Detection of Non-Technical Losses(NTL) in Electric Distribution Network by Applying Machine Learning

DIMF G P
Research Scholar
Dept. of Information Technology, Manonmaniam Sundaranar University
Tirunelveli – 627 012, Tamilnadu, India

Dr. P. Kumar
Assistant Professor
Department of Information Technology/Information Technology and
Engineering, Manonmaniam Sundaranar University, Tirunelveli, 627012

Abstract.

This article discusses Non-Technical Loss (NTL) in power utilities and how to deal with it. Benefits of electric power utilities have changed as a result of non-technical loss. At the same time, due to widespread distribution, dishonest users and regular consumers have many similar consumption tendencies. Electricity theft, unauthorised connections, faulty metering, and billing issues are examples of non-technical losses. The NTL detection method's accuracy has to be improved urgently. Non-Technical Loss is identified using a classification scheme and data mining tools. Implement an intelligent computational technique to detect non-technical losses and choose the characteristic that is most closely related using information from a database of customer profiles. This work developed a machine learning-based non-technical loss detecting method. The performance of the suggested model is confirmed by the experimental results. These results unequivocally demonstrate that the proposed detection model outperforms other current methods in terms of accuracy, precision, recall, F1 score, and AUC score.

Keywords. Random Forest, Decision Tree, SVM, Data Mining, Non-Technical Loss

1. INTRODUCTION

Power networks are essential to the prosperity of any nation. Unfortunately, non-technical losses have a negative impact on these networks (NTL). Any utility faces a serious problem with NTL. Although it is estimated that these losses cost energy suppliers throughout the globe billions of dollars annually, eliminating NTL might increase revenue, profit, and grid dependability. As a result, the government is enthusiastic about NTL spending. The demand for electrical energy is rising as a result of urbanisation and rising living standards. The transition to electricity is being made using finite fossil fuels. The two categories of losses that happen throughout the production, transmission, and distribution of energy are technical and non-technical. Technical losses are brought on by internal resistance in the transformer, generator, and transmission lines. These losses range from 9 to 2 percent in Golden and Min[1] systems to 2 to 6 percent in inefficient systems, accounting for around 1-2 percent of the overall efficient energy distribution in Western Europe. According to Antmann[3], non-technical losses include power theft, problems with metre reading, record-keeping, accounting, and infrastructure failure or damage.

Theft of power might fall under one of the following categories: •A metre that is malfunctioning or damaged.

•Staying away from metering apparatus.

•A supply source that is unmetered.

•Measurement mistakes brought on by human and technological factors.

• Illegal actions, such tampering with metres.

The authors of references [14,15] claim that these losses cause around $1 billion in financial losses globally. This page was created because, when compared to technical losses, these losses account up a significant portion of overall losses.

To address the aforementioned difficulty using artificial intelligence, a great deal of effort has gone into implementing machine learning and deep learning techniques. Classification and clustering methods have been used to categorise existing machine learning techniques [5,6–8]. While human feature extraction is still required for existing machine learning detection methods, this suggests that they are unable to handle high-dimensional data such as standard deviation (SD), maximum, and lowest consumption statistics. To extract 2D attributes from smart metre data, manual feature extraction is a laborious and ineffective process. However, the bagging and random feature selection advantages of two machine learning methods are combined in the random forest (RF) classifier. As a result, there are several challenges associated with utilising machine learning to discover non-technical losses, such as class imbalance, data quality, technique comparison, feature design, and selection.

The structure of this document is as follows: In section 2, related works of literature were referenced. The results, as well as future study directions, are presented in Section 3.

2. RELATED WORK

In this part, we've covered the many NTL causes, the economic impact of NTL, the percentage of NTL in various nations, and the key works of literature on identifying non-technical losses.

In order to identify electricity and gas NTL, Coma-Puig et al. (2016) [3] utilised a range of machine learning strategies. They found that a single gradient boosted machine (GBM) outperformed any ensemble or classifier. Naive Bayes (NB), AdaBoost (AB), KNN, DT, NN, SVM, RF, and GBM were among the methods used.

Using data from a Chinese electrical firm, Zheng et alstudy .'s on deep convolutional neural networks in 2017 [18] (CNN). SVM, Random Forest (RF), Logistic Regression, and TSR are used to compare the results of deep convolutional neural networks (Three Sigma Rule). The classifiers listed before were outperformed by Deep CNN.

The 5K Brazilian Industrial Customer data collection is used by Ramos et al. (2013) [6]; each customer profile has 10 parameters, such as the maximum demand, demand billed, installed power, and so on. SVM, NN, and K-Nearest Neighbors (KNN) test accuracy is 0.9628, 0.9448, and 0.9620, respectively.

M. Hasan and colleagues (2019) [7] Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) were used for classification. The issue of class disparity is also addressed using the SMOTE approach. The simulation, which used a 10,000 customer data set, produced an 89 percent test accuracy. The remaining layers carried out the LSTM operation, whereas the first four hidden layers, each with 20 features, carried out the convolutional operation.

3. PROPOSED METHODOLOGY

The suggested method for NTL detection in the electric distribution organisation is explained in this section. We then discussed the necessity for specific performance assessment criteria for NTL detection. This proposed technique must be carried out in many stages. The next subsections provide a detailed description of these tactics.

Figure 1 depicts the suggested task's flow. The model is fed the information on the consumer's consumption metrics that was obtained from the service provider. The information is sorted using customer categories and meteorological data. Data cleansing, missing value imputation, and data transformation are done via feature engineering, which is also used for data pre-processing.

The division code and SDO code, among other attributes in the collected data set, are useless. These inconsequential attributes will be removed from the dataset. The k fold cross-validation strategy is used to train and test the model on the pre-processed dataset. There is an issue with data imbalance since there are many less defaulters in the data set than there are regular customers. The performance of the suggested model might be hampered by the data imbalance issue. To balance the data, we used the Megatrend Diffusion Function (MTDF) method.

Four machine learning classifiers are used in the proposed model. Several performance assessment measures acquired from test results offer a sufficient foundation for discovering various traits that identify the best classifiers to detect NTL [23].

A. Data Collection and Analysis

Without a dataset, NTL cannot be reliably detected. As a result, an actual dataset is obtained from a Tuticorin, Tamilnadu, India-based power distribution company. Statistics on consumer consumption from July 2019 to December 2020 are included in the data set. In all, there are 48754 records of monthly consumption. Utilizing customer groupings and weather kinds, the data is reviewed..

B. Data Pre-processing

There are 72 characteristics in the raw data that was received via the distribution tool. However, some of the parts have been shown to be worthless. For instance, all consumers have the same feature division code, SDO code, and load unit. As a result, certain functionalities may not be accessible. Other aspects that are employed for unique identification include "serial number" and "account id." Therefore, only one of them may be used to establish identification. Replace the null value in this step with the proper feature value. There were 19 criteria in all that were used by the suggested method.

Many characteristics in the obtained data contain some incorrect values during data pre-processing; these are known as outliers. In this study, the outliers are corrected using Equations 1 and 2.

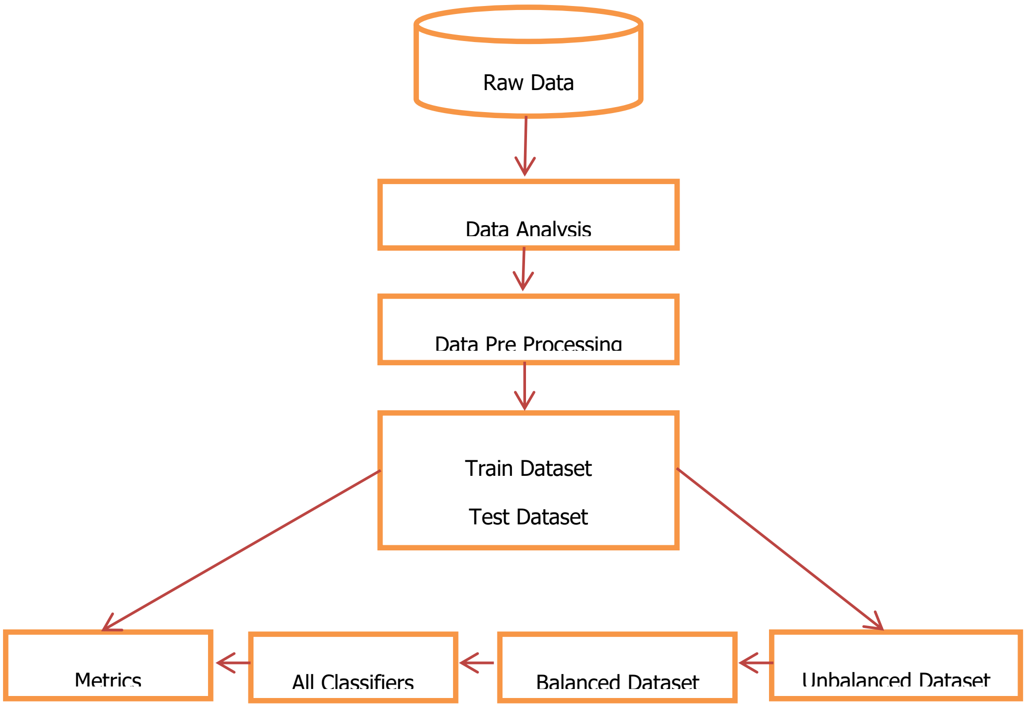

Figure 1. Flowchart of NTL detection

$$F(z) = \frac{x-\mu}{\sigma} \quad \text{--(1)}$$

The Z score, x, and mean values are the Z score, current feature value, standard deviation, and mean value, respectively. Upper(Z) and lower(Z) are determined as threshold values based on the kind of feature and standard deviation once the Z score has been established. In the end, Eq (2) is used to find and eliminate outliers.

$$f(x) = \begin{cases} x, \ upper(Z) >= x >= lower(Z) \\ mean(x), \quad else \end{cases} \quad \text{------------------ (2)}$$

The raw data acquired has several features with a broad range of values. As a result, normalisation is required before training and testing data can be used. Using Eq. (3), the current feature value, Vx, is normalised, where min(Vx) and max(Vx) are the current feature's lowest and highest values, respectively.

$$F(Vx) = \frac{\mathrm{Vx- min(Vx)}}{\mathrm{Max(Vx) – min(Vx)}} \quad \text{---------(3)}$$

C. Feature Selection

The raw data that was retrieved has 72 characteristics, although not all of them are required. As a result, relevant attributes from the master data are chosen right away. A feature's utility is determined by its prediction error; if the prediction error varies when the feature's value changes, the feature is helpful; if not, it is not. The 19 main attributes that were chosen as a consequence of this technique.

In addition to these characteristics, other metrics based on metre data, such as creditworthiness, are produced here (CWT). This function ranges from 1 to 5, depending on the customers' ignorance of or tardiness in paying the bill, the healthy consumer flag, the overload, the metre read comment, and the aberrant load consumption rate. This feature categorises customers into two groups: normal and aberrant. Five distinct CWT types are used in the proposed study, and their calculations are as follows:

$Overload = MDI - Load$ -----------------------(4)

The raw data that was retrieved has 72 characteristics, although not all of them are required. As a result, relevant attributes from the master data are chosen right away. A feature's utility is determined by its prediction error; if the prediction error varies when the feature's value changes, the feature is helpful; if not, it is not. This approach led to the identification of 19 essential traits.

In addition to these characteristics, other metrics based on metre data, such as creditworthiness, are produced here (CWT). This function ranges from 1 to 5, depending on the customers' ignorance of or tardiness in paying the bill, the healthy consumer flag, the overload, the metre read comment, and the aberrant load consumption rate. This feature categorises customers into two groups: normal and aberrant. The five distinct forms of CWT employed in the proposed study are computed as follows:

$$CWT2 = \begin{cases} 1, & HCF = yes \\ 5, & else \end{cases} \quad \text{--------} \; (5)$$

$$CWT3 = \begin{matrix} 5, & MRR = CDF \\ 1, & else \end{matrix} \quad \text{--------} \; (6)$$

CWT3 is determined by the metre read remark (MRR) status; if the MRR status is Ceiling Defective (CDF), CWT3 is 5, otherwise 1. Based on the cost of late payments, CWT4 (LPSC). By normalising it between 1 and 5, the LPSC value is transformed into a CWT4 value. Utilizing the load consumption rate, CWT5 is determined. These five CWTs may be used to categorise consumers in a number of ways.

D. Generation of train and test datasets

The train and test datasets are built using the k fold cross-validation method. The train dataset is used to train the recommended model's parameters, and the test data set is used to evaluate it. Data balancing is accomplished using the Megatrend Diffusion Function (MTDF) methodology.

E. Classification

Four classifiers were used for NTL detection, training, testing, and application. Using normalised data, these four classifiers—logistic regression, random forest, decision tree, and SVM—are trained and assessed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The suggested model is written in Python 3.6 and operates on an Intel Core i3 processor running at 3.4 GHz with 4.0 GB of RAM. Using Sci-kit Learn, the logistic regression, decision tree, SVM, and RF are all coded [16].

A. Performance Metrics

A discrete two-class classification problem is discovered for NTL. Each customer is thus placed into either the abnormal or normal category. Confusion matrices are produced during classifier validation. For NTL detection, four confusion matrices are used: a true positive matrix (TP), a false negative matrix (FN), a false positive matrix (FP), and a true negative matrix (TN) (TN). The number of customers who are properly categorised as normal, wrongly classified as abnormal, truthfully classified as normal, and correctly classified as abnormal is what these matrices are characterised as. The precision with which a classifier predicts the TP and TN values is another crucial assessment factor. Utilize Eq. (7) to evaluate accuracy; it is as follows..

$$Accuracy = \frac{\text{True Positive + True Negative}}{\text{True positive +True Negative + False Positive + False Negative}} \quad \text{---------}(7)$$

The proposed model's performance metrics are measured using some additional crucial evaluation metrics, as given in Eqns. (8) to (12).

$$\text{True Positive Rate(TPR)} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{True Negative (TNR)} = \frac{TN}{TN + FP} \quad (9)$$

$$\text{False Negative Rate(FNR)} = \frac{FN}{TP + FN} \quad (10)$$

$$\text{False Positive Rate(FPR)} = \frac{FP}{TN + FP} \quad (11)$$

$$\text{Positive Predictive Value(PPV)} = \frac{TP}{TP + FP} \quad (12)$$

The total number of power theft users that the classifier system successfully identified is determined using the recall or true positive rate (TPR). It becomes simpler to identify NTL as the TPR value rises.

The accuracy, recall, and F1 score for the classifiers for linear regression, support vector machines (SVM), decision trees (DT), and random forests (RF) are shown in Tables 2 through 5. A comparison of several classifiers based on accuracy, recall, and F1 score is shown in Table 6. Figure 2 shows its efficacy in contrast. The decision tree's accuracy, recall, and F1 score are 0.97, 0.97, and 0.97, respectively. With values of 0.98, 0.98, and 0.98 for recall, accuracy, and F1 score, respectively, Random Forest exceeds previous studies. Table 1 displays this difference. The accuracy, recall, and F1 score for both groups (normal and defaulter customers) are almost comparable, as shown in Tables 2–5, proving that the suggested model solves the issue of data imbalance.

Table 1. Classificationscoreof Different Classifiers

	Normal consumer	Defaulter consumer	Average/Total
Logistic Regression			
Precision	0.76	0.76	0.76
Recall	0.76	0.74	0.76
F1 Score	0.74	0.74	0.75
Decision Tree			
Precision	0.98	0.98	0.98
Recall	0.98	0.98	0.98
F1 Score	0.98	0.98	0.98
Support Vector Machine			

Precision	0.91	0.84	0.87
Recall	0.83	0.92	0.88
F1 Score	0.87	0.88	0.88
Random Forest(RF)			
Precision	0.98	0.98	0.99
Recall	0.97	0.98	0.97
F1 Score	0.99	0.99	0.99
Classification Score of LR, SVM, DT, and RF			
Logistic Regression	0.76	0.76	0.75
SVM	0.87	0.88	0.88
Random Forest	0.99	0.97	0.99
Decision Tree	0.98	0.98	0.98

A high harmonic mean score indicates the value of accuracy and memory. For the Decision Tree and Random Forest, the harmonic means of this suggested model are 97% and 98 percent, respectively.

This parameter's range is from +1 to -1. If the dominance value is close to +1, the classifier is accurate in the positive class. Additionally, the negative classifier seems to be accurate with a dominance score that is close to 1. The arithmetic mean and dominance of this suggested model are 98 percent and 0.006, respectively, as shown in Table 2.

Table 2: Parameter for LR, SVM, DT and Random Forest

Parameters	Linear Regression	Support Vector Machine	Decision Tree	Random Forest
Accuracy in %	0.76	0.87	0.98	0.99
Mean(Arithmetic)	0.76	0.86	0.98	0.99
Mean(Hormonic)	0.76	0.86	0.98	0.99
FNR	0.238	0.201	0.038	0.024
TNR	0.841	0.954	0.987	0.989
FPR	0.325	0.096	0.042	0.031
TPR	0.76	0.88	0.98	0.99
Domi	0.018	-0.041	0.002	0.008
AUC Score	0.845	0.894	0.98	0.99
MCC	0.54	0.896	0.96	0.97

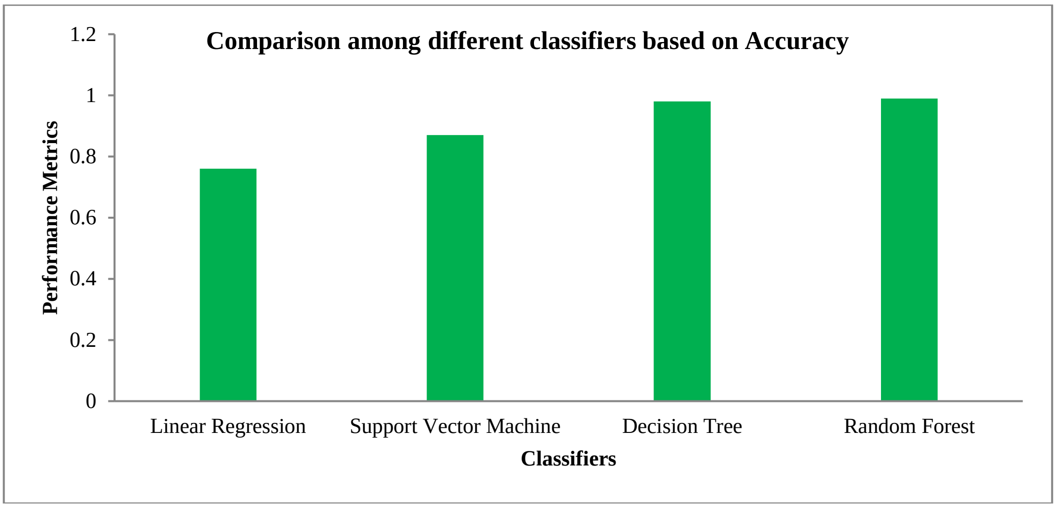

Fig.2 Comparison among different classifiers based on Accuracy

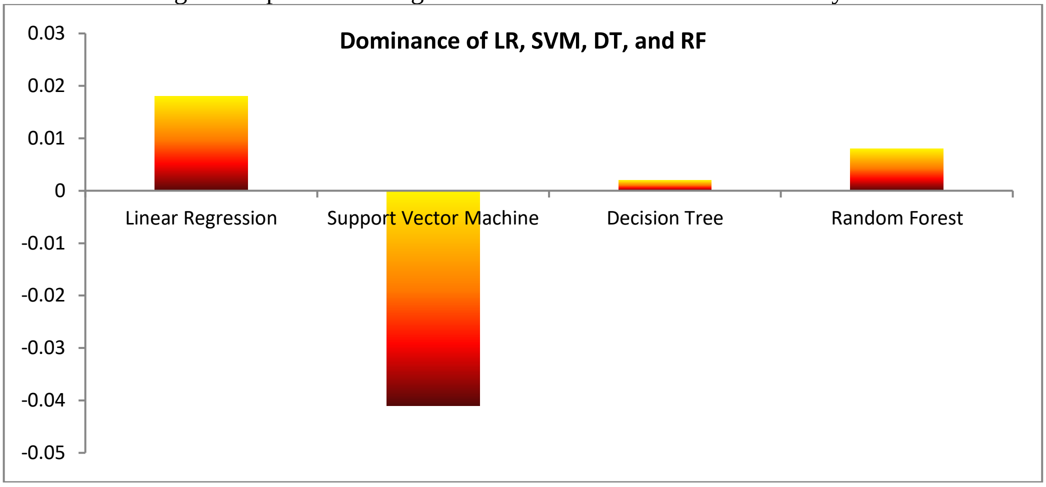

Fig 3. Dominance of LR, SVM, DT, and RF

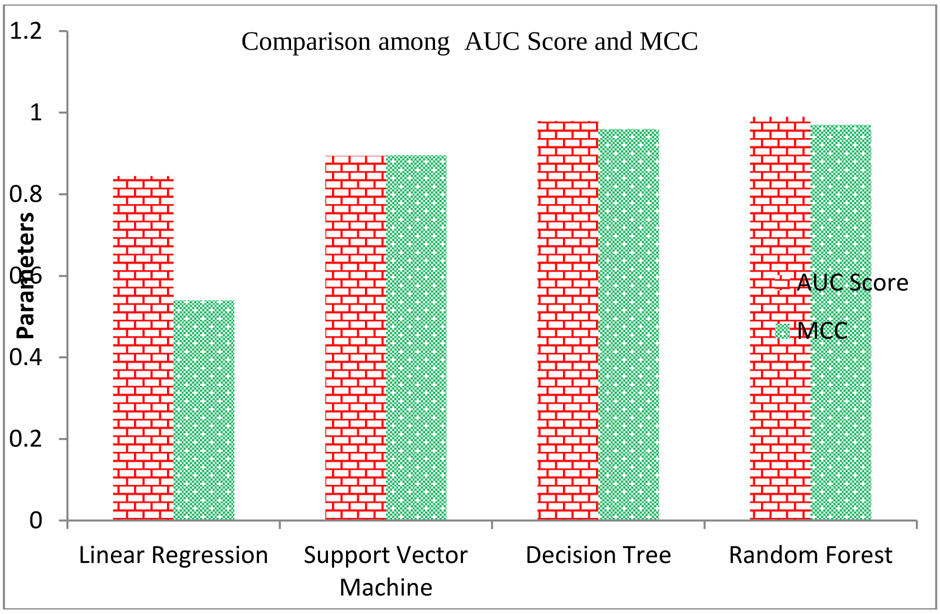

Fig 4. Comparison among AUC Score and MCC

Table 3 contrasts LR, SVM, DT, and random forest based on accuracy, arithmetic mean, harmonic mean, TPR, FPR, TNR, FNR, dominance, and AUC score. A comparison of classifiers utilising dominance metrics is shown in Figure 3. Figure 4 compares classifiers based on the TPR, TNR, and harmonic mean. The ROC curves for these three different classifiers are shown in Figure 5. The three terms are (LR, SVM, and RF). SVM, LR, and Figure 5 illustrates how Random Forest performs better than the opposition. The performance of the proposed model has been compared with that of earlier research, as shown in Table 5. Table 5 demonstrates that the new model performs better than earlier studies in terms of accuracy, recall, precision, and AUC score.

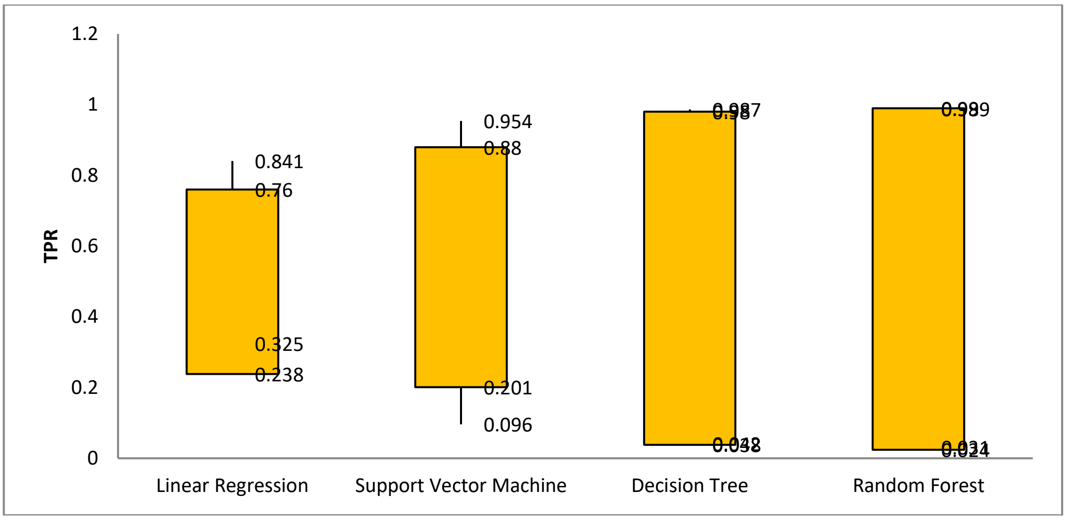

Figure 5. Stock diagram for LR, DT, and RF

Table 3. Comparisonbetweentheproposedschemeandexistedworks

Reference	Model	Accuracy	Recall	Precision	AUC
11	SVM(Gauss)	0.86	0.77	-	-
13	SVM+Fuzzy	-	0.72	-	-
14	SVM-FIS	0.72	-	-	-
15	FuzzyClassification	0.745	-	-	-
17	SVM	0.60	0.53	-	-
18	FuzzyLogic	0.55	-	-	-
19	CNN,LSTM	0.89	0.87	0.90	-
21	FuzzyClustering	-	-	-	0.741
22	Wideand Deep CNN	0.9404	-	-	-
23	DTcoupledSVM	0.925	-	-	-
25	(SVM,OPF, C4,5 tree)	0.862	0.64	0.544	-
26	CNN,LSTM	0.966	-	-	-
Our Proposed	LogisticRegression	0.75	0.75	0.75	0.749
	SVM	0.854	0.86	0.86	0.854
	DecisionTree	0.97	0.97	0.97	0.968
	RandomForest	0.98	0.98	0.98	0.98

This approach will provide the service provider a major edge in terms of identifying NTL. It will improve their capacity to detect NTLs while also saving them money, which is another important concern.

5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

A machine learning-based mechanism for recognising NTL is proposed in this study. This study made use of a data set from a distribution company in Tuticorin, Tamil Nadu, India. Theproposedworkconsidersthevariouschallengingissues,includingclassimbalance,dataquality,comparisonofdifferentmethods,featuredescription,andselection. The suggested model outperforms other existing studies in terms of accuracy, precision, recall, F1 score, and AUCscore, according to the comparison study. ThisworkwilldrasticallygiveaheavybenefittotheserviceprovidertodetectNTL. ItwillimprovetheirabilitiesforNTLdetectionandtheenormoussavingsofrevenuelosseswhichisalsoa seriousconcern. This investigation made clear how difficult it is to get a complete data collection from a distribution organisation. So,thereisaneedforpubliclyavailablearealdataset,whichcanbehelpinginthisareaofstudy.

REFERENCE

[1] Tooraj Jamasb, Tripta Thakur, Baidyanath Bag, "Smart electricity distribution networks, business models, and application for developing countries," Energy Policy,Vol. 114, Pages 22-29, 2018.

[2] Baidyanath Bag, Tripta Thakur,"A Utility Initiative based Method for Demand Side Management and Loss Reduction in a Radial Distribution Network Containing Voltage Regulated Loads," IEEE International Conference on Electrical Power and Energy Systems, Bhopal, Pages 52- 57, 2016.

[3] Government of India, Power Finance Corporation Ltd. Report on, " The Performance of State Power Utilities for the years 2016 “. [Online] Available: http://www.pfcindia.com/.

[4] Government of India, Power Finance Corporation Ltd. Report on, " The Performance of State Power Utilities for the years 2017, " [Online] Available: http://www.pfcindia.com/

[5] Venkatesh T, Trapti Jain, "Synchronized measurements-based wide-area static security assessment and classification of power systems using case based reasoning classifiers," Computers and Electrical Engineering, Vol. 68, Pages 513-525, 2018.

[6] Messinis, G.M. and Hatziargyriou, N.D.,"Review of non-technical loss detection methods, " Electric Power Systems Research, Vol.158, Pages 250-266, 2018.

[7] Villar-Rodriguez, E., Del Ser, J., Oregi, et al., "Detection of non-technical losses in smart meter data based on load curve profiling and time series analysis, " Energy, Vol.137, Pages 118-128, 2017.

[8] Viegas, J.L., Esteves, P.R., Melício, R., et al., "Solutions for detection of non-technical losses in the electricity grid: A review," Renewable and Sustainable Energy Reviews, Vol. 80, Pages 1256-1268, 2017.

[9] Gaur,V. and Gupta,E., "The determinants of electricity theft: An empirical analysis of Indian states," Energy Policy, Vol.93, Pages 127-136, 2016.

[10] Xiao, Fei, and Qian Ai, "Electricity theft detection in smart grid using random matrix theory," IET Generation, Transmission & Distribution, Vol. 12, Issue 2, Pages 371-378, 2018.

[11] Jokar, P., Arianpoo, N., Leung,V.C.M.,"Electricity theft detection in AMI using customer’s consumption patterns," IEEE Transaction on Smart Grid, Vol.7, Issue 1, Pages 216–226, 2016.

[12] Jindal, A., Dua, A., Kaur, K., et al.,"Decision tree and SVM-based data analytics for theft detection in smart grid," IEEE Transaction on Industrial Informatics, Vol. 12, Issue 3, Pages 1005–1016, 2016.

[13] P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "-e challenge of non-technical loss detection using artificial intelligence: a surveyficial intelligence: a survey," International Journal of Computational Intelligence Systems, vol. 10, no. 1, pp. 760–775, 2017.

[14] O. Rahmati, H. R. Pourghasemi, and A. M. Melesse, "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran region, Iran," CATENA, vol. 137, pp. 360–372, 2016.

[15] J. B. Leite and J. R. S. Mantovani, "Detecting and locating nontechnical losses in modern distribution networks," IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 1023–1032, 2018.

[16] M. Ismail, M. Shahin, M. F. Shaaban, E. Serpedin, and K. Qaraqe, "Efficient detection of electricity theft cyber attacks in ami networks," in Proceedings of the IEEE Wireless Communications and Networking Conference, Barcelona, Spain, April 2018.

[17] R. Mehrizi, X. Peng, X. Xu, S. Zhang, D. Metaxas, and K. Li, "A computer vision based method for 3D posture estimation of symmetrical lifting," Journal of Biomechanics, vol. 69, no.1, pp. 40–46, 2018.

[18] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," IEEE Transactions on Industrial Informatics, vol. 14, no. 4, pp. 1606–1615, 2018.

[19] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter datafication from smart meter data," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 2593–2602, 2019.

[20] Saeed, M.S.; Mustafa, M.W.B.; Sheikh, U.U.; Salisu,S.;Mohammed, O.O. Fraud Detection for Metered Costumers in Power Distribution Companies Using C5.0 Decision Tree Algorithm. J. Comput. Theor. Nanosci. 2020, 17, 1318–1325.

[21] Hasan, M.; Toma, R.N.; Nahid, A.A.; Islam, M.M.; Kim, J.M. Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. Energies 2019, 12, 3310

[22] C. C.O. Ramos, A. N. Sousa, J. P. Papa, and Alexandre X. Falcao. A new approach for nontechnical losses detection based on optimum-path forest. IEEE Transactions on Power Systems, 2011, pp. 181–189.