Peertechz

ENGINEERING GROUP

TRENDS IN
Computer Science and Information Technology
OPEN ACCESS JOURNAL

ISSN: 2641-3086 DOI: https://dx.doi.org/10.17352/tcsit

Research Article

Detection of electricity theft in developing countries-A machine learning approach

Leonardo Grant[1,2]*, Haniph Latchman[1,3] and Kolapo Alli[1]

[1]Faculty of Engineering, the University of the West Indies (Mona), Kingston, Jamaica

[2]Jamaica Public Service Co. Ltd, Kingston, Jamaica

[3]Electrical and Computer Engineering, University of Florida, Gainesville Florida, USA

Received: 27 July, 2023
Accepted: 17 August, 2023
Published: 18 August, 2023

***Corresponding author**: Leonardo Grant, Faculty of Engineering, the University of the West Indies (Mona), Kingston, Jamaica, Tel: +1 604 790 5738; E-mail: leonardo.a.grant@gmail.com

ORCiD: https://orcid.org/0000-0002-7820-8042

Keywords: AutoML; Feature extraction; Machine learning; Non-Technical losses

https://www.peertechzpublications.org

Check for updates

Abstract

In developing countries, energy theft negatively affects the growth of utilities through loss of revenue and damage to the grid. The size and variety of the utility data set require extracting meaningful features to counter theft, which is difficult and computationally expensive. Recent developments have made machine learning more accessible to researchers, enabling its application in big data analysis for power utilities. Through greater access to training resources, as well as commercial and open-source machine learning tools, it has become easier to test large sets of data against various algorithms and automate many of the processes such as data cleaning and feature extraction, a procedure known as Automated Machine Learning (AutoML). These tools, along with frequent data collection by utilities, lend themselves to the use of machine learning to solve power grid issues such as anomaly detection. This paper focuses on feature extraction from monthly consumption records, previous investigations, and other customer information to detect power anomalies critical in the detection of theft. Using AutoML, features were extracted, and models were then trained and tested on data gathered from investigations. The results show that by using machine-learning algorithms, anomaly detection can be 4 times more effective than present manual detection techniques, increasing from 10% to 40% while reducing the number of unnecessary audit investigations by 61%.

Abbreviations

NTL: Non-Technical Losses; ML: Machine Learning; AMI: Advanced Metering Infrastructure; DFS: Deep Feature Synthesis; TSFresh: Time Series Feature extraction based on Scalable Hypothesis tests; GLM: Generalized Linear Model; GBT: Gradient Boosted Tree

Introduction

Electricity generation and distribution enabled rapid development throughout the 20th century, powering technology in the fields of information, communication, and computing. As technologies such as the internet, computers, and microprocessors advanced, the power grid remained relatively unchanged until the early 21st century when the incorporation of those and other technologies, in the form of grid sensors, communication protocols, and data analysis, are now being used to tackle some of the greatest challenges faced by modern power grids. In developing countries, particularly in Latin America and the Caribbean, one of those is the problem of energy loss, in particular Non-Technical Losses (NTL). This can be defined as the illegal abstraction (use, waste, or diverting) of electricity without paying for it [1]. While this is a serious issue globally, it is extremely pervasive in poorer nations where these losses can sometimes account for up to 50% of the energy produced, and this is energy that ends up being paid for by regular customers or the government through subsidies [2]. This is an issue that costs Latin America up to USD 17 billion and the island of Jamaica, in particular, USD 244 million annually [3]. A solution to this problem suitable to the conditions in the developing world is vital for the growth, development, and economic equity of electricity in these countries.

Approaches to NTL detection have been researched extensively in a variety of different countries. The solutions

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

generally fall into 3 categories: 1) Data-oriented, which focuses on analyzing primarily consumption data on an entity to determine the likelihood of energy theft by that entity. 2) Network oriented, which uses grid sensors and other specialized equipment in conjunction with network topography to detect NTL. 3) Hybrid, which is a combination of data and network-oriented approaches [4]. The method best suited for a particular utility depends on the level of electrical infrastructure already available in the country or the utility's ability to quickly implement the associated system. In many countries, that process has already been completed or is currently underway, requiring between 3 to 15 years, depending on a variety of factors such as size, population, and financial resources [5]. The required infrastructure, advanced communication, meter technology, and dedicated sensors are very costly [6]. This puts those methods out of the reach of countries with little or no existing smart grid development or whose deployment is slowed by lack of capital, whether due to low investment or high theft [7]. A study of the implementation of smart meters puts the cost of the infrastructure at USD 54 million per 1 million power meters [5]. Since utilities cannot wait to upgrade the entire grid before they tackle the overwhelming power theft issue, a data-oriented approach is the primary solution available to many developing countries.

As part of the technical and business processes involved in providing electricity to customers, large volumes of data have been generated by utilities, ranging from monthly information, such as billing and consumption data, to fixed information, such as customer details and the results of previous investigations [3]. Through the increased use of communication and computational technology in every process, from the metering to the billing cycle, researchers have begun to store and analyse that data on a large scale. The increased data processing power and increased access to Machining Learning (ML) tools and techniques have given rise to the use of big data analytics to mine customer data to detect and reduce energy theft, as is well described in work done by Guerrero et al and Han, et al. [8-10]. This approach is the most popular due to the low cost of implementation and has been used with both monthly meter readings and interval data from Advanced Metering Infrastructure (AMI) systems in conjunction with machine learning algorithms, with varying degrees of success. Whereas some implementations perform better than others do, it is worth noting that all of them perform better than their respective manual approaches to NTL detection.

Automated NTL detection is of particular interest for many researchers, due to the increased volume of customer data available, especially customer meter information. In the past, the only interaction with the customer was a monthly consumption reading and the bill that followed, along with some demographic information. With smart meters, the utility is now able to view a customer's consumption with a much higher resolution, moving from once a month to every 15 minutes. That, along with various meter events that may be registered by the meter, gives the utility insight into not just a customer's energy use but also the various states the meter goes through at the customer's location. This high-frequency interval data lends itself to big data analytics and has been used in studies by numerous researchers, as seen in some works [10-16], which use 15-minute to hourly interval data. Medium-resolution data in the form of daily readings are used in a few studies [17-20]. Other researchers have also used even lower resolution monthly consumption and applied similar analytical techniques to large volumes of data, which, though less accurate, in many cases still perform better than random guessing and human experts using the same data [21-31], have also used even lower resolution monthly consumption and applied similar analytical techniques to large volumes of data, which, though less accurate, in many cases still perform better than random guessing and human experts using the same data.

In NTL detection, various techniques have been used but most are a variation with some improvements on core methods, with researchers sometimes combining methods to improve the likelihood of detection. Techniques broadly fall into two categories, namely supervised and unsupervised learning.

A. Supervised learning

Many utilities routinely conduct customer visits, allowing them to assign a label to a portion of their customer base. Supervised learning relies on labelled data, where customers are designated as either normal or anomalous, to train algorithms to detect NTL. It is the most commonly used data-oriented approach. The most popular algorithms used are Support Vector Machines [11,14-16,18,23-25,27-29,32], Random Forest [16,21,25-28], Boosted algorithms [19,20,22,28,30,31], Neural Networks [17,18,22,27], K nearest neighbour [25,27,28,32] and Logistic Regression [25,27,32].

B. Unsupervised learning

These methods work without any labels in the training data to learn from the information hidden within it, allowing for more complicated processing tasks [33]. The labels are only required once the algorithms need to be evaluated. This tends to be used in clustering or outlier detection algorithms, such as Optimum-Path Forest (OPF), k-means, Self-Organizing Maps (SOM), Multivariate Gaussian Distribution (MGD), and Local Outlier Factor (LOF), to detect the anomalous customers in the data set [12,13]. For example, the OPF was able to outperform other clustering algorithms such as k-means, in terms of its accuracy score, when trained for anomaly detection [12]. MGD also performed well, achieving the best F-measure on one of the datasets [12] and performing the best overall [13].

C. Challenges in data-oriented approaches

Class imbalance and evaluation metrics, covariate shift, data quality, feature selection, scalability and comparability of various methods are some of the challenges faced by researchers [34]. The current literature focuses primarily on just one or two of these challenges, primarily class imbalance. Most of this research is being done in highly developed countries that feature significantly less NTL. This means that much of this data can often be less practically applicable to high NTL areas that would most benefit from the implementation of NTL detection and mitigation strategies.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

Methodology

This research focuses on, *class imbalance*, in conjunction with *feature selection* and *evaluation metrics* [34], *Class imbalance* has received attention in work done by the authors in [16-20] and [24-28] and involves using different sampling techniques with a training data set. Synthetic Minority Oversampling Technique (SMOTE), which generates samples of the minority class, and Radom Undersampling which ignores part of the majority class, are the main method to balance the data set. *Feature selection* has received less attention, with different features, such as averages, sums, and standard deviations, commonly extracted from consumption data, being used by various researchers. No standardized feature list or extraction method exists. Figure 1 shows a flow chart of the methodology proposed for our implementation of data-oriented NTL detection. This proposed solution is to address these specific challenges by introducing techniques to deal with *class imbalance*, *feature extraction*, and *evaluation metrics*.

A. Class imbalance

The most common method of handling class imbalance is sampling, particularly under-sampling, where the data set is artificially balanced by removing samples of the dominant class from the training set. This reduces the number of training examples but can still increase the performance of the algorithm. Previous research has shown that the increase in NTL proportion improves the performance of the algorithm when tested on an unbalanced data set and reduces the likelihood of predicting false negatives to maintain high accuracy. This can be seen in [15-22] and [24-28], all works focused on data set imbalance, some using sampling and others by the performance metrics. To handle the class imbalance, the models were trained at 10%, 33%, and 50% NTL. Each was then tested on a data set with 10%, 20% and 50% NTL, which represent low, average, and high loss areas respectively.

Sampling is a key part of the approach used in this research to manipulate anomalies in the training and test sets. This data set maintained an 80:20 train/test ratio with the percentage of anomalies varying within each set while the ratio remained constant. For example, if 100 audits were sampled, 80 would be in the training set and 20 in the test set. If the training set had 10% irregularities 8 of the 80 audits would be anomalous, if it has 50% irregularities, 40 of the 80 would be anomalous. For the test set if 10% irregularities 2 of the 20 audits would be anomalous, if it has 50% irregularities, 10 of the 20 would be anomalous.

B. Feature extraction

This research also focuses on the use of automated extraction to address the feature selection problem, and this distinguishes it from previous approaches which were manual in nature. This has clear benefits for the two main groups that would be involved in the development of the NTL models, domain experts, and machine learning experts. Domain experts are empowered to create ML processes without in-depth knowledge of data science, while ML experts can automate tedious tasks, reducing their workload and improving efficiency. Three of the most popular automated open-source feature extraction algorithms were chosen to take advantage of their ease of implementation as well as their unique characteristics. The three methods were: (1) Deep Feature Synthesis using the Featuretools Python library, (2) Time-Series Feature Extraction based on Scalable Hypothesis tests using the Tsfresh Python library, and (3) RapidMiner's Time Series Feature Extraction plugin.

1) **Deep Feature Synthesis (DFS):** Deep Feature Synthesis was first proposed by Kanter and Veermachaneni in 2015, its main advantage is its ability to generate features from relational data sets and present these as a rich feature space [35]. As most companies rely on relation databases for information storage this allows the creation of features automatically from various related sources. It has also been used for other anomaly detection systems, such as an intrusion detection system using convolutional neural networks [36].

2) **Time Series Feature extraction based on Scalable Hypothesis tests (Tsfresh):** The Tsfresh library is specifically designed for supervised learning, combining feature extraction with filtering important features [37], removing the redundant features, and leaving the ones most closely related to the predicted label. It achieves this through a combination of hypothesis testing and feature significance testing. Tsfresh has an advantage when used with time series sensor data but can also be used for high interval data with some success, and this is useful as consumption data which is typically used for NTL detection is a time series data set. Xu [38] used the Tsfresh library with the XGBoost algorithm (a form of gradient boosting), with monthly consumption data to predict future consumption with high accuracy.

3) **RapidMiner's Time Series Feature Extraction:** RapidMiner combines data preparation, feature

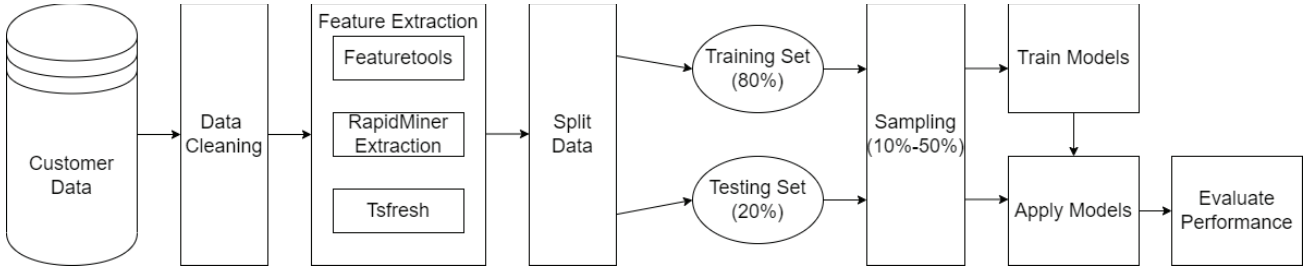

Figure 1: Proposed NTL Detection Strategy.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

extraction, and easy deployment of various machine learning algorithms into one unified platform [39]. It has been used to design autonomous data mining processes [40] in analytical chemistry for predicting the linear retention indices among various compounds [41], in medicine for patient classification in relation to liver disease, and in aircraft classification using air traffic data [42]. The main advantage of using the time series extension is direct integration with the tool that is being used for the rest of the machine learning process, creating a single automated pipeline from data collection to model evaluation. Using this method, we extracted features from four 3-month periods, each representing a quarter of a year.

The ancillary data of the customer account was used as the base table, joined to the consumption, billing, and exceptions tables by an identification number was then used to extract the features of interest from the data.

Each feature extraction method produced its own unique feature matrix with 125 features in the case of DFS, 550 in the case of Tsfresh, and 50 in the case of RapidMiner's Time Series Feature Extraction. In all cases, RapidMiner's feature score was used to select the features most highly correlated with the label. The feature vectors can be expressed in the form $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$.

C. Data

The current research uses data related to customers' accounts to train a model to predict NTL. The data is based on historical records from 454,000 audits done on residential and small commercial accounts between 2016 and 2021. These audits provide an indicator of whether or not an anomaly was found, as well as the classification of the type of anomaly found. Preliminary analysis of the data showed significant increases in the number of audits, standardization of labelling, and anomaly categorization starting in May 2018. This made cleaning and preparation of the data easier, thus the data set was further reduced to the last audit done on accounts between May 2018 and May 2021, to take advantage of the increased data quality. The resulting data consisted of 246,000 audits with 24,500 anomalies representing a precision of 10% using manual account selection and domain knowledge only. For the 246,000 accounts, four distinct data sets were collected: consumption, billing, exception, and ancillary.

The consumption data was analysed to find the minimum time period during which models still performed well. And it was found that the last 12 months of consumption in kWh fit that profile. Additionally, the reading type (estimated or actual read), and the number of days in the billing cycle to account for the difference between readings were also used. Consumption is the primary data type used in NTL detection. Many features can be extracted from this data, but additional information can also be gained by analysing the consumption of a given area as seen in Ref [21,22,25] and [26]. The billing data contained the last 12 months of bills and the payments made to those bills were collected. Exceptions represented any special event on the meter, such as excessively high or low readings over the last 12 months. Ancillary data provide additional information about the accounts, such as the location and type of account.

Figure 2 shows the average consumption of normal customers over a year compared to the average of all customers in the same community. It can be observed that the average consumption of all customers in a community is slightly lower than the average of just the normal customers. This is expected as anomalous customers would skew the overall average downwards due to their full consumption not being registered on their meters.

Changes in the customer's energy usage, such as sharp declines, may be indicative of theft but also might be indicative of some change in behaviour or conditions. Figure 3 shows the average consumption of anomalous customers over a year compared to that of all customers in the same community and reveals a wide gap between those anomalous customers and average neighbourhood consumption. By adding neighbourhood consumption, we can differentiate changes in consumption that occur due to factors that affect everyone, such as higher consumption that occurs during Christmas or summer, and then focus on changes caused by individual behaviour.

D. Evaluation metrics

Metrics were specifically selected based on their relevance to the problem of NTL detection particularly that of a highly imbalanced data set. Metrics such as accuracy may have deficiencies including high false-negative rates but still appear accurate since most accounts are normal, thus failing to detect anomalies [3]. On the other hand, Using the *True Positives* (TP),

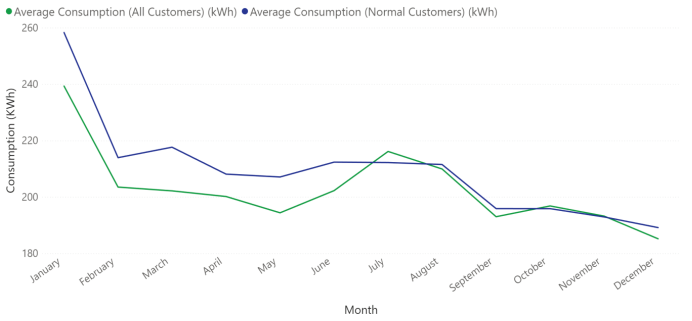

Figure 2: Average Consumption of Normal Customers *vs.* Average Consumption of All Customers over a Year.

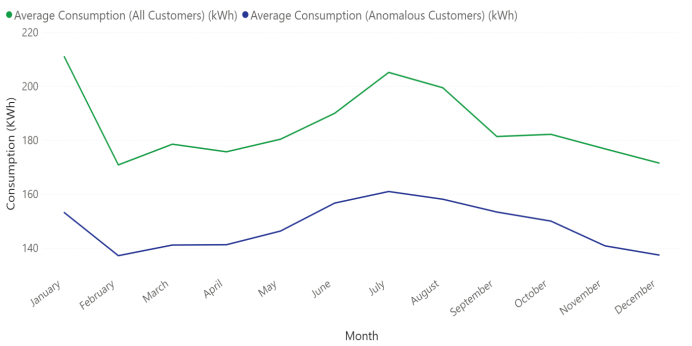

Figure 3: Average Consumption of Anomalous Customers *vs.* the Neighbourhood Average over a Year.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

False Positives (FP), *True Negatives* (TN), and *False Negatives* (FN), we can calculate the following more meaningful metrics:

- ***Precision*** – the fraction of all investigations done at a customer's premises that detected actual anomalies.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Recall, also known as the **detection rate** – the fraction of the total anomalies present in the data set that the model correctly predicts as anomalous.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

- ***Area Under the Curve (AUC):*** A performance metric for binary classifiers; it ranges between 0 and 1, with an AUC > 0.5 indicating the prediction is better than a random guess [24]. It is represented as a plot of the Recall against the specificity. Hosmer *et al* further break down AUC scores into 5 categories, = 0.5 which is considered random, 0.51 to 0.69 which is poor but slightly better than a coin toss, 0.7 to 0.79 being acceptable, 0.8 to 0.89 classified as excellent, and 0.9 and above is outstanding [43].

- ***Recoveries***: an underused metric, though they can be considered the most important one for the utility. The audit stops future energy theft but also customers are back billed for the energy extracted without payment; hence a model that can correctly identify a few accounts with large recoveries may be more useful than one that identifies a larger number of low recovery accounts from a financial perspective [13].

E. Algorithms

Four (4) algorithms were tested in this experiment, namely, Deep Learning, Generalized Linear Model (GLM), Gradient Boosted Tree (GBT), and Naïve Bayes Model.

1) **Deep learning:** RapidMiner's implementation of Deep Learning is a multi-layered feed-forward Artificial Neural Network (ANN) that employs stochastic gradient descent using backpropagation and is often used in applications such as image and speech processing [44]. This implementation's architecture consists of one input layer, two hidden layers with 50 neurons each, and one output layer, it uses the rectified linear activation function to calculate the output of each node. The authors in Ref. [9] use ANNs to mine artifacts from customer data that are used as features for a Rule-Based Expert System. ANNs have also been used in several other NTL studies [6]. The model seeks to minimize the error function

$$E = \frac{1}{2}\sum_i (t_i - y_i)^2 \quad (3)$$

Where *ti* represents the target and *yi* represents the output. The algorithm is represented in Figure 4.

2) **Generalized Linear Model (GLM):** GLMs are statistical models derived from a weighted linear regression [45]. Through generalization, they expand their scope to deal with nonlinear data by transforming the nonlinear relationship between input and output into a linear one. This is done by using a function to do the transformation [46] and is implemented using binomial regression as the transformation function. The general form of that function being

$$g(\Pi_i) = \sum_{j=0}^{p} \beta_j * x_{ij} \quad (4)$$

Where $g(.)$ is the link function, βj is the regression coefficients, x_{ij} is the regression variables and π_i is the conditional expectation of y on $X = x_i$.

3) **Gradient Boosted Tree (GBT):** GBTs are an improvement on decision trees using boosting, a method that aims to strengthen weak learners [47]. These weak learners are tree ensemble models created by combining decision trees sequentially to create a stronger learner [48]. This is done through the construction of additive regression models which fit those learners to the gradient of a loss function [49]. Boosting approximates the function that maps feature to the output by the expansion of the form

$$F(x) = \sum_{m=0}^{m} \beta_m h(x; a_m) \quad (5)$$

Where $\{\beta_m\}^M$ represents the expansion coefficient and $h(X; a)$ is chosen as a simple function of X with the set of parameters a = {a1, a2, ...}.

4) **Naive bayes:** Naïve Bayes is a classification algorithm that works on the assumption that features are independent of each other and, though this is rarely true in the real world, the algorithm generally has lower error rates than more complex classifiers [50]. It's based on conditional probability theory, modelling the posterior probability and estimating the class density

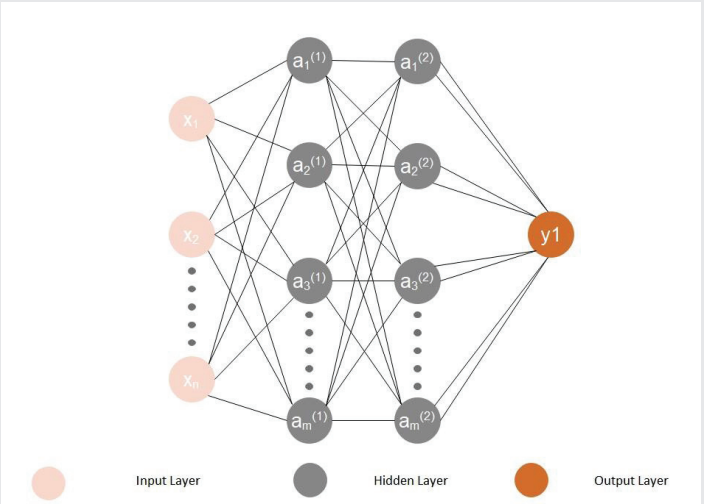

Figure 4: Diagram of a Multi-Layered Feed-Forward Artificial Neural Network (ANN).

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

of the predictors using Bayes' rules [51]. That posterior probability is denoted by

$$P(C \mid X) = \frac{P(C)\Pi_{i=1}^{n} P(x_i \mid C)}{P(X)} \quad (6)$$

Where C is the class of an observation X assuming the features X_1, X_2,..., Xn are conditionally independent.

F. Implementation

Data from a relational database consisting of customer information from 2016 to 2021 was extracted and anonymized. This data was then pre-processed in the RapidMiner platform to clean and reduce the data to the interval between May 2018 and May 2021. The consumption data was then further reduced to the last 12 months before the most recent audit. Python scripts for Featuretools and Tsfresh were then created to complete the feature extraction. The RapidMiner Time Series Feature Extraction plugin was also used to create a feature extraction process that split the data into 4-month windows.

Model training, testing, and evaluation processes were then set up in RapidMiner to automate the rest of the machine learning process. Once the features were extracted the data was randomly under-sampled at 10%, 33%, and 50% NTL proportions for both training and testing data sets. This data was then used to train each model at each NTL proportion. The test data set was then used to evaluate the model at each NTL proportion and the result was recorded.

Results

Distinguishable trade-offs between recall and precision can be seen in most algorithms. Figure 5 shows that for all models, there is an increase in recall as the NTL proportions increase in the training data set, except for Naïve Bayes. In the latter case, there is a decrease for models trained on the RapidMiner and Tsfresh extracted features but a slight increase when Featuretool's features are used. Naïve Bayes produced the highest recall with the Tsfresh features, in some cases, it was as high as 0.998 but this was done by predicting most of the accounts as anomalous causing the corresponding precision to fall to 0.101. That approach would trigger many unnecessary audits resulting in wasted resources.

The Precision did decrease with the increase in the NTL proportions as seen in Figure 6, but in all cases, it was still greater than what had been achieved by manual account selection. GLM achieves the highest Precision also with Tsfresh; the model was able to score 0.944 with a Recall of 0.013 and detected a very small subset of anomalies very well but missed most of the others.

Figure 7 shows that this trend reverses when tested on data sets with higher proportions of NTL. Due to the larger number of anomalous accounts, a model trained on high NTL proportions can detect losses better.

The ability of various models to detect an anomaly as measured by the AUC can be seen in Figure 8 and generally remained the same when GLM and GBT are the algorithms used, no matter what feature extraction is selected. For Deep Learning, the highest AUC was achieved by training models at 33%, close to the 26% at which losses stand in the real world. The AUC produced by Naïve Bayes is less than that of all the other algorithms.

The Recoveries follow the same general trend as Recall; therefore a better way to look at Recoveries is not just as a percentage of the total energy recovered but as how much energy is recovered per audit done. Figure 9 shows a general trend towards high recoveries per audit when trained at lower NTL proportions. The GLM shows one such case detecting a few instances of NTL without expending too much effort, in which a few high-loss accounts were identified.

Figure 10 shows the comparison of the confusion matrix for GBT and Deep Learning both trained at 33% NTL and tested at 20% NTL using the Featuretools extraction method. These

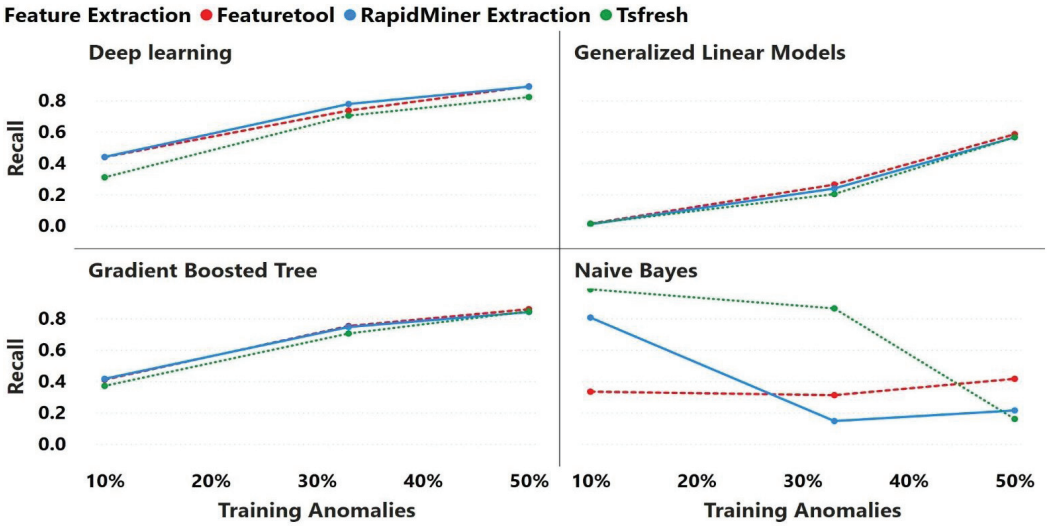

Figure 5: The Recall Score of Algorithms and Feature Extraction Methods at Each NTL Training Proportion Used.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

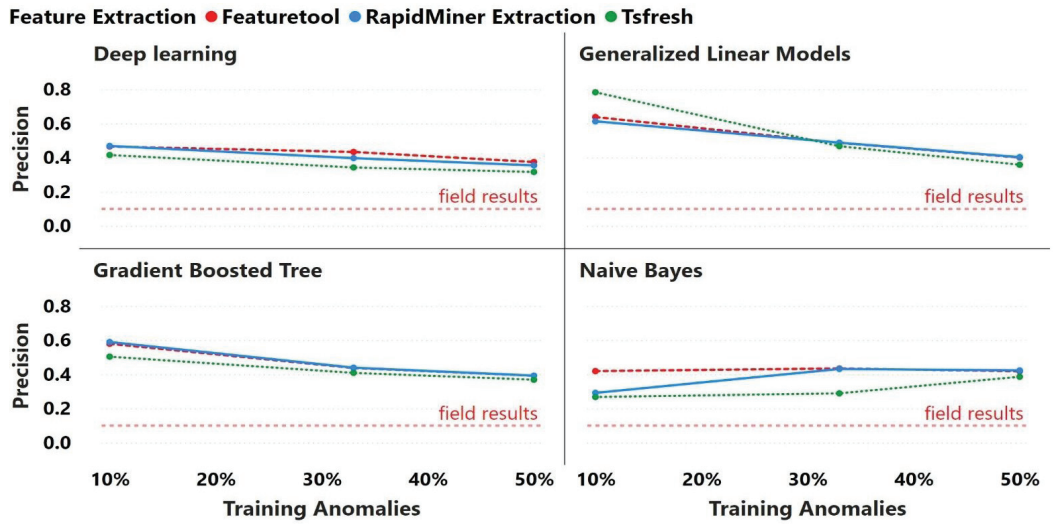

Figure 6: The Precision Score of Algorithms and Feature Extraction Methods at Each NTL Training Proportion Used.

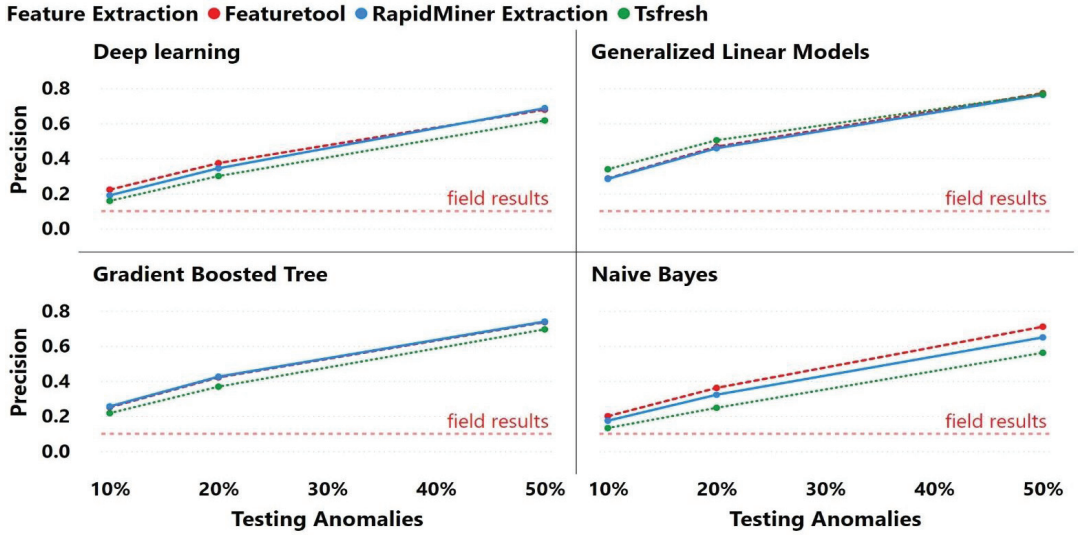

Figure 7: The Precision Score of Algorithms and Feature Extraction Methods at Each NTL Testing Proportion Used.

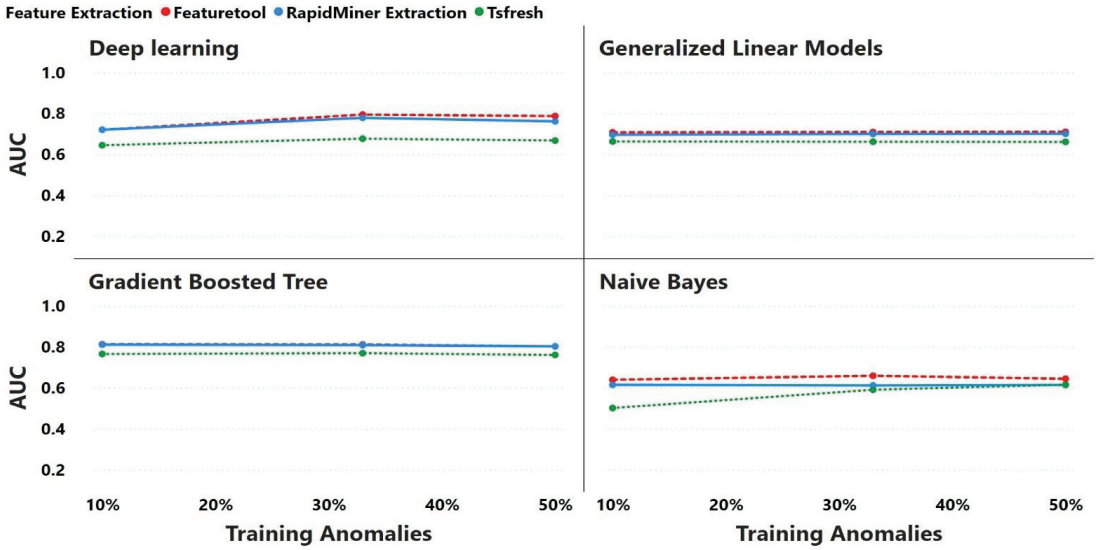

Figure 8: The AUC Score of Algorithms and Feature Extraction Methods at Each NTL Training Proportion Used.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

were the best-performing algorithms with GBT outperforming Deep Learning only slightly. GBT had higher True and False Positive scores, indicating higher Recall, this means in the real world it would trigger more investigation which could be useful for aggressive NTL reduction campaigns. Deep Learning on the other hand had higher True and False Negatives scores, indicating higher Specificity, which might be useful for a more targeted loss reduction strategy.

Discussion

RapidMiner's easy deployment of machine learning algorithms and automated feature extraction together are a powerful combination for AutoML. Featuretools was overall the best feature extraction library across all algorithms, and this may be attributed to Deep Feature Synthesis (DFS) and its ability to generate rich and meaningful features based on the relationship between the data sets. RapidMiner extraction also performed well but still fell behind Featuretools across all algorithms, except GBTs where its performance was about the same. GBT was the best-performing algorithm with most of the models that used the Featuretools and RapidMiner extraction, achieving an AUC > 0.8, which is considered excellent [43]. Some Deep Learning models also achieved this AUC of 0.8 or greater but still performed slightly worse than GBT on that metric. Deep Learning and GBT when combined with Featuretools and RapidMiner extraction produced good models, whereas other models did not perform as well.

Table 1 shows the best-performing algorithms at each of the three levels of sampling used for the testing and training data sets. The difference between recoveries and recall reflected how close the values are for each model with a minimum of -0.09, a maximum of 0.08, and an average of 0.006. Both values increase when the models are trained at higher NTL proportions increasing the number of false positives. Precision in inverse proportion decreases as training NTL proportions increase reflecting the trade-off between it and recall. AUC is generally the highest when trained at 33% NTL but the differences between models trained with the same feature extraction methods are very small. For Deep Learning, GLMs, and Naïve Bayes, on average the algorithms trained at 33% NTL had the highest AUC among all other models for the same algorithms with the second-best models being trained at 50% NTL. GBTs however saw sign drops in AUC as the training proportions increased with the best model being trained at 33% NTL and then 50%. GBT was the overall best algorithm when looking at AUC. For Deep Learning, GLMs, and GTBs, all models saw an increase in recall when trained at higher NTL proportions with models trained at 50% NTL scoring the best. Naïve Bayes was the only algorithm that experienced a decrease in recall when trained at higher NTL proportions with its best models being those trained at 10% NTL. It is worth noting that changing the testing proportions did not significantly influence the Recall as the training proportions did. Deep Learning had the best Recall and Recovery especially when trained at 50% NTL proportions, indicating that it might be better at detecting accounts with high

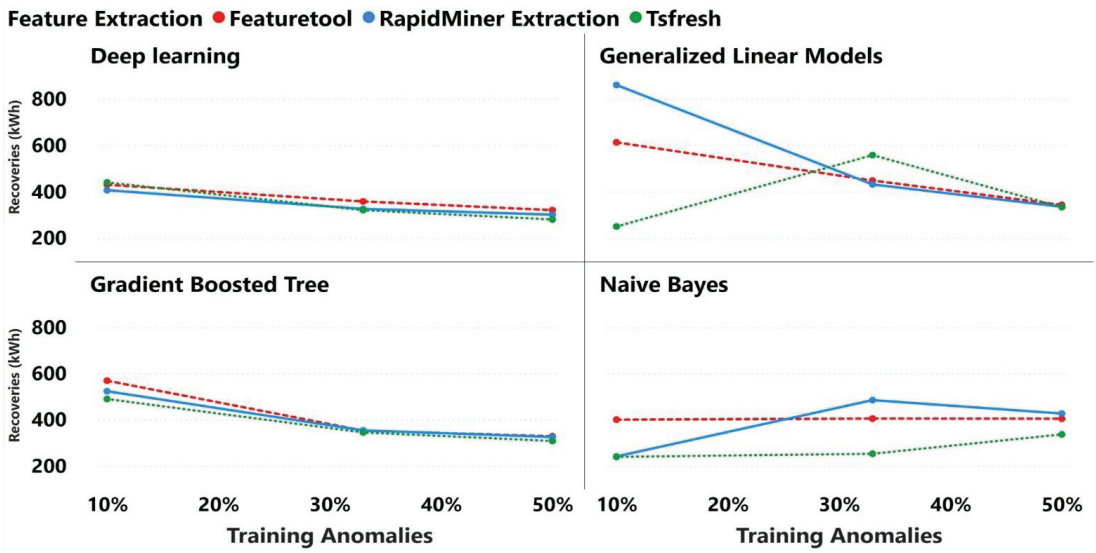

Figure 9: The Recoveries (kWh) per Audit of Algorithms and Feature Extraction Methods at Each NTL Training Proportion Used.

		GBT	
		Actual	
		Negative	Positive
Predicted	Negative	6,897	615
	Positive	2,987	1,856

		Deep Learning	
		Actual	
		Negative	Positive
Predicted	Negative	7,235	774
	Positive	2,649	1,697

Figure 10: The Confusion Matrix for the Top 2 Algorithms.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

Table 1: Comparison of the Algorithms Used to Detect NTL in This Experiment Trained and Tested at Different NTL Proportions.

Algorithm			Recall	AUC	Precision	Recoveries	Recovery Recall Difference
Deep Learning	33%	20%	0.78	0.79	0.39	0.75	0.03
Deep Learning	33%	50%	0.80	0.79	0.68	0.74	0.06
Deep Learning	50%	20%	0.89	0.79	0.31	0.87	0.02
Deep Learning	50%	50%	0.89	0.78	0.64	0.84	0.05
Generalized Linear Models	33%	20%	0.26	0.71	0.44	0.26	0
Generalized Linear Models	33%	50%	0.26	0.71	0.77	0.29	-0.03
Generalized Linear Models	50%	20%	0.57	0.71	0.35	0.58	-0.01
Generalized Linear Models	50%	50%	0.58	0.71	0.68	0.59	-0.01
Gradient Boosted Tree	33%	20%	0.75	0.81	0.39	0.70	0.05
Gradient Boosted Tree	33%	50%	0.75	0.81	0.71	0.67	0.08
Gradient Boosted Tree	50%	20%	0.86	0.81	0.33	0.82	0.04
Gradient Boosted Tree	50%	50%	0.86	0.80	0.66	0.81	0.05
Naive Bayes	33%	20%	0.34	0.64	0.35	0.30	0.04
Naive Bayes	33%	50%	0.26	0.70	0.75	0.29	-0.03
Naive Bayes	50%	20%	0.32	0.64	0.37	0.24	0.08
Naive Bayes	50%	50%	0.16	0.66	0.65	0.25	-0.09

NTL at the location. Looking at the Precision, all algorithms experienced an increase when tested at higher NTL proportion with models tested at 50% NTL having the greatest Precision. We observed that algorithms trained at 33% NTL or even as low as 10% and tested at 50% NTL had higher Precision but generally had the lowest Recall. A similar trend was observed for GLMs had the best Precision especially when trained at lower NTL associated with the worst Recall.

Multiple researchers have taken monthly consumption and other customer data, performed some processing, and then used it, combined with various algorithms, to train and test models. Table 1 shows the best-performing model for each algorithm based on their AUC, which we consider the most important metric for this problem due to the nature of unbalanced data sets. There were, however, models that scored higher in the other metrics but had lower AUC, such as GLMs trained at 10% that had Precision scores as high as 0.94 and Naïve Bayes models also trained at 10% with a Recall as high as 0.99. GBT when trained at 10% NTL and tested at 10% NTL had the best AUC. In most cases, the best-performing models were trained at the high NTL proportions except for the GBT model. Table 2 shows that when compared against the best-performing models of other research done using monthly consumption and some that used higher resolution daily consumption, our models performed better than the previous research in key metrics such as AUC, Recall, and Precision. It is worth noting that a Random Under-Sampling (RUS) Boosting algorithm tested on a data set with a 50% NTL was among the best we could find in the literature; we were however able to match its AUC and outperform its Recall and Precision scores with other models, though they did not match RUS Boosting algorithm's AUC. We consider these highly important metrics for utilities in their fight against NTL. Our model performance was also comparable to research done using higher-resolution daily readings, indicating a potential for significant improvement if the models are trained on similarly high-quality data.

Contribution

The paper advocated for the use of Automated Feature Extraction, model training, and testing as a part of a new paradigm of AutoML. This increases the effectiveness and scale of machine learning as it allows for the easy preparation of data, training of models, and testing across a wide range of parameters. Through this approach, we are able to quickly determine the effectiveness of models on a given problem through a process of rapid iterations using different configurations of training data, test data, and algorithms. This is ideal for a problem such as NTL detection where the ground truth can vary greatly over a geographic location and one-size-fits-all approaches may inadequate.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

Table 2: Comparison of the Algorithms Used to Detect NTL in this Experiment with the Research that Used a Similar Methodology

Paper	Algorithms	NTL Training Proportion	NTL Testing Proportion	Recall	AUC	Precision	Recoveries
Current	Deep Learning	33%	10%	0.78	0.81	0.24	0.63
Current	Generalized Linear Models	50%	10%	0.6	0.72	0.2	0.6
Current	Naive Bayes	33%	50%	0.26	0.7	0.75	0.29
[18]	Neural Network	50%	4%	0.55	0.65	0.08	-
[20]	Random Undersampling Boosting Algorithm	50%	50%	0.65	0.82	0.89	-
[21]	Random Forest	50%	11.6%*	0.649	0.646	0.19	
[22]	Extreme Gradient Boosting	11.8%*	11.8%*	0.48	0.76	0.38	-
[23]	Support Vector Machine	11.2%*	11.2%*	0.6	-	-	-
[24]	Support Vector Machine	60%	20%	0.75	0.55	-	-
[25]	Random Forest	40%	3%	-	0.63	-	-
[26]	Random Forest	70%	70%	-	0.75	-	-

(* Indicates Percentage Was Not Explicitly Stated but Assumed Based on # of Anomalies in the Data Set).

This paper also focused on Recovery, which was a key metric that other researchers have overlooked. For a problem such as NTL detection where not all energy theft is the same, looking at the impact of detection from the perspective of energy recovered is important for real-world applications. The detection of a larger amount of energy stolen with a smaller number of customers has a greater effect on NTL than a smaller amount of energy from a larger number of customers. Previous research which only focused on the number of anomalous accounts detected, missed these insights.

Overall, this paper presents a novel approach to the problem of NTL detection by largely automating part of the machine learning process such as the feature extraction, sampling, training, and evaluation of the models. It improves on the use of classical machine learning algorithms and outperformed other models trained on similar data. The inclusion of Recovery as a key metric also lends itself to improving the practical value of this approach to be focused on what is a real-world priority to utilities but is often not considered by researchers.

Conclusion

Though the recent increased adoption of smart grid technology has led to the use of more sophisticated approaches to NTL detection in developed countries, the data-driven approach still represents the most attractive model in the developing world. This is further aided by machine learning tools which allow for the easy training and deployment of models. Class imbalance can also be addressed with sampling which improves Detection Rate, though at the expense of Precision. Though issues of feature selection still affect the data-oriented approaches, this can be overcome using techniques such as Deep Feature Synthesis (DFS) and Time Series Feature Extraction, which generate rich feature sets from available data. These strategies, when applied in the real world, can result in a significant increase in efficiency, particularly in feature extraction using Featuretools and the Gradient Boosted Tree algorithm. This can further aid in segmenting the grid by the percentage of NTL and then training algorithms on these data set segments with NTL levels similar to where they will be deployed.

As more of the local grid is converted to use smart meters, additional data will also be available for training and testing. Further work involving more grid-related data on NTL at a given section of the grid, as well as sensor data from the newly installed metering points, can be considered. Further testing using Tsfresh might also be helpful as it is designed for high-interval sensor data.

Acknowledgment

The authors wish to thank the Jamaica Public Service Company Ltd for access to the data used in this research, and Andrew Manison for valuable editorial and administrative assistance.

References

1. Dick A. Theft of electricity - how UK electricity companies detect and deter. European Convention on Security and Detection. 1995.

2. Antmann P. Reducing technical and non-technical losses in the power sector. Background Paper for the WBG Energy Strategy. Energy Unit World Bank. 2009; 1-34.

3. Grant L, Latchman H. A Data-Oriented Approach to the Problem of Power Grid Non-Technical Losses in Developing Countries. In: IMCIC'21: The 12th International Multi-Conference on Complexity, Informatics and Cybernetics. Orlando, Florida, USA: Proceedings; 2021.

4. Messinis GM, Hatziargyriou ND. Review of non-technical loss detection methods. Electric Power Systems Research. 2018; 158: 250-266.

5. Nimbargi S, Mhaisne S, Nangare S, Sinha M. Review on ami technology for smart meter. In: 2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology ICAECCT. 2016.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

6. Saeed MS, Mustafa MW, Hamadneh NN, Alshammari NA, Sheikh UU, Jumani TA. Detection of Non-Technical Losses in Power Utilities—A Comprehensive Systematic Review. Energies. 2020; 13(18):4727.

7. Ponce-Jara MA, Ruiz E, Gil R, Sancristóbal E, Pérez-Molina C, Castro M. Smart Grid: Assessment of the past and present in developed and developing countries. Energy Strategy Reviews. 2017; 18:38-52.

8. Guerrero JI, Matos AP, Personal E, León C, Biscarri J, Biscarri F. Intelligent Information System as a Tool to Reach Unaproachable Goals for Inspectors: High-Performance Data Analysis for Reduction of Non-Technical Losses on Smart Grids. In: The Fifth International Conference on Intelligent Systems and Applications. 2016; 83-87.

9. Guerrero JI, Monedero I, Biscarri F, Biscarri J, Millan R, Leon C. Non-technical losses reduction by improving the inspections accuracy in a power utility. IEEE Transactions on Power Systems. 2018; 33(2):1209-1218.

10. Han SY, No J, Shin J, Joo Y. Conditional abnormality detection based on AMI data mining. IET Generation, Transmission & Distribution. 2016 Sep; 10(12):3010-3016.

11. Depuru SSSR, Wang L, Devabhaktuni V. Support vector machine-based data classification for detection of electricity theft. In: 2011 IEEE/PES Power Systems Conference and Exposition. 2011.

12. Júnior LAP, Ramos CCO, Rodrigues D, Pereira DR, Souza AND, Costa KAPD. Unsupervised Non-Technical Losses Identification through Optimum-Path Forest. Electric Power Systems Research. 2016; 140: 413-423.

13. Messinis GM, Hatziargyriou ND. Unsupervised Classification for Non-Technical Loss Detection. In: 2018 Power Systems Computation Conference (PSCC) 2018.

14. Messinis GM, Rigas AE, Hatziargyriou ND. A Hybrid Method for Non-Technical Loss Detection in Smart Distribution Grids. IEEE Transactions on Smart Grid. 2019; 10(6):6080-6091.

15. Buzau MM, Tejedor-Aguilera J, Cruz-Romero P, Gomez-Exposito A. Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. IEEE Transactions on Smart Grid. 2019; 10(3):2661-2670.

16. Qu Z, Li H, Wang Y, Zhang J, Abu-Siada A, Yao Y. Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. Energies. 2020; 13(8):2039.

17. Pereira J, Saraiva F. Convolutional neural network applied to detect electricity theft: A comparative study on unbalanced data handling techniques. International Journal of Electrical Power & Energy Systems. 2021; 131:107085.

18. Figueroa G, Chen YS, Avila N, Chu CC. Improved practices in machine learning algorithms for ntl detection with imbalanced data. In: 2017 IEEE Power & Energy Society General Meeting. 2017.

19. Hussain S, Mustafa MohdW, Jumani TA, Baloch SK, Alotaibi H, Khan I. A novel feature engineered-CatBoost- based supervised machine learning framework for electricity theft detection. Energy Reports. 2021; 7:4425-4436.

20. Avila NF, Figueroa G, Chu CC. NTL detection in electric distribution systems using the Maximal Overlap Discrete wavelet-packet transform and Random Undersampling Boosting. IEEE Transactions on Power Systems. 2018; 33(6):7171-7180.

21. Massaferro P, Marichal H, Martino MD, Santomauro F, Kosut JP, Fernandez A. Improving electricity non-technical losses detection including neighborhood information. In: 2018 IEEE Power & Energy Society General Meeting (PESGM). 2018.

22. Massaferro P, Martino JMD, Fernandez A. NTL detection: Overview of classic AND DNN-based approaches on a labeled dataset of 311k customers. In: 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). 2021.

23. Nagi J, Yap KS, Tiong SK, Ahmed SK, Mohammad AM. Detection of abnormalities and electricity theft using genetic Support Vector Machines. In: TENCON 2008 - 2008 IEEE Region 10 Conference. 2008.

24. Glauner P, Boechat A, Dolberg L, State R, Bettinger F, Rangoni Y. Large-scale detection of non-technical losses in imbalanced data sets. In: 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). 2016.

25. Glauner P, Meira JA, Dolberg L, State R, Bettinger F, Rangoni Y. 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. 2016.

26. Meira JA, Glauner P, State R, Valtchev P, Dolberg L, Bettinger F. Distilling provider-independent data for general detection of non-technical losses. In: 2017 IEEE Power and Energy Conference at Illinois (PECI). 2017.

27. Ghori KM, Awais M, Khattak AS, Imran M, Fazal-E-Amin, Szathmary L. Treating Class Imbalance in Non-Technical Loss Detection: An Exploratory Analysis of a Real Dataset. IEEE Access. 2021; 9:98928-98938.

28. Lee J, Sun YG, Sim I, Kim SH, Kim DI, Kim JY. Non-Technical Loss Detection Using Deep Reinforcement Learning for Feature Cost Efficiency and Imbalanced Dataset. IEEE Access. 2022; 10:27084-27095.

29. González Rodríguez R, Jiménez Mares J, Quintero M. CG. Computational Intelligent Approaches for Non-Technical Losses Management of Electricity. Energies. 2020 May 11; 13(9):2393.

30. Coma-Puig B, Carmona J. A Human-in-the-Loop Approach Based on Explainability to Improve NTL Detection. In: 2021 International Conference on Data Mining Workshops (ICDMW). Auckland, New Zealand: IEEE; 943-950. https://ieeexplore.ieee.org/document/9679878/

31. Coma-Puig B, Carmona J. Non-technical losses detection in energy consumption focusing on energy recovery and explainability. Mach Learn. 2022; 111(2):487–517.

32. Han SY, No J, Shin J, Joo Y. Conditional Abnormality Detection Based on Ami Data Mining. IET Generation, Transmission & Distribution. 2016; 10(12):3010-6.

33. Saeed MS, Mustafa MW, Hamadneh NN, Alshammari NA, Sheikh UU, Jumani TA. Detection of Non-Technical Losses in Power Utilities-A Comprehensive Systematic Review. Energies. 2020; 13(18):4727.

34. Glauner P, Meira JA, Valtchev P, State R, Bettinger F. The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey. International Journal of Computational Intelligence Systems. 2017; 10(1):760.

35. Kanter JM, Veeramachaneni K. Deep feature synthesis: Towards automating data science endeavors. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2015.

36. Al-Turaiki I, Altwaijry N. A Convolutional Neural Network for Improved Anomaly-Based Network Intrusion Detection. Big Data. 2021 Jun;9(3):233-252. doi: 10.1089/big.2020.0263. PMID: 34138657; PMCID: PMC8233218.

37. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package. Neuro Computing. 2018; 307:72-77.

38. Xu J. Research on Power Load Forecasting based on Machine Learning. In: 2020 7th International Forum on Electrical Engineering and Automation (IFEEA). 2020; 562-567.

39. Khan S, Bhardwaj S. Time series forecasting of Gold Prices. In: Advances in Intelligent Systems and Computing. 2018; 63-71.

40. Ilieva R, Angelov M. Template for Building Manageable Data Mining Autonomous Process with RapidMiner Studio. In: 2021 XXX International Scientific Conference Electronics (ET). Sozopol, Bulgaria: IEEE; 2021; 1-5. https://ieeexplore.ieee.org/document/9580103/

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067

41. Dossin E, Martin E, Diana P, Castellon A, Monge A, Pospisil P, Bentley M, Guy PA. Prediction Models of Retention Indices for Increased Confidence in Structural Elucidation during Complex Matrix Analysis: Application to Gas Chromatography Coupled with High-Resolution Mass Spectrometry. Anal Chem. 2016 Aug 2;88(15):7539-47. doi: 10.1021/acs.analchem.6b00868. Epub 2016 Jul 22. PMID: 27403731.

42. Marzukhi S, Awang N, Alsagoff SN, Mohamed H. RapidMiner and Machine Learning Techniques for Classifying Aircraft Data. Journal of Physics: Conference Series. 2021; 1997(1):012012.

43. Hosmer DW, Lemeshow S, Sturdivant RX. Assessing the fit of the model. Applied Logistic Regression. 2013; 153-225.

44. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436-44. doi: 10.1038/nature14539. PMID: 26017442.

45. Nelder JA, Wedderburn RW. Generalized linear models. 1992; 547-563. (Springer Series in Statistics).

46. Joshi AV. Deep Learning. Machine Learning and Artificial Intelligence. 2019; 117-126.

47. Arunadevi J, Ramya S, Raja MR. A study of classification algorithms using Rapidminer. International Journal of Pure and Applied Mathematics. 2018; 119(12):15977-15988.

48. Truong VH, Vu QV, Thai HT, Ha MH. A robust method for safety evaluation of steel trusses using Gradient Tree Boosting algorithm. Advances in Engineering Software. 2020; 147:102825.

49. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis. 2002; 38(4):367-378.

50. Frank E, Trigg L, Holmes G, Witten IH. Machine Learning. 2000; 41(1):5-25.

51. Zhang Y, Ma Y. Application of supervised machine learning algorithms in the classification of sagittal gait patterns of cerebral palsy children with spastic diplegia. Comput Biol Med. 2019 Mar;106:33-39. doi: 10.1016/j.compbiomed.2019.01.009. Epub 2019 Jan 16. PMID: 30665140.

Citation: Grant L, Latchman H, Alli K (2023) Detection of electricity theft in developing countries-A machine learning approach. Trends Comput Sci Inf Technol 8(2): 038-049. DOI: https://dx.doi.org/10.17352/tcsit.000067