Available *online* at : http://jnte.ft.unand.ac.id/

Jurnal Nasional Teknik Elektro

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

ANN-Based Electricity Theft Classification Technique for Limited Data Distribution Systems

Monister Yaw Kwarteng, Francis Boafo Effah, Daniel Kwegyir, and Emmanuel Asuming Frimpong

Department of Electrical and Electronic Engineering, KNUST, Kumasi, Ghana

ARTICLE INFORMATION

Received: January 13,2023
Revised: March 30, 2023
Available online: March 31, 2023

KEYWORDS

non-technical loss, electricity theft detection, artificial neural networks, synthetic minority oversampling

CORRESPONDENCE

Phone: +233(0241)216264

E-mail: fbeffah.coe@knust.edu.gh

A B S T R A C T

Electricity theft has been a challenge for distribution systems over the years. Theft presents a massive cost to the system operators and other issues such as transformer overloading, line loading, etc. It has become crucial for measures to be implemented to combat illegal electricity consumption. This work sought to develop an artificial neural network-based electricity theft classifier for distribution systems with limited data, i.e., systems that can only provide consumption data alone and no auxiliary data. First, a novel data pre-processing method was proposed for the systems with consumption data only. Again, synthetic minority oversampling is employed to deal with the unbalance problem in the theft detection dataset. Afterwards, an artificial neural network (ANN)-based classifier was proposed to classify customers as normal or fraudulent. The proposed method was tested on actual electricity theft data from the Electricity Company of Ghana (ECG) and its performance compared to random forest (RF) and logistic regression (LR) classifiers. The proposed ANN-based classifier performed exceptionally by producing the best results over RF and LR regarding precision, recall, F1-score, and accuracy of 99.49%, 100%, 99.75%, and 99.74%, respectively.

INTRODUCTION

One of the significant concerns of distribution system operators is to reduce the overall losses of the distribution network to an acceptable level while maintaining maximum possible efficiency for delivery to consumers at proper voltages, frequency, and reliability. It is estimated that 40% of the total system losses of the electrical power network come from the distribution network, which is a vital link connecting the transmission network to the consumer. A significant portion of these losses is due to commercial (non-technical) losses, which represent energy supplied that is not billed, despite being reflected in the price paid by distribution system operators to transmission network operators [1], [2]. Non-technical losses represent a huge revenue loss to the system operators and countries since customers who steal power do not pay. It also adds to system operators' operational costs, such as paying penalties for transmission and distribution losses beyond regulatory bodies' thresholds.

Estimates show that utilities worldwide lose more than $25 billion yearly due to NTLs [3]. Specifically, Germany, Spain, and Italy lose about 504, 426, and 408 million Euros annually, representing 4.7%, 7.8%, and 6.3%, respectively, of their distribution revenue due to NTLs [4]. This clearly shows that a bulk amount of power is lost at the distribution end and therefore demands attention for it to be reduced, if not eliminated.

Conventional efforts by most power system operators worldwide to combat illegal consumption have been inspecting consumer premises to ascertain unlawful activity. The inspections are costly as personnel are trained, paid, and sent to such areas to carry out the checks. Some areas are also not easily accessible, especially in low-income countries.

Many methods have been proposed to combat and detect illegal consumption. These methods are grouped under theoretical studies, state-based (hardware) solutions, and machine-learning techniques. The theoretical studies and hardware solutions are expensive to implement. Machine learning (ML) models are currently used to detect and classify electricity consumers as normal or fraudulent [5], [6]. This method hinges on the characterisation and description of the software or an algorithm, which helps estimate and detect non-technical losses from consumer consumption data. The challenge with ML methods or models is that they perform poorly when the dataset is highly unbalanced, has low resolution with fewer attributes, and has a low relationship between input and output classes. Unfortunately, most distribution systems, especially in developing countries (such as Ghana), can only provide consumption data with low resolution (such as monthly

https://doi.org/10.25077/jnte.v12n1.1072.2023

consumption data) and little or no auxiliary data (e.g., latitude, longitude, altitude, and economic activity code) of customers. This causes machine learning models, such as artificial neural networks (ANNs) and random forests (RFs), to perform poorly in electricity theft detection and classification. Additionally, these features are costly to extract, which most distribution systems worldwide need help to afford. Finally, unbalanced data naturally exists in electricity theft detection datasets, with more normal customers than fraudulent customers. Therefore, there is a need for efficient data pre-processing methods to deal with these problems in customer data from distribution systems with limited data attributes, low resolution, and unbalanced data. This will improve the performance of machine learning models for electricity theft detection and classification.

Some works have been done over the years which employ machine learning methods to classify customers as usual or fraudulent. An ensemble extreme learning machine (ELM) algorithm based on ANN was proposed to detect abnormal electricity consumption in the network data [7]. The authors achieved an accuracy of 93.02% over the support vector machine (SVM) and K-nearest neighbour (KNN). The method also proved more accurate than ensemble SVM and KNN, with an accuracy of 97.51%. However, the algorithm requires more data variables which can be expensive to extract and that most distribution networks might need help to provide. Authors in [8] proposed a hybrid classifier model based on multi-layer perceptron (MLP) and long short-term memory (LSTM) for electricity theft detection. The authors achieved 54.5% area under the precision-recall curve (PR-AUC) when 80% of the data was used for training. However, the model requires more auxiliary data, which most distribution networks cannot provide. Again, the method did not address the unbalance problem in theft detection data. Also, the classifier produced a higher false positive rate, which is undesirable for machine learning classifiers.

Authors in [9] employed a convolutional neural network (CNN) and LSTM approach to classify customers as normal or fraudulent. The authors leveraged the consumption signature in time-series consumption data to build a CNN-LSTM model to classify the customers in the smart grid dataset. The proposed model achieved an overall accuracy of 88.82% before applying the synthetic minority oversampling technique (SMOTE) and 89.14% after using SMOTE. However, the overfitting and data duplication of SMOTE was not considered and dealt with. Authors in [10] proposed an ensemble learning method (ELM) algorithm based on ANN. The method achieved an accuracy of 93.02% over SVM. However, the algorithm requires more data attributes than just consumption, such as customer economic factors and longitude, which cannot be provided by some distribution systems and is also very expensive to extract. Again, they did not deal with the problem of unbalance in the theft detection dataset. In [11], authors employed a modified wavelet transform and random under-sampling boosting (RUSBoost) to detect fraudulent customers. The RUSBoost was used to handle the unbalance in the theft detection dataset. The method achieved a precision and recall of 81.87% and 82.63%, respectively. The RUSBoost reduced the data size, resulting in the models underfit. Again, the method requires auxiliary data, which most distribution systems cannot provide.

From the literature, it is clear that proposed machine learning methods for detecting electricity theft in distribution networks require more than just the consumption data of customers. Therefore, this work presents an ANN classifier with a novel SMOTE-based pre-processing method to classify electricity customers as normal or fraudulent. This will aid system operators in apprehending customers who steal power faster and reduce the cost of operation. The proposed method deals with the problem of class unbalance in the theft detection dataset and the problem of low resolution in the dataset from distribution systems with limited data. The contributions of this paper are:

- A comprehensive data pre-processing method has been proposed utilising a combination of statistical methods (quartiles, confidence interval), SMOTE and conditional formatting to improve the learning ability of ANN for electricity theft classification.
- The proposed pre-processing method improves the relationship between the input and output electricity theft classification data by dealing with the problem of low resolution and unbalance.
- An ANN-based electricity theft classification model has been developed that utilises the proposed data pre-processing technique to classify electricity customers as usual or fraudulent.

The remaining sections of this paper are as follows. Section 2 presents methodological steps to develop the novel pre-processing method and the ANN-based classifier. The implementation of the proposed method and the various software used for the research are also described in detail in Section 3. Results and analysis of the proposed ANN-based method are presented in Section 4, while Section 5 concludes.

METHOD

In most distribution systems, there is limited data in terms of data quantity and attributes such as; customer location, i.e., longitude, latitude, altitude, and customer economic activity index. Again, these data have many missing data points that translate into data incompleteness and hence present inefficiencies in ML classifiers. Also, most machine learning classifiers, such as ANN, need to learn better on sufficient data with fewer attributes; therefore, a new data pre-processing method for such a distribution system is presented. The novel pre-processing method proposed is shown in Figure 1.

Sample size determination

The minimum number of customers (regular and suspicious) needed to train and test the proposed theft classifier is determined using sample size calculation for a general population according to (1) [12]. This is done due to insufficient data in power distribution systems with limited data quantity and attributes.

$$n = \frac{N\left(z_{\alpha}/2\right)^2}{\left(z_{\alpha}/2\right)^2 + 4Ne^2} \quad (1)$$

https://doi.org/10.25077/jnte.v12n1.1072.2023

In (1), N is the population size (number of electricity customers living in a case study area), $Z_\alpha/2$ is the Z-score, α is 5%, and $4Ne^2$ is the margin of error. The sample calculated is used for training and testing the proposed classifier.

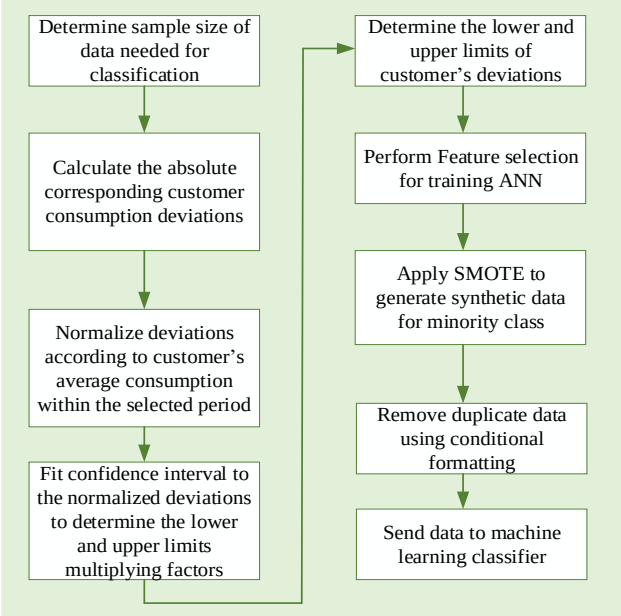

Figure 1: Proposed pre-processing model

Determination of absolute corresponding consumption deviations of customers

Absolute corresponding consumption deviations of a customer's previous and present consumption is determined to capture how a customer's consumption increases or decreases over time. The deviations are calculated for each customer in the sample size for a period *T*. In this work, monthly deviations are considered. Hence, the monthly absolute corresponding deviation d_{ij} of each customer *i* is calculated using (2).

$$d_{ij} = \left| C_{i(j+1)} - C_{ij} \right| \tag{2}$$

In (2), d_{ij} is the absolute monthly deviation of the current month's consumption $C_{i(j+1)}$ from the previous month's consumption C_{ij}. This is done for all customer consumptions in the period being considered. For *T* periods of customer consumption, there will be *T-1* deviations to be calculated. This captures how a customer's consumption changes within the period being considered.

Normalization of customers' corresponding deviations using average consumption

The deviations of each customer are normalized based on their respective average consumption according to (3).

$$d_{norm}(i) = \frac{d_{ij}}{C_{ave}(i)} \tag{3}$$

where $C_{ave}(i)$ is the average consumption of customer *i* and d_{ij} is the deviation.

Fitting of the confidence interval to customer normalized deviations

The confidence interval is fitted to the normalized deviations using (4). This is to confidently determine the range in which an unknown deviation of a customer in the selected dataset will fall and to confidently select appropriate lower and upper limit multiplying factors to determine outlier data points in customers' consumption data.

$$\bar{\alpha} - \left|\frac{t_a}{2}\right| \frac{s}{\sqrt{n}} < x_i < \bar{\alpha} + \left|\frac{t_a}{2}\right| \frac{s}{\sqrt{n}} \tag{4}$$

In (4), ***s*** is the standard deviation of the customer deviations, ***n*** is the number of customers in the datasets, $t_a/2$ is taken as 0.5 for a 99% confidence interval, $\bar{\alpha}$ is the mean of the customer deviations and x_i is the unknown customer deviation.

Determination of lower and upper limits of customer's consumption deviations

For every customer *i*, the first quartile Q_{1i}, third quartile Q_{3i} and the interquartile range IQR_i for their absolute consumption deviations for a period *T* are determined. These determine the lower and upper limits of the customer's deviations using (5) and (6). Each customer *i* is assigned an upper limit and a lower limit interval as:

$$lower\ limit = Q_{1i} - (\alpha_l \times IQR_i) \tag{5}$$

$$upper\ limit = Q_{3i} - (\alpha_u \times IQR_i) \tag{6}$$

The constants α_l and α_u are taken as the lower and upper limits of the confidence interval determined for the sample datasets. Hence;

$$\alpha_u = \alpha + \left|\frac{t_a}{2}\right| \frac{s}{\sqrt{n}} \tag{7}$$

$$\alpha_l = \alpha - \left|\frac{t_a}{2}\right| \frac{s}{\sqrt{n}}. \tag{8}$$

These are used because they confidently represent intervals for the corresponding deviations of the customers.

Feature selection for neural network training

Feature selection involves identifying relevant attributes in the data that will impact training performance. In this work, the BestFirst algorithm [13] in Weka software [14] is applied to the available attributes in the data to select the best. The customer data up to this point will have the following attributes; monthly customer deviations, customer normalized monthly deviations, customer consumption first quartile, third quartile, and inter-quartile range values. All these can be used for training, but there is a need to perform feature selection to select the best attributes which will give the best classification accuracy.

Generation of synthetic data using SMOTE

After feature selection, SMOTE is applied to the selected features to generate synthetic data for the minority class (suspicious customers) to balance the dataset of the selected attributes in terms of suspicious and normal electricity customers. SMOTE does this by choosing the K-Nearest-Neighbours of the minority class and interpolating new samples based on this using Euclidean Distances. The following steps create a new data point in the minority class. This process is carried out in Jupyter Notebook using python and the imblearn library [15].

Step 1: For each minority datapoint X_0 in the suspicious customer data, pick one of its K-nearest neighbours X belonging to the overall customer data.

Step 2: Create a new variable Z on a random point on the line segment connecting the datapoint and the selected neighbour as $Z = X_0 + w(X - X_0)$, where w is a uniform random number in the range [0,1].

Condition formatting to remove duplicate data

The condition formatting capability of EXCEL enables users to highlight or format data based on a certain condition set. This concept removes duplicate data after generating synthetic data for the minority class using SMOTE. A simple condition is set, which highlights the duplicated minority data point. The flow chart shown in Figure 2 removes duplicate data points. After conditional formatting, the resulting dataset is sent to the machine learning classifier for classification.

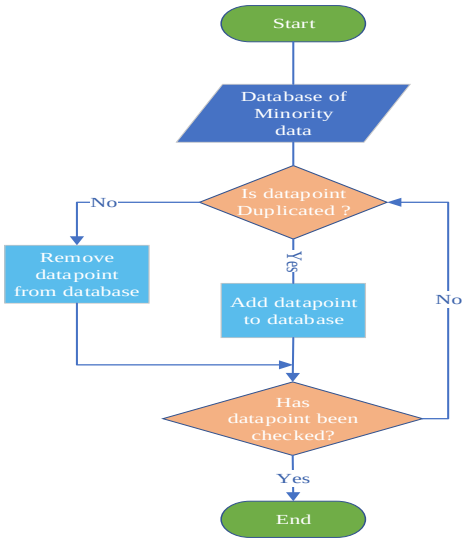

Figure 2: Flow chart for conditional formatting

Proposed ANN-based Electricity Theft Classifier

The proposed ANN-based theft classifier after data pre-processing is shown in Figure 3. The processes involved in the classification are outlined next.

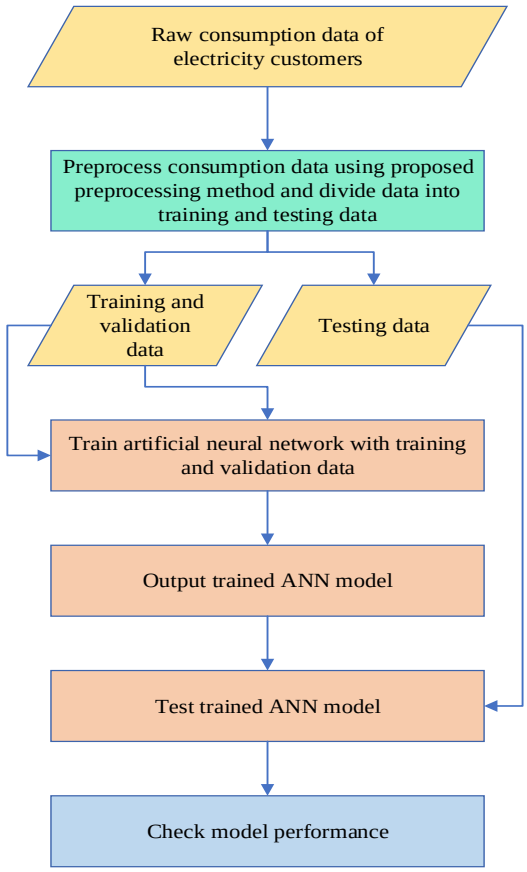

Figure 3: Proposed ANN-based classifier

Step 1: The customer sample size for the case study distribution system is calculated, and consumption data is obtained for the number determined.

Step 2: The proposed pre-processing method is applied to the consumption data of the sample size determined. The data attributes selected are then divided into training, validation and testing data for neural network training.

Step 3: The training and validation dataset is supplied to the ANN for training.

Step 4: The neural network is trained for a specified number of iterations or epochs until appropriate accuracy is obtained. The trained model is outputted for further testing with the testing dataset.

Step 5: The model performance is checked, and further prediction is done.

Implementation, simulation and testing of the proposed classifier

The implementation of the proposed model is summarized in Figure 4. The implementation is grouped under 5 headings; data collection, data pre-processing, feature selection, data classification, and validation of the proposed classifier. The proposed classifier for detecting suspicious consumption in a distribution network with limited data attributes was tested and validated using consumption data from the Electricity Company of Ghana, Dansoman District, Accra. Portions of the network in the district have a distribution transformer (DT) meter installed with 1245 customers on it. Twelve months of consumption data for each customer in the sample size was used. A sample size was used for the study because only some of the 1245 customers had

https://doi.org/10.25077/jnte.v12n1.1072.2023

consumption data available for the 12 months. The data included consumption data from customers who were apprehended as stealing power and regular customers as well. The sample size was determined to be 294 customers. Of these, 42 were electricity thieves (customers apprehended as stealing power), and 252 were regular customers (customers who have not been detained before). For each of the 294 customers, their ten months of consumption for 2018 were recorded from January to October for the study. The dataset was then made up of 294 rows and 10 columns.

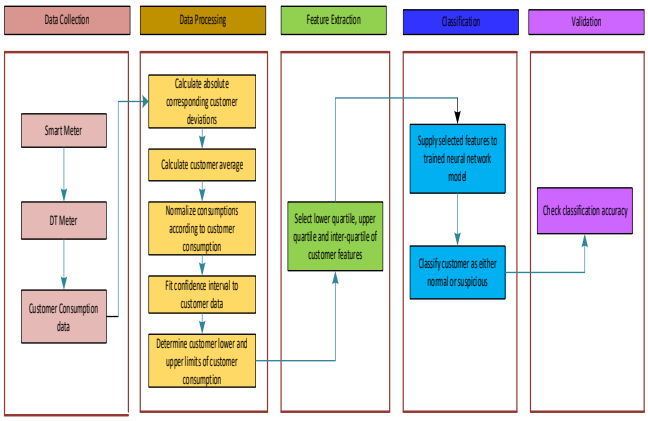

Figure 4: Implementation of proposed electricity theft classifier

The steps employed to apply the proposed model to the case study datasets are outlined below;

Step 1: After the sample size was determined, the absolute monthly corresponding deviations of each of the 294 customers were calculated. For each customer, nine deviations were determined for the ten months.

Step 2: The average consumption of 294 customers for the ten months was determined and used to normalise each customer's deviation.

Step 3: To choose the appropriate multiplying factors, α_u for the upper limit and α_l for the lower limit, a 99% confidence interval is fitted for the normalised deviation using EXCEL 2019. The results are shown in Table 1.

Table 1: Results of Confidence Interval Fitting

Parameter	**Value**
Sample mean	0.32
$Z_\alpha/2$	-1.98
Sample standard deviation (**s**)	0.64
Number of samples (**n**)	2646.00
Margin of error	0.02
lower limit	0.30
upper limit	0.35

From Table 1, the upper limit and lower limit factors for the case study dataset are $\alpha_u = 0.35$ and $\alpha_l = 0.30$, respectively. These determine each customer's consumption's lower and upper limit intervals according to (10) and (11).

$$lower\ limit = Q_{1i} + (0.30 \times IQR_i) \quad (10)$$

$$upper\ limit = Q_{1i} + (0.35 \times IQR_i) \quad (11)$$

Each customer's lower and upper limits are determined, and feature selection is done for neural network training.

Step 4: Feature selection determines which relevant features should be used to train the neural network. This is done with Weka software, a machine learning tool, and the BestFirst algorithm was used. Three attributes were supplied to the algorithm. These are the customer inter-quartile range, customer consumption lower and upper limits and their respective class labels (0=normal and 1=suspicious). The results of the feature selection are shown in Figure 5. The BestFirst algorithm selected customer consumption deviation upper limit (UB) and lower limit (LB) as the best features for training the neural network. LB captures the lower limit of customers' corresponding deviations, and UB captures the upper limit of customers' consumption deviations.

```
=== Run information ===

Evaluator:     weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:        weka.attributeSelection.BestFirst -D 1 -N 5
Relation:      New train data(IQRLBUB)
Instances:     292
Attributes:    4
               IQR
               UB
               LB
               output
Evaluation mode:    evaluate on all training data
```

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 6
        Merit of best subset found:    0.569

Attribute Subset Evaluator (supervised, Class (numeric): 4 output):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3 : 2
                     UB
                     LB
```

Figure 5: Results of feature selection in Weka software

Figures 6 and 7 show the plots of the dataset's normal and suspicious customers' lower and upper limits. The interval plot in Figure 6 indicates that the two customer groups are statistically different. This is true since the two groups have different means, and their confidence intervals do not overlap regarding lower bound values. The lower bound values of regular customers are higher (between 0.05 and 0.10) than those of suspicious customers. The interval plot in Figure 7 indicates that the two customer groups are statistically different. This is true since the two groups have other means, and their confidence intervals do not overlap regarding upper bound values. The upper bound values of regular customers are lower compared to suspicious customers.

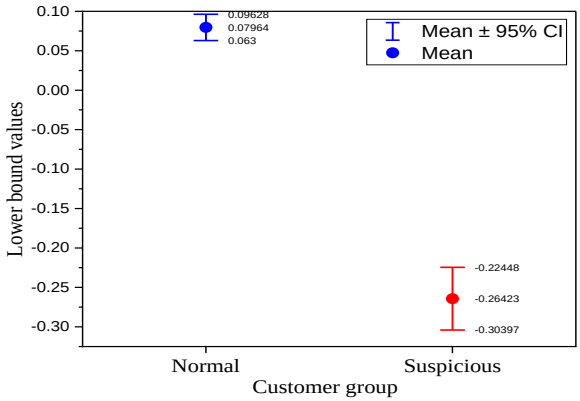

Figure 6: Interval plot of upper bound values of regular and suspicious customers

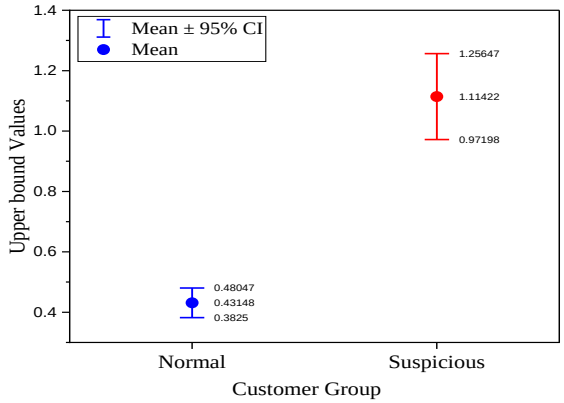

Figure 7: Interval plot of lower bound values of normal and suspicious customers

Step 5: Synthetic minority oversampling is carried out in Jupyter Notebook using Python programming language to increase the number of minority classes and the overall data from 294 to 375.

Neural Network Training

Training of neural networks was done in MATLAB using the pattern recognition application. The application classifies input data into target categories. It performs classification using a two-layered feedforward network with sigmoid hidden and SoftMax output neurons. The network was trained with scaled conjugate gradient backpropagation. The 375 data were divided into training, testing and validation of 80%, 10% and 10%, respectively. This division is done based on best practices in the literature [4]. This is done stratified to ensure that the same percentage of suspicious customers are present in each dataset (training, validation and testing). The neural network structure used for the training is shown in Figure 8. The network used for the training has 2 inputs representing the lower and upper limits of customer monthly consumption deviations, 10 hidden layers and 1 output layer representing the class label as a normal or suspicious customer. It should be noted that the same percentage division is used for training the neural network when the data is unbalanced. The model is trained for 1000 epochs.

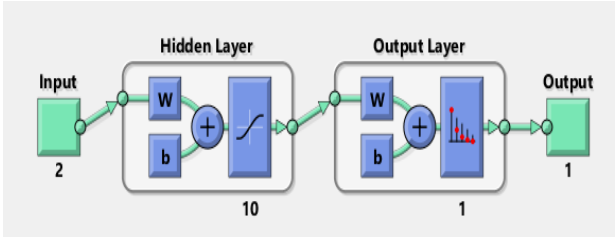

Figure 8: Neural network architecture for training the proposed classifier

Performance Evaluation of Proposed ANN-Based Classifier

The performance of the proposed classifier is evaluated using precision, recall, F1-score, and ROC-AUC. A confusion matrix is employed to determine the precision, recall and F1-Score. The various portions of the matrix are defined below.

i. True positive (TP): these are dishonest consumers accurately predicted as dishonest.
ii. True Negative (TN): these are honest consumers accurately predicted as honest.
iii. False Negative (FN); these are honest consumers predicted as electricity thieves.
iv. False Positive (FP): these are dishonest consumers predicted as honest consumers.

Precision shows the number of honest customers correctly classified by the model and is determined using (12).

$$precision = \frac{TP}{TP + FP} \tag{12}$$

Recall shows the number of positives correctly identified by the model and is determined using (13).

$$recall = \frac{TP}{TP + FN} \tag{13}$$

More than these are needed to evaluate classification performance with an unbalanced dataset; hence F1-score and ROC-AUC are added. F1-score is helpful as a measure for binary classification problems where the distribution of labels is unbalanced and is calculated by the weighted harmonic mean of precision and recall using (14).

$$f_1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{14}$$

ROC-AUC gives a graphical representation of a model to evaluate its classification performance. The classifier having ROC-AUC close to 1 has better performance. Again, the performance of the classifier is compared to the performance of logistic regression (LR), and random forest (RF), two well-known and accurate machine learning classifiers. Firstly, the performance is checked when the classifier is tested with data not balanced with SMOTE and when the data is balanced with SMOTE. Finally, RF and LR are all tested with the same case data.

https://doi.org/10.25077/jnte.v12n1.1072.2023

RESULTS AND ANALYSIS

The performance of the proposed ANN-based classifier is presented in this section. Also, the performance of this classifier is compared to LR, RF and other classifiers in the literature. The results are presented under four sub-headings next.

Performance of proposed Classifier when Tested with Unbalanced and Balanced Case Study Data.

Figure 9 compares the performance of the proposed classifier using unbalanced data versus balanced data with SMOTE.

For the classifier using unbalanced data, the precision is 43.47%, the recall is 90.91%, the F1-score is 58.82%, and the accuracy is 90.47%. For the classifier using balanced data with SMOTE, the precision is 99.49%, the recall is 100%, the F1-score is 99.75%, and the accuracy is 99.74%. Overall, the classifier using balanced data with SMOTE performs much better than the unbalanced data classifier. This shows that balancing the data can significantly improve the classifier's performance.

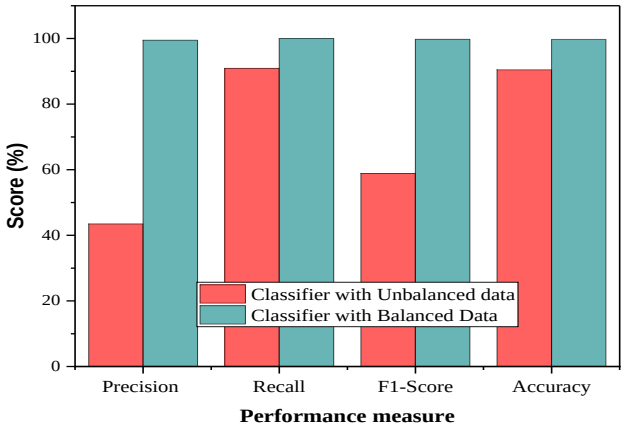

Figure 9: Comparison of performance of the proposed model with balanced and unbalanced data

Comparison of ROC-AUC Curves of Proposed Classifier for Balanced and Unbalanced Data

The ROC-AUC, which measures the strength of the proposed classifier to distinguish between regular and suspicious customers, is shown in Figure 10. The closer the value is to 1, the greater the ability of the classifier to perform the classification.

The proposed classifier produced a higher AUC value of 0.997 when trained and tested with balanced data than when trained and tested with unbalanced data. For the classifier without SMOTE, the AUC was 0.907. This validates the claim that ML classifiers perform well on balanced and well-processed data.

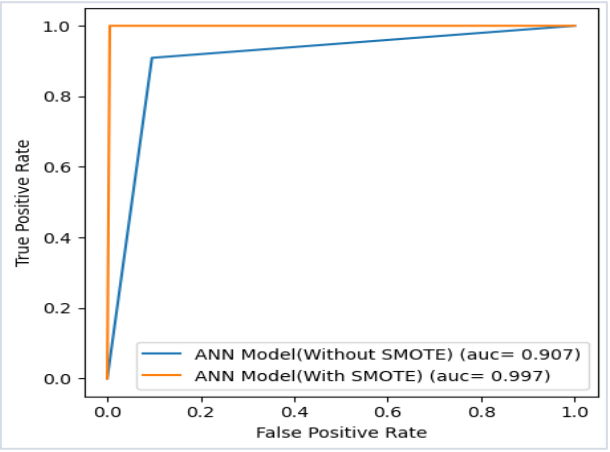

Figure 10: ROC-AUC Curve of Proposed Classifier for Balanced (WITH SMOTE) and Unbalanced (WITHOUT SMOTE) Data

Performance of Proposed Classifier Compared to Logistic Regression and Random Forest when Tested with Unbalanced Data

Figure 11 compares the performance of the proposed classifier with unbalanced data to two other classifiers: random forest and logistic regression.

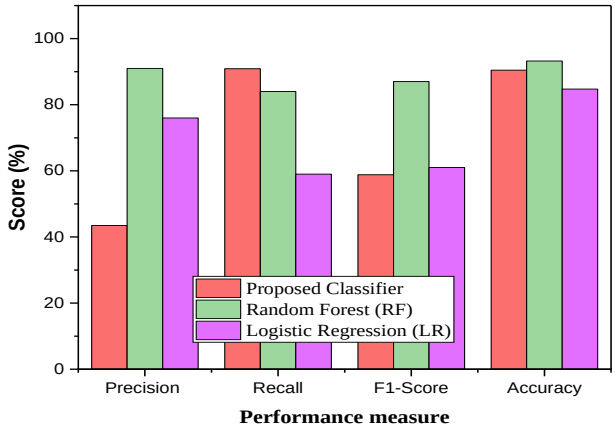

Figure 11: Comparison of Proposed Classifier to Random Forest and Logistic Regression for Unbalanced Data

For the proposed classifier, the precision is 43.47%, the recall is 90.91%, the F1-score is 58.82%, and the accuracy is 90.47%. For RF, the precision is 91.00%, the recall is 84.00%, the F1-score is 87.00%, and the accuracy is 93.22%. For LR, the precision is 76.00%, the recall is 59.00%, the F1-score is 61.00%, and the accuracy is 84.75%. Overall, the proposed classifier performs the worst among the three classifiers. RF has the highest accuracy and F1-score, while LR has the highest precision. However, the proposed classifier has the highest recall among the three classifiers. This also shows that ANN is more prone to poor performance on unbalanced data than RF and LR.

Comparison of ROC-AUC Curves of Proposed Classifier with RF and LR for Unbalanced Data

The ROC-AUC curves of the classifiers (RF, LR and proposed ANN-based classifier) are compared in Figure 12.

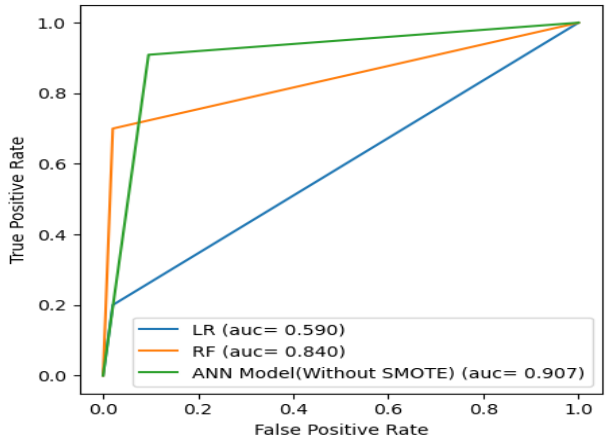

Figure 12: Comparison of AUC values for RF, LR and ANN-based classifiers for unbalanced data

The proposed classifier had the best AUC value of 0.907 compared to 0.840 and 0.590 for RF and LR, respectively. This shows that, although the proposed ANN-based classifier has low precision, it still had the most excellent capability to distinguish between the two classes (regular and suspicious customers) of the data compared to the others.

Performance of Proposed Classifier Compared to Logistic Regression and Random Forest when Tested with Balanced Data

Figure 13 compares the performance of the proposed classifier with balanced data to two other classifiers: random forest and logistic regression.

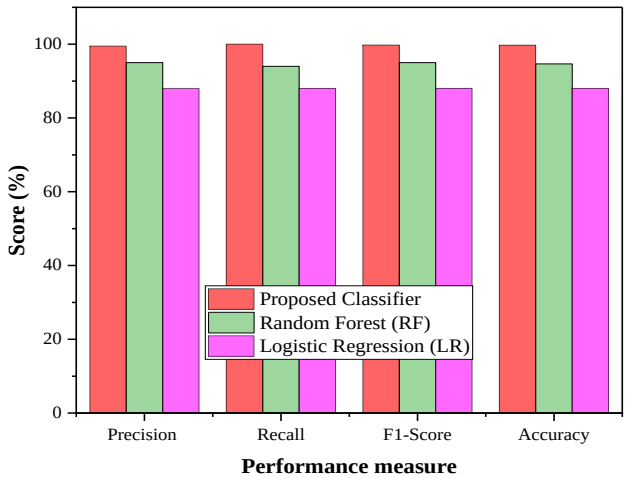

Figure 13: Comparison of the proposed classifier to RF and LR for balanced data

For the proposed classifier, the precision is 99.49%, the recall is 100%, the F1-score is 99.75%, and the accuracy is 99.74%. For RF, the precision is 95.00%, the recall is 94.00%, the F1-score is 95.00%, and the accuracy is 94.67%. For LR, the precision is 88.00%, the recall is 88.00%, the F1-score is 88.00%, and the accuracy is 88.00%. Overall, the proposed classifier performs the best among the three classifiers. It has the highest precision, recall, F1-score, and the second highest accuracy (behind RF). These results suggest that the proposed classifier performs strongly when using balanced data.

Comparison of ROC-AUC Curves of Proposed Classifier with RF and LR for Balanced Data

Figure 14 shows ROC-AUC curves comparing the performances of the proposed ANN-based classifier, RF and LR.

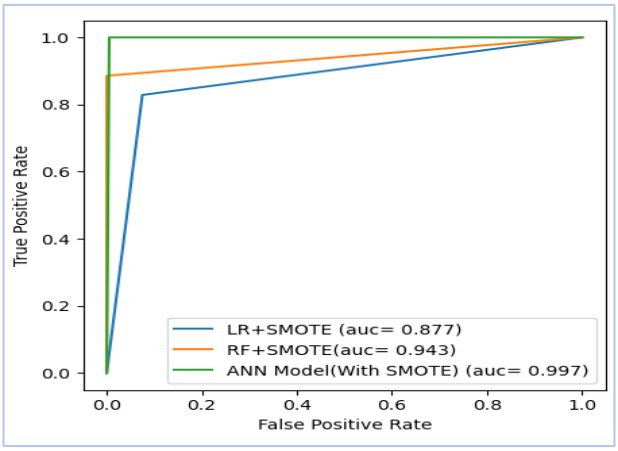

Figure 14: Comparison of AUC values for RF, LR and ANN-based classifiers for balanced data

From Figure 14, the proposed ANN-based classifier gave the best AUC value of 0.997 compared to 0.877 and 0.943 of LR and RF, respectively. Again, there was a general improvement in the ability of the classifiers to classify the customers as normal and suspicious after applying SMOTE to the dataset.

CONCLUSIONS

An ANN-based classifier using a novel data pre-processing method for distribution systems with limited data has been proposed. The performance of the proposed classifier was assessed by testing it on balanced and unbalanced theft classification datasets and comparing it to RF and LR classifiers. It was concluded that the proposed classifier generally performed exceptionally well when trained and tested with a balanced dataset compared to the unbalanced dataset. It obtained precision, recall, F1-score, and accuracy of 43.47%, 90.91%, 58.82%, and 90.47%, respectively, on unbalanced data and 99.49%, 100%, 99.75%, and 99.74% precision, recall, F1-score, and accuracy on balanced data, respectively. This represents an average improvement of 28.83%.

Again, the performances of the proposed ANN classifier, LR, and RF could have been better when trained and tested with unbalanced data. However, RF performed better than LR and the proposed ANN-based classifier. RF produced an average score of 88.81% compared to 70.19% for LR and 70.93% for the proposed ANN-based classifier for precision, recall, F1-score, and accuracy, respectively. However, the proposed model had the best AUC value of 0.907 compared to 0.840 and 0.590 for RF and LR, respectively. This shows that, although the proposed ANN-based classifier has low precision, it still had the most significant

https://doi.org/10.25077/jnte.v12n1.1072.2023

capability to distinguish between the data's two classes (regular and suspicious customers) compared to the others. Also, RF, LR, and the proposed ANN-based classifier generally performed well when trained and tested with balanced data. However, the proposed ANN-based classifier produced the best results over RF and LR with precision, recall, F1-score, and accuracy of 99.49%, 100%, 99.75%, and 99.74%, respectively. This represents an increase of 56.02%, 9.09%, 40.93%, and 9.27% in the performance of the ANN-based classifier in terms of precision, recall, F1-score, and accuracy, respectively, over the situation when the model was tested with unbalanced data. The ANN-based model gave the best AUC value of 0.997 compared to 0.877 and 0.943 of LR and RF, respectively.

In conclusion, the performance of a classifier can be significantly affected by the balance of the data it is trained on. When using unbalanced data, the proposed classifier performed worse than the random forest and logistic regression classifiers. However, when using balanced data, the proposed classifier performed the best among the three classifiers, with the highest precision, recall, F1-score, and second-highest accuracy. These results suggest that balancing the data can significantly improve the performance of a classifier, and the proposed classifier is a strong performer when using balanced data.

REFERENCES

[1] P. R. Babu and B. Sushma, "Operation and control of electrical distribution system with extra voltage to minimize the losses," *Proc. 2013 Int. Conf. Power, Energy Control. ICPEC 2013*, pp. 165–169, 2013, doi: 10.1109/ICPEC.2013.6527643.

[2] L. Marques, N. Silva, I. Miranda, E. Rodriges, and H. Leite, "Detection and localisation of nontechnical losses in low voltage distribution networks," *IET Conf. Publ.*, vol. 2016, no. CP711, 2016, doi: 10.1049/cp.2016.1079.

[3] A. Hatem Tameem Alfarra, B. Amani Attia, and C. S. M. El Safty, "Nontechnical loss detection for metered customers in alexandria electricity distribution company using support vector machine," *Renew. Energy Power Qual. J.*, vol. 1, no. 16, pp. 468–474, 2018, doi: 10.24084/repqj16.353.

[4] M. Hashatsi, C. Maulu, and M. Shuma-Iwisi, "Detection of electricity theft in low voltage networks using analytics and machine learning," *2020 Int. SAUPEC/RobMech/PRASA Conf. SAUPEC/RobMech/PRASA 2020*, 2020, doi: 10.1109/SAUPEC/RobMech/PRASA48453.2020.9041117.

[5] M. Madrigal, J. J. Rico, and L. Uzcategui, "Estimation of Non-Technical Energy Losses in Electrical Distribution Systems," *IEEE Lat. Am. Trans.*, vol. 15, no. 8, pp. 1447–1452, 2017, doi: 10.1109/TLA.2017.7994791.

[6] I. Bula, V. Hoxha, M. Shala, and E. Hajrizi, "Minimizing non-technical losses with point-to-point measurement of voltage drop between 'SMART' meters," *IFAC-PapersOnLine*, vol. 49, no. 29, pp. 206–211, 2016, doi: 10.1016/j.ifacol.2016.11.103.

[7] Z. Fang, Q. Cheng, L. Mou, H. Qin, H. Zhou, and J. Caol, "Abnormal electricity consumption detection based on ensemble learning," *9th Int. Conf. Inf. Sci. Technol. ICIST 2019*, pp. 175–182, 2019, doi: 10.1109/ICIST.2019.8836863.

[8] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Hybrid Deep Neural Networks for Detection of Non-Technical Losses in Electricity Smart Meters," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1254–1263, 2020, doi: 10.1109/TPWRS.2019.2943115.

[9] M. Nazmul Hasan, R. N. Toma, A. Al Nahid, M. M. Manjurul Islam, and J. M. Kim, "Electricity theft detection in smart grid systems: A CNN-LSTM based approach," *Energies*, vol. 12, no. 17, pp. 1–18, 2019, doi: 10.3390/en12173310.

[10] C. Tsai, K. Chiang, H. Hsieh, C. Yang, J. Lin, and Y. Chang, "Feature Extraction of Anomaly Electricity Usage Behavior in Residence Using Autoencoder," 2022.

[11] N. F. Avila, G. Figueroa, and C. C. Chu, "NTL Detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7171–7180, 2018, doi: 10.1109/TPWRS.2018.2853162.

[12] "Power and Sample Size Determination." https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_power/bs704_power_print.html (accessed May 31, 2022).

[13] "Best First Search Algorithm in AI | Concept, Algorithm and Implementation." https://www.mygreatlearning.com/blog/best-first-search-bfs/ (accessed Jan. 10, 2023).

[14] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench," *Data Min.*, pp. 553–571, 2017, doi: 10.1016/b978-0-12-804291-5.00024-6.

[15] "Over-sampling methods — Version 0.10.1." https://imbalanced learn.org/stable/references/over_sampling.html#smote-algorithms (accessed Jan. 10, 2023).