detecting non-technical losses in big data sets. In: *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pp. 253–261

[27] Gers F A, Schraudolph N N and Schmidhuber J 2002 Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research.* 3: 115–143

[28] Hochreiter S and Schmidhuber J 1997 Long short-term memory. *Neural Computation.* 9(8): 1735–1780

[29] Yildirim Ö 2018 A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Computers in Biology and Medicine.* 96: 189–202