Technoscience Journal for Community Development in Africa 2:1 (2021) 81–91
ISSN 2659-0573

Research Article

Fraud detection in customers' electricity consumption in Nigeria using machine learning approach

Sulaiman Olaniyi Abdulsalam[*,1], Micheal Olaolu Arowolo[2], Ronke Babatunde[1], Musa Isiaka[1], Shakirat Oluwatosin Haroon Sulyman[3]

[1] Department of Computer Science, Kwara State University, Malete, Nigeria
[2] Department of Computer Science, Landmark University, Omu-Aran, Nigeria
[5] Department of Information and Communication Science, University of Ilorin, Ilorin, Nigeria

Abstract: Electricity theft is estimated to have cost Nigeria billions of Naira over the years. Electric utilities use data analytics to discover unusual consumption patterns and possible fraud in order to prevent electricity theft. This work uses data analysis to detect electricity theft, as well as a measure that uses this threat model to compare and evaluate anomaly detectors. This study employs machine learning algorithms to categorize fraud detection in customers' electricity use, as data mining techniques has helped multiple companies and sectors better their various types of technology. Support Vector Machine (SVM) and C4.5 Decision Tree classification algorithms were used to detect fraud using consumer electricity use data. The accuracy of SVM and C4.5 was 63.4 percent and 65.9%, respectively. As a result, the Map-Reduced-ANOVA with SVM attained an accuracy of 77.5 %.

Keywords: Fraud detection; Machine learning; classification; support vector machine; decision tree

1. Introduction

An electrical company's fundamental function is to provide customers with high-quality electrical energy in a secure environment in return with an economic outlook. Electricity companies are facing financial and technical challenges in Electrical power system planning, control, and operation. For optimal electric power planning and operation Systems, the current and future electricity charges need proper assessments (Hambali et al., 2017).

Nigeria has joined the rest of the world in the past few years in privatizing the economic sector. Therefore, Electricity generation and distribution companies have not been excluded from privatization and deregulate. These have brought more attention to the issue of precise electric load prediction in the regional and the national systems of power. There was a need for suitable solutions to satisfy consumers. Assessment of current and future electricity charges by electricity companies, to have Optimum scheduling and proper electric power system configuration (Hambali et al., 2017). Electricity theft is a severe utility issue, with dishonesty from users. It prompts for the financial identification of clients with anomalies (Jeyaranjani & Devaraj, 2018).

Fraud describes cases in which a threat actor gains unauthorized access to use electricity, the importance of machine learning and data science cannot be overemphasized, predicting and detecting frauds correctly on a given financial transaction dataset is of the essence for evaluating effectiveness vis-à-vis frauds detection in electricity consumption. Previous research on the detection of power theft has carried out a series of detection procedures. The specific is where the supplied power and billed power differ. All customers belonging to the area are deemed, suspect. The drawback of the previously discussed work is that the identification of power theft was carried out based on the assumption

* Corresponding author:
Email: sulaiman.abdulsalam@kwasu.edu.ng

KWASU PRESS AND PUBLISHING

that the clients are suspected of being a fraud. This case could identify the potential client as fraudulent clients. This study is motivated to incorporate the bogus data on customer power consumption into essential information on energy consumption. The algorithms of machine learning are used to analyze the information that clusters and then classify the client. Data of customers are discriminated against as genuine and fraud based on their pattern of use (Glauner, 2017).

Several methodologies based on data-mining have been developed to identify anomalies in energy consumption. Clustering techniques have been successfully applied as an algorithm with applicable reduction data-space requirements, a useful tool for detecting natural consumer groups with the same energy-consumption profiles. It turns out that clustering is the most appropriate technique for modelling and identifying the different energy consumption profiles over an electrical utility, in particular the blurry clustering (Angelos et al., 2011). There are recurrent variations in the behavior of produced data with huge volume of data. precise identification of information of interest which are responsible for causing fraud are imperative. Several existing schemes have employed filtrations such as dimensionality reduction models.

In this study, the machine learning approach for fraud detection in electricity consumption is proposed. Energy consumption data for customers are to be classified using SVM and Decision tree by training the data to build a model that forms the basis for the decision-making classification of normal and abnormal behaviours. Also, a MapReduce-ANOVA is proposed to select relevant features and classified using the SVM and Decision tree classifiers. A comparative analysis is carried out on the models from the obtained results. The technique also suggests punishment for detected fraudulent customers to ease monitoring for utility officials.

2. Related works

In the study of electricity theft detection, the advent of smart meters with its continued provision of real-time customer consumption data has necessitated the applications of big data analytics and machine learning. Machine learning applications, including Support Vector Machines (SVM) and Artificial Neural Networks (ANN), are often used. Training and modelling datasets have reported some works using several classification strategies (Jokar et al., 2016).

An electricity theft detection framework based on a universal prediction algorithm was suggested using a Universal Anomaly Detection (UAD) framework that uses the Lempel-Ziv universal compression algorithm to achieve real-time detection in a smart grid setting. Several network parameters can be monitored through the anomaly detections. Still, this framework monitors data on energy consumption, rate of change in data on energy consumption, date stamp, and time signatures. Usual and abnormal classification of data behaviour, the Lempel-Ziv algorithm is used to assign occurrence probability to the compressed data of the parameters being monitored. This framework can learn the typical behaviours of smart meter data and give alerts based on deviation from this probability during any detected anomaly. Also suggested within the framework is a forced aggressive measure as a means of applying fines to fraudulent customers (Otuoze et al., 2019).

Electric Power Load Forecast Using Decision Tree Algorithms, an up-to-date experiment was proposed using Decision Tree Algorithms Classification and Regression Tree CART, (Reduced Error Pruning Tree) REPTree and Decision Stump for electric load forecasting. The work revealed that REPTree Decision Tree Technique is suitable for forecasting electrical load and has outperformed other algorithms for the decision tree. This work will be of considerable use to Yola / Jimeta Power Transmission Company and others involved in the power transmission, generation, distribution, and marketing industries to enable them to forecast electrical power charges and to provide timely advice and decisions (Hambali et al., 2017).

The use and comparison of two ANN algorithms (MLP and RBF) and SMO algorithms have been proposed for the Artificial Neural Network approach for electrical load forecasting in power distribution companies. The results were interpreted then; the models obtained were analyzed to determine the pattern in the model of load forecasting. The experimental analysis was conducted using the option of 10-fold cross-validation tests. Results showed that the Multilayer-Perceptron (MLP) model gives 86 percent accuracy with Mean Absolute error (MAE) of 0.016, Radial Base Function (RBF) was 76 percent accurate with MAE of 0.030, and Sequential Minimal Optimization (SMO) 85 percent accuracy with MAE of 0.090 indicating an adequate level of electrical load forecast (Hambali et al., 2017).

Approach to identify machine learning algorithms was proposed to Suspect customers using the usage pattern for customer power. To this end, the Machine Learning Algorithm is used. Customer trustworthiness is verified and is chosen for the theft program. This analysis is done by tweaking the actual data from the Smart Meter to create fraudulent data. The ANN classification model is developed using a supervised

learning algorithm which helps discriminate against client's profile based on their genuine activity and their fraudulent use of electricity. The simulation result shows that the proposed system is useful in the highly accurate identification of the suspects. Absolute error (MAE) of 0.016, Radial Base Function (RBF) was 76 percent accurate with MAE of 0.030, and Sequential Minimal Optimization (SMO) 85 percent accuracy with MAE of 0.090 indicating an adequate level of electrical load prediction (Jeyaranjani & Devaraj, 2018).

The detection of electricity theft using external detection algorithms in customer consumption was analyzed using the feasibility of application Outlier's detection algorithms to improve the safety of AMI by detecting the theft of electricity. We are exploring the performance of various existing outer detection algorithms on a real dataset (use of energy from consumers). The results show the feasibility of using outliers' algorithms in AMI security and also the effectiveness of using these methods for theft detection in the electricity consumption datasets (Yeckle & Tang, 2018).

A computational technique for the classification of electricity consumption profiles was proposed to detect and identify abnormalities in customer consumption in power distribution systems. The methodology consists of 2 steps. In the first, a Fuzzy clustering based on C-means is performed to find consumers with similar consumer profiles. A fuzzy classification is then carried out using a fuzzy membership matrix and the Euclidean distance to the cluster centres. The distance measures are then normalized and ordered, resulting in a unitary index score, where potential fraudsters or users with irregular consumption patterns have the highest score. The approach has been tested and validated on a real database, showing good performance in fraud detection and measurement defect tasks (Angelos et al., 2011).

A Convolutional Neural Network (CNN) is programmed to learn the features from large and changing smart meter data by convolution and down-sampling operations between various hours of the day and different days. Besides, a dropout layer is introduced to delay the possibility of overfitting, and in the testing process, the back-propagation algorithm is applied to change network parameters. And then, depending on the features collected, the random forest (RF) is conditioned to determine when the user steals electricity. The grid search algorithm is adopted to decide optimum parameters to construct the RF into the hybrid model. Finally, tests are performed based on actual evidence on energy usage, and the findings demonstrate that the proposed model of detection outperforms other models in terms of precision and performance (Li et al., 2019).

The attribute ranking impact on the detection of credit card fraud was suggested by the use of credit card fraud datasets (Taiwan and European bank) collected from UCI and ULB repositories containing 30 000 and 284 807 transactions respectively. The rating of features on dataset sets is conducted using the method of correlation analysis. Four classifier algorithms are developed and implemented on data graded by raw and function. The classification algorithms are implemented in MATLAB. Specificity, specificity, Matthew's correlation coefficient, resilience, accuracy, and balanced classification rate are the output metrics used in determining the impact of the four classifiers on the raw and feature rated datasets. Results from the comparative study show that classifiers for decision tree variants outperform naïve Bayes, enable radial function approaches for the vector and neural network, respectively. The graded function and raw data sets of the data from the European credit card fraud reported the highest output metrics for decision trees. The paper explores the impact of rating features of two imbalanced credit card fraud data using a filter method on four machine learning techniques (Awoyemi et al., 2018).

Classification models based on decision trees and vector support machines (SVM) are developed and applied to problem detection of credit card frauds. This study is one of the first to compare SVM performance and decision tree methods in the detection of credit card fraud to a real dataset (Sahin & Duman, 2011).

A new method for extracting features and model fusion technologies to solve the problem of identifying basic medical insurance fraud is introduced. The second-level extraction algorithm function proposed in this paper will effectively extract essential features and improve the accuracy of subsequent algorithms in forecasting. A sample division approach based on the idea of sample proportion equilibrium is proposed to solve the problem of unbalanced simulation distribution in the medical insurance fraud finding scenario. A new training and fitting model fusion algorithm (tree hybrid bagging, Bagging) is proposed based on the above methods of extraction and sample division of the function. This method makes fair use of the balanced tree model algorithm based on Boosting to fuse and eventually achieves the effect of improving the accuracy of detection of simple medical insurance frauds (Gong et al., 2020).

A supervised technique to detect meter irregularities and consumer fraudulent activity (meter tampering) is proposed, which primarily but not only feeds on meter information. Their system detects anomalous meter readings based on models developed on past data using

machine learning techniques. Unlike other previous studies, the results of field checks can be implemented incrementally to expand the inventory of fraud and non-fraud trends, thereby increasing model consistency over time and theoretically adjusting to evolving fraud patterns. The entire device was built for a company supplying power and gas and has since been used to conduct many field checks, for substantial gains in fraud identification relative to previous reviews using simplified techniques (Coma-Puig & Carmona, 2018).

A device to track power theft is proposed using a combination of a convolution neural network (CNN) and a long-term memory model (LSTM). CNN is a method commonly used to automate the detection of characteristics and the labelling process. Since the power consumption signature is time-series data, we were led to create a smart grid data classification model based on the CNN LSTM (CNN-LSTM). A novel data pre-processing algorithm has also been applied in this work to measure missing instances in the dataset, based on local values relative to the missed data point. In comparison, the number of users of electricity theft was comparatively low in this dataset which may have made the model unreliable in detecting users of theft. This class disequilibrium scenario was resolved through the generation of synthetic data. Finally, the findings obtained indicate that the proposed scheme may with good accuracy identify both the majority class (regular users) and the minority class (users of electricity theft) (Hasan et al., 2019).

A short-term method of detecting fraud based on the Support Vector Machine (SF-SVM) is introduced, the system only needs to collect and archive a limited amount of the user's recent electricity usage data to identify troublesome users. Having a small volume of data will minimize data storage and reduce the costs of transferring data remotely. Also, protection should be best protected for consumers. The system periodically gathers grid and consumer data on energy usage over a specified period. When the device senses that a threshold increases the difference between the amount of energy generated by the regional grid and the amount of electricity used by consumers, the mechanism switches to a suspect state and causes fraud detection. The framework implements deep learning algorithms to retrieve user data attributes, and ultimately identify suspect users. The results of the simulation demonstrate that anomalous users are successfully observed by the system (Xiong et al., 2019).

3. Methodology

The proposed methodology used a small percentage of the Nigerian Electricity Power distribution of low-voltage consumption customers residing in Kwara State, Nigeria, who use the monthly kWh interval data collected from the Nigerian electricity database for more than 300 customers. Decision tree algorithms and SVM are used for the Information Discovery and Data Mining process. The proposed system uses the Ibadan Electricity Distribution Company, Ilorin Branch, IBEDC Baboko office Nigeria's customer data sets for electricity distribution. There are 300 instances and eight attributes in the Customer dataset, table 1 and figure 1 respectively shows the dataset features.

Pre-processing: This step includes the preparation of datasets before the execution of data mining techniques. Currently, the typical methods of pre-processing, such as data cleaning, variables translation, and data partitioning, were applied. Additional procedures such as the collection of attributes and the re-balancing of data were also employed to solve high dimensionality issues and imbalanced data which could be present in the dataset.

Data Mining: Data mining algorithms are implemented using C 4.5 and are used and contrasted with SVM. The obtained models are analyzed to determine the performance of the experiment

The insignificant and irrelevant features degrade the fetching of the qualities of the analysis. It is essential to analyze the dataset. The input data are preprocessed using MapReduce-ANOVA. The selected features are classified using SVM and Decision Tree.

3.1. The experiments

In this study, MATLAB was used as an experimental tool, MATLAB is an ease use of the device for implementing Machine learning approaches, it is a data analytics technique that teaches computers to do what comes naturally with the capability of classification among others (Paluszek & Thomas, 2017), figure 2 shows the framework approach for the study.

3.1.1. MapReduce-ANOVA

MapReduce is a programming model, for processing datasets in a dispersed method consisting of mapping, shuffling and reducing steps. The algorithm input is a matrix, with total feature set Y numbers and dataset sample numbers X. ANOVA is an analysis of variance used in computing multiple means of the dataset values and visualizing the significant differences in the multiple sample means (Kumar et al., 2016)], algorithm 1 shows the MapReduce based ANOVA procedure.

Table 1: Dataset Attribute Descriptions

ATTRIBUTES DESCRIPTORS		
Attributes	Data type	Description
Customer's Name	Nominal	Customer's Name of account
Active Terrain	Nominal	Consumption Purpose
Meter Number	Numeric	Meter Reading
Balance	Numeric	Customer's Account balance
Supply Reference	Numeric	Customer's Supply Number
Consumption	Numeric	Customer Consumption Rate
Monthly Charges	Numeric	Customer Charges
Status	Nominal	Customer Account Status

POWER HOLDING COMPANY OF NIGERIA PLC
BABOKO DISTRICT ILORIN
SAMPLE BILLING & COLLECTION DATA (RESPONSE RATE: AVERAGE 60-70%)

S/N	NAME	ADDRESS	ACCOUNT NUMBER	CURRENT BILLING(N)	CURRENT BILLING(N)	CURRENT BILLING(N)	CURRENT BILLING(N)	CURRENT BILLING(N)	BILLING AMOUNT (N)	BILLI AMOUN
	SURULERE SERVICE CENTRE			JANUARY	FEBRUARY	MARCH	APRIL	MAY JUNE	JULY	AUGU
1	Intercontinental Bank	Taiwo Rd. Ilorin	97/21/91/1118-01	37,821.32	19,080.92	38,248.16	35,047.69	26,655.91	23,816.28	27,9
2	Stadium Complex	Taiwo Rd. Ilorin	97/21/91/1127-01	93,458.80	38,887.20	87,377.69	46,655.42	44,857.56	36,470.42	53,6
3	Kwara Ind. Ltd.	Oko-Erin	97/21/91/1145-01	115,031.56	95,946.19	107,352.91	159,009.12	80,676.39	83,037.34	145,8
4	Kwara State Govt.	Stadium	97/21/91/1154-01	95,026.37	71,020.32	88,161.68	81,918.27	75,674.87	69,431.46	64,9
5	Lubcon Ltd.	Taiwo Rd. Ilorin	97/21/91/1305-01	11,021.85	11,021.85	11,021.85	11,021.85	11,021.85	11,021.85	19,2
6	Access Bank Plc	Stadium Complex	97/21/91/1314-01	81,208.49	38,717.99	118,803.98	66,165.91	63,740.55	58,790.64	74,3
7	The Medical Officer Unilorin	Gen. Hosp. Ilorin	97/21/91/0004-01	349,903.26	283,328.96	223,637.40	325,335.47	311,830.79	209,312.71	205,0
8	Faculty of H'Science	Lagos Road	97/21/91/0068-01	21,530.88	17,154.90	22,559.67	28,573.02	31,644.90	28,347.10	25,9
9	The Med officer Uni.Tech.(UITH)	Lagos Road	97/21/91/0077-01	262,970.51	304,049.66	394,005.15	440,081.78	418,513.41	280,328.20	278,3
10	Bank of the North	Baboko Mkt. Rd.	97/21/91/0120-01	51,320.41	36,195.69	35,757.07	45,245.52	39,418.38	33,311.35	39,4
11	Union Bank	Ilorin	97/21/91/0362-01	58,712.48	46,785.84	43,440.53	61,949.98	52,225.56	47,344.03	54,5
12	Univ. Teaching Hosp.(UITH)	Ilorin	97/21/91/0380-01	253,291.19	173,559.96	230,498.42	209,777.72	189,064.26	168,450.91	156,1
13	The Prin. Sch. Of Nurs.	Ilorin	97/21/91/0406-01	210,009.56	144,014.85	191,150.80	174,009.15	188,984.57	168,378.46	157,9
14	Inland Bank Plc.	New Mkt. Baboko	97/21/91/0479-01	54,904.21	42,102.99	41,189.25	48,077.61	47,677.46	37,667.88	47,4
15	M.T.N.	Agbo-Oba	97/21/91/0522-01	-	-	11,021.85	11,021.85	18,663.33	-	-
16	First Bank Plc.	Abdul Azeez	97/21/91/0531-01	48,939.11	40,576.56	40,637.67	51,970.93	49,789.36	41,877.60	50,5
17	Kwara Poly	Ilorin	97/21/91/1412-01	71,419.95	158,070.15	322,169.40	185,311.35	200,163.60	172,377.70	166,0
18	Univ. of Ilorin	Lagos Road	97/21/91/1421-01	13,459.95	32,876.55	664,785.45	476,665.80	315,648.90	301,331.95	157,1
19	Kwara Coll. Of Educ.	Lagos Road	97/21/91/1430-01	228,708.90	191,252.25	127,351.35	315,431.55	277,974.90	184,331.95	242,8

Figure 1. Customer Consumption Data from IBEDC

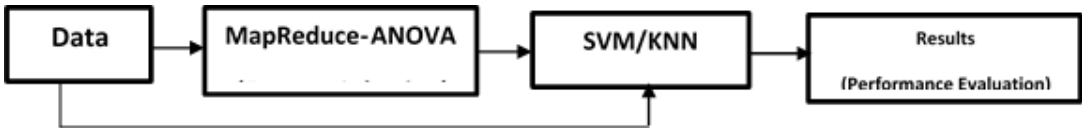

Figure 2: A framework of the proposed approach

3.1.2. C4.5 Algorithm

C4.5 is a decision tree algorithm, as an inheritor of ID3. Machine learning approaches allow decision trees to be the most accessible. Because of this, they are a way to identify phenomena that looks like a series of if-then statements arranged in a tree. Classification starts based on maximum node outputs named root node down the tree to the leaves. The target groups of our dataset are leaves. By answering issues down the decision tree, it is easy to implement classification. What the algorithm does is usually to partition the data into two or more sets. The information and entropy advantages, filtering is done based on the most relevant attributes to make as possible distinct classes (Khine & Khin, 2020), algorithm 2 shows the C4.5 procedure.

3.1.3. SVM algorithm

Support Vector Machines (SVM) are supervised learning algorithms that have proven to be better than specific other learning algorithms with attendants. SVM is a series of algorithms proposed to solve regression and classification problems. SVM has found use by differentiating different classes by way of a hyperplane to solve quadratic programming problems that have inequality restrictions and linear equality. The line takes full advantage of that. Although the SVM may not be as fast as other classification methods, owing to its ability to model multidimensional borderlines that are not sequential or straightforward, the algorithm derives its power from its high precision. SVM is not readily vulnerable to a scenario where a model is overly

complicated, such as having multiple parameters relative to observation amounts. These attributes make SVM the perfect algorithm for use in the fields of automated handwriting recognition, text categorization, speech recognition, etc. (Dada et al., 2019), algorithm 3 shows the SVM procedure.

3.2. Confusion matrix

The confusion matrix summarizes Algorithm efficiency. From this can be grasped the meaning of what algorithm is doing right and what is doing wrong. Confusion matrix rows are expected class, while rows are true class (Patil et al., 2018). The predicted class includes True Positive (TP), True Negative (TN), False

Algorithm 1: MapReduce-based ANOVA (Kumar et al., 2016)

Input: Y x X Matrix, X = number of features; Y= number of samples.
Output: Top P features.
1: Begin MR Job
2: MAP (M ()):
3: for each feature fi do
4: Calculate the value of BMS $i = 1; 2; \ldots; Y$
5: Calculate WMS values.
6: Calculate F-value $(Fi = BMS = WMS)$
7: Calculate p-value (pi) corresponding to each F-value using the F-distribution curve.
8: Emit $\{I, Fi; pi\}$
9: end for
10: REDUCE (R()):
11: for each feature fi do
12: if pi < 0:05 then
13: Select the feature, called fsi.
14: else
15: Discard the feature.
16: end if
17: end for
18: Emit $\{fs1, fs2, fs3, \ldots\}$
19: End

Algorithm 2: C4.5 Decision Tree (Jindal et al., 2016)

Input: Features
Output: Decision tree (DT)
1: Collect the data
2: Extract the features
3: **while** $(N_{at} \geq 2)$ **do** ▷ N_{at} = Number of features
4: Calculate total entropy (E) using eq. 2
5: **for** $(i = 1; i \leq N_{at}, i++)$ **do**
6: Calculate E_i using eq. (3)
7: Calculate IG_i using eq. (4)
8: Initialize $Max = 0$
9: **if** $(IG_i \geq Max)$ **then**
10: Max = IG_i
11: k = i
12: **end if**
13: **end for**
14: Make k as parent node
15: Split the attributes on decision attribute k
16: $N_{at} = N_{at} - 1$
17: **end while**

Positive (FP), and False Negative (FN), (Arowolo et al., 2020).

- The accuracy represents the TP+TN/ TP+TN+FP+FN
- The precision represents ratio of true positives (TP) and actual positives (TP + FP) = TP/ TP+FN.
- The recall or true positive rate (TPR) is ratio of true positives (TP) and actual positives (TP + FN).

4. Results and discussion

Decision tree C 4.5 and SVM classification models were proposed in this study for fraud detection for the distribution of electricity and performance evaluation of the models using both 10-fold cross-validation method-based classification accuracy and performance metrics.

Decision tree C 4.5 is known for its robust, accurate, and efficient method in classification model, because of its ability to split the dataset into groups for easy prediction (Damanik et al., 2019) which has been helpful and yielded a good result in this study. SVM, on the other hand, proves a high-performance of sensitivity which can still be considered as efficient.

Classification is one of the essential areas in machine learning, due to the huge variety of problems that can be stated as different or specific classification tasks. This study shows how there is a large number of applications and issues in various aspects of classification systems, which can be solved successfully with classification algorithms. The improvement of techniques for classification is, without a doubt, the most critical research line in the area, which will produce promising results in applications in the nearest future. In this sense, some problems that are currently tackled as classification problems, exploiting the continuity of the data, could be tackled in the future as exceptional cases of classification. For the future intelligent electrical network will be stated as classification problems and attacked with some of the algorithms described in this review (or improvements of them). Figure 5 shows the result output for the classification of power consumption in Nigeria using SVM. SVM achieved 63.4% overall accuracy. 86 features were selected using the MapReduced-based ANOVA. Figure 3, 4, 5, and 6, shows the confusion matrix for the performance evaluation.

The decision tree classification algorithm uses Simple Tree, the result is shown in figure 6 below, and its product outperformed the SVM. The decision tree has proven to be a better method in achieving the aim of this study.

In this study, classification of the power consumption, in Baboko locality of Kwara State Nigeria, was classified using MapReduce-ANOVA

Algorithm 3: SVM (Dada et al., 2019)

Input Sample Email Message x to classify
A training set *S*, a kernel function, $\{c_1, c_2, \ldots c_{num}\}$ and $\{\gamma_1, \gamma_2, \ldots \gamma_{num}\}$.
Number of nearest neighbours k.
for $i = 1$ to *num*
set $C=C_i$;
for $j = 1$ to q
set $\gamma=\gamma$;
produce a trained SVM classifier $f(x)$ through the current merger parameter (C, γ);
if ($f(x)$ is the first produced discriminant function) **then**
keep $f(x)$ as the most ideal SVM classifier $f^*(x)$;
else
compare classifier $f(x)$ and the current best SVM classifier $f^*(x)$ using *k*-fold cross-validation
keep classifier with a better accuracy.
end if
end for
end for
return Final Email Message Classification (Spam/Non-spam email)
end

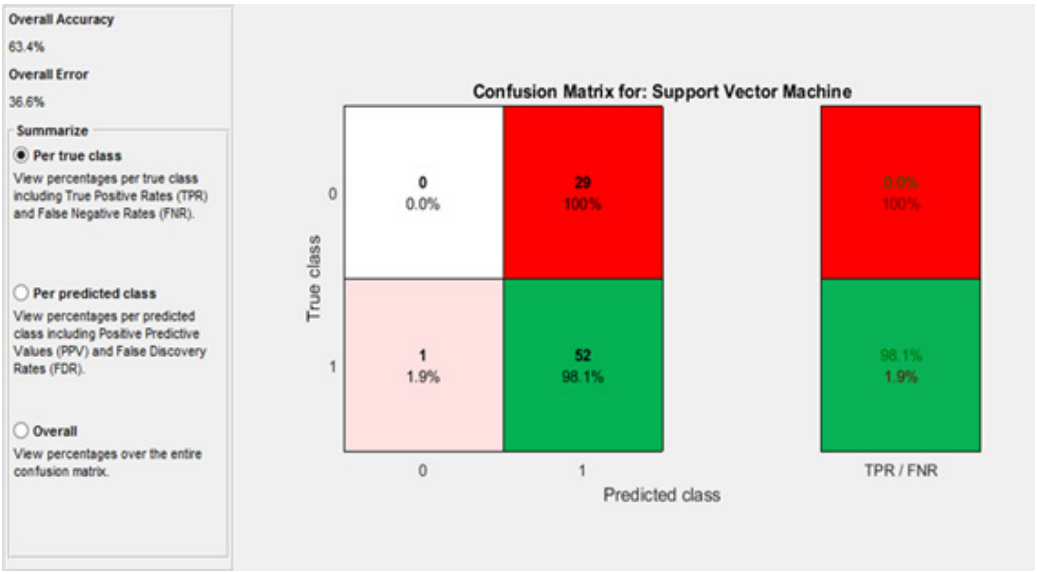

Figure 3: Confusion Matrix for the classification of power consumption using SVM
TP= 52; TN=1; FP=29; FN=0
Accuracy rate = 0.6341; Sensitivity Rate = 0.9811; Specificity Rate =34.5 Precision Rate = 64.2.

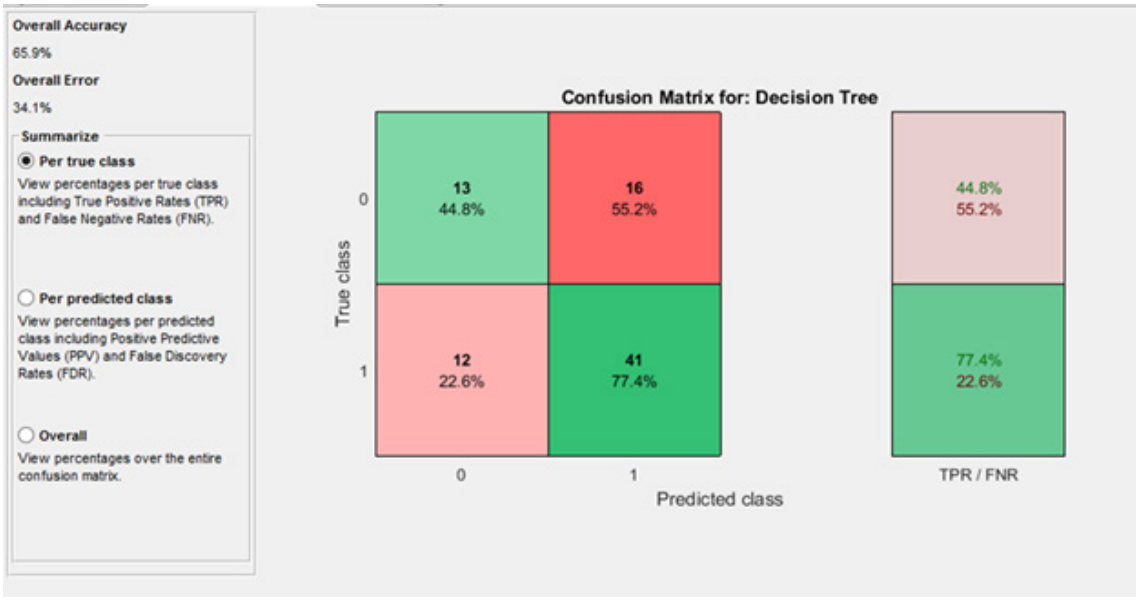

Figure 4: Confusion matrix result for decision tree
TP= 41; TN= 13; FP= 16; FN= 12
Accuracy rate = 0.659; Sensitivity Rate = 0.8039; Specificity Rate = 0.52; Precision Rate = 0.7735

with the Decision tree and SVM, it was proven that the classification efficiency of the decision tree outperformed SVM, hence the MapReduce-ANOVA with SVM achieved 77.4 Accuracy, the MapReduced-ANOVA helped in fetching out relevant information that can be helpful in decision making for engineers in the field of electricity supply and power consumption. It will also regulate the theft in electricity distribution. A comparative table is shown in table 2.

Table 2. Comparative Table for the Evaluation of the Experiment

Performance Metrics	SVM Classifier	Decision Tree Classifier	MapReduce-ANOVA + SVM	MapReduce-ANOVA + Decision Tree
Accuracy (%)	63.4	65.9	77.4	72.6
Sensitivity (%)	98.1	80.39	92.5	75
Specificity (%)	34.5	52	50	68.2
Precision (%)	64.2	77.35	77.1	81.1

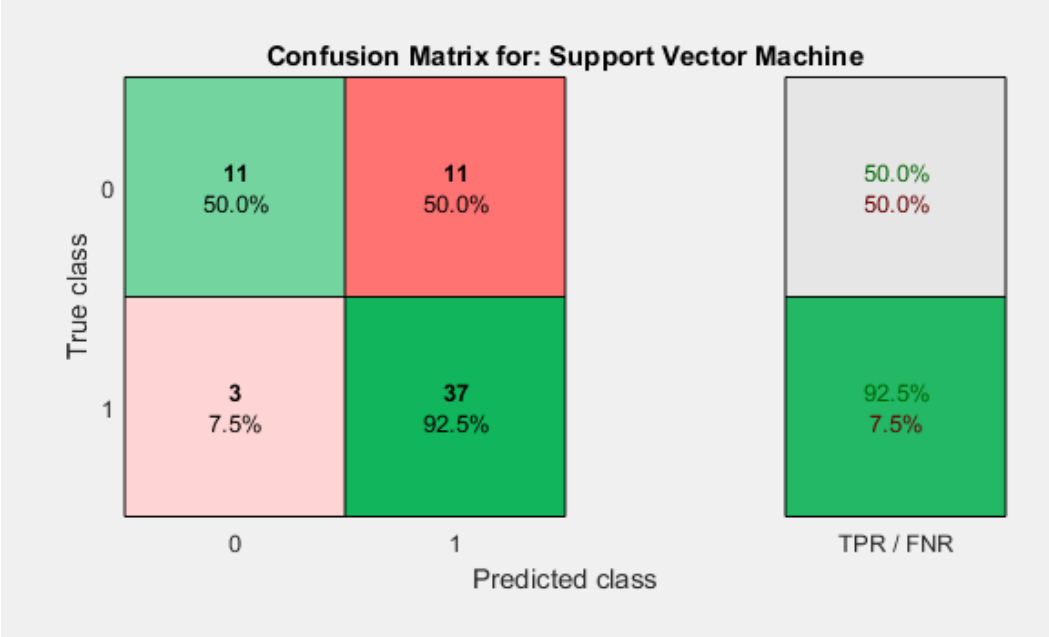

Figure 5: Confusion matrix for the classification of power consumption using MapReduce-ANOVA +SVM TP=37; TN=11; FP = 11; FN = 3

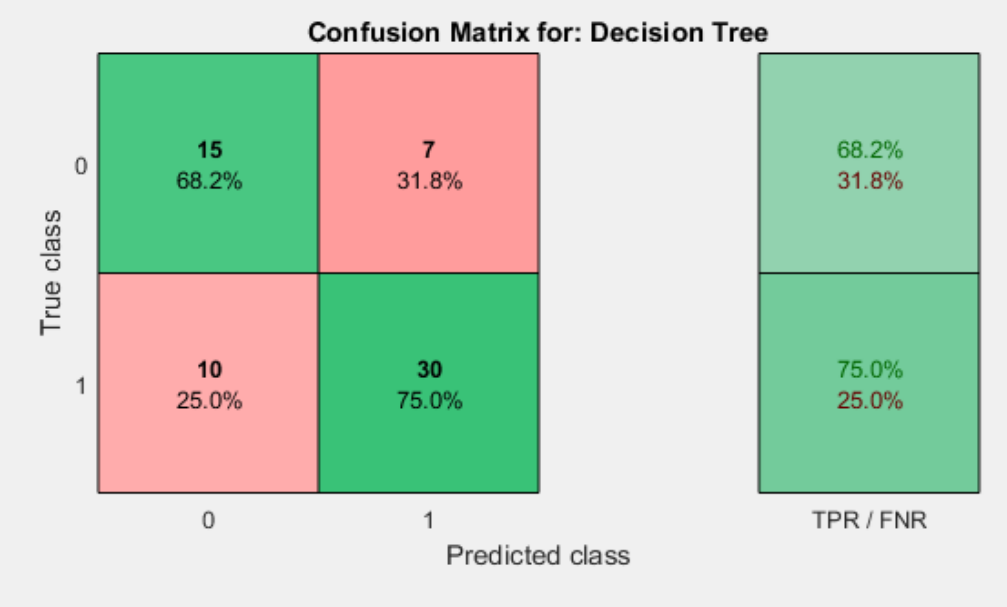

Figure 6: Confusion matrix for the classification of power consumption using MapReduce-ANOVA + Decision Tree P=30; TN=15; FP=7; FN=10

In this study, a novel MapReduced based ANOVA with 0.5 significant value was proposed as a procedure for detection of fraud in electricity distribution in localities in Nigeria. It also addresses the problem associated with noisy data and fetching for relevant information that can enhance data analysis for relevant decision making. It is evident that most classification decisions are often biased, leading to misclassification. MapReduced-ANOVA proposes solution to enhance the data classification using SVM and Decision Tree. The knowledge shows that SVM classifier performs better with 77% compared to the decision tree with 73%. The proposed method performs better compared to other state-of-the-art algorithms. The significance of this study is solving misclassification problems from a perspective of effective and efficient approach to accurately classifying datasets, which can help practitioners in the field to make relevant decisions, that will help them with achieving efficient output of dispensing and controlling electricity theft in the community. This study proves to solve the mitigation of the study.

5. Conclusion

In this study, a machine learning approach is carried out using classification algorithms to carry out the irregularity's affection for electric power consumption in Nigeria. The study used efficient classification algorithms SVM and Decision tree. The decision tree was able to achieve 65% accuracy, and SVM achieved 62% performance accuracy. The SVM sensitivity proves to be high and can still be considered highly efficient. Mapreduce based ANOVA was a great approach in helping to fetch out relevant information that can enhance the classification results. The results obtained show that the power consumption sector subject has to do more in customer reviews and satisfaction so as not to lose their revenue generation. Hence, this study has helped to improve its management of power supply and revenue protection. Although the outcome of the course is favourable, working on a more extensive data set is vital before using the SVM classifier for prediction to be put into practice in the real-world context. The concept of data mining has also been actualized during the project has the case study dataset was subject to clustering techniques. The classification with the Decision tree behaved quite well as it was able to reach a 65% accuracy which shows confidence in its knowledge discovery from the dataset used.

References

Angelos, E. W. S. D., Saavedra, O. R., Cortes, O. A., & Souza, A. N. (2011). Detection and Identification of Abnormalities in Customer Consumptions in PowerDistribution Systems. *IEEE Transactions on Power Delivery*, *26*(4), 2436–2442.

Arowolo, M. O., Adebiyi, M., Adebiyi, A., & Okesola, O. (2020). PCA Model For RNA-Seq Malaria Vector Data Classification Using KNN And Decision Tree Algorithm. *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, 1–8. https://doi.org/10.1109/ICMCECS47690.2020.240881

Awoyemi, J., Adetunmbi, A., & Oluwadare, S. (2018). Effect of Feature Ranking on Credit Card Fraud Detection: Comparative Evaluation of four techniques,". *2nd International Conference on Information and Communication Technology and Its Applications.*, 140–147.

Coma-Puig, B., & Carmona, J. (2018). A quality control method for fraud detection on utility customers without an active contract. *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 495–498. https://doi.org/10.1145/3167132.3167384

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa., O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*. https://doi.org/10.2405844018353404

Damanik, I. S., Windarto, A. P., Wanto, A., Poningsih, Andani, S. R., & Saputra, W. (2019). Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm. *Journal of Physics: Conference Series*, *1255*, 012012. https://doi.org/10.1088/1742-6596/1255/1/012012

Glauner, P. (2017). The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey. *International Journal of Computational Intelligence Systems*, *10*(1), 760–775.

Gong, J., Zhang, H., & Du, W. (2020). Research on Integrated Learning Fraud Detection Method Based on Combination Classifier Fusion (THBagging): A Case Study on the Foundational Medical Insurance Dataset. *Electronics*, *9*(6), 894. https://doi.org/10.3390/electronics9060894

Hambali, M. A., Saheed, Y. K., Gbolagade, M. D., & Gaddafi, M. (2017). Artificial Neural Network Approach for Electric Load Forecasting in Power Distribution Company. *Academia Journal Universiti Teknologi*, *6*(2), 23–33.

Hambali, M., Akinyemi, A., Oladunjoye, J., & Yusuf, N. (2017). Electric Power Load Forecast Using Decision Tree Algorithms. *Computing, Information Systems, Development Informatics & Allied Research Journal*, *7*(4), 29–42.

Hasan, M. N., Toma, R. N., Nahid, A.-A., Islam, M. M. M., & Kim, J.-M. (2019). Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. *Energies*, *12*(17), 3310. https://doi.org/10.3390/en12173310

Jeyaranjani, J., & Devaraj, D. (2018). Machine Learning Algorithm for Efficient Power TheftDetection using Smart Meter Data. *International Journal of Engineering and Technology, 7*(3), 900–904.

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid. *IEEE Transactions on Industrial Informatics*, *12*(3), 1005–1016. https://doi.org/10.1109/TII.2016.2543145

Jokar, P., Arianpoo, N., & Leung, V. . (2016). Electricity Theft Detection in AMI Using Customers' Consumption Patterns. *IEEE Trans. Smart Grid*, *7*(1), 216–226.

Khine, A. A., & Khin, H. W. (2020). Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree. *2020 IEEE Conference on Computer Applications(ICCA)*, 1–4. https://doi.org/10.1109/ICCA49400.2020.9022815

Kumar, M., Rath, N. K., & Rath, S. K. (2016). Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. *Journal of Biomedical Informatics*, *60*, 395–409. https://doi.org/10.1016/j.jbi.2016.03.002

Li, S., Han, Y., Yao, X., Yingchen, S., Wang, J., & Zhao, Q. (2019). Electricity Theft Detection in Power Grids with

Deep Learning and Random Forests. *Journal of Electrical and Computer Engineering*, *2019*, 1–12. https://doi.org/10.1155/2019/4136874

Otuoze, A. O., Mustafa, M. W., Sofimieari, I. E., Dobi, A. M., Sule, A. H., Abioye, A. E., & Saeed, M. S. (2019). Electricity theft detection framework based on universal prediction algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *15*(2), 758–768.

Paluszek, M., & Thomas, S. (2017). *MATLAB Machine Learning*. Apress. https://doi.org/10.1007/978-1-4842-2250-8

Patil, S., Nemade, V., & Soni, P. . (2018). Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. *Procedia Comput. Sci.*, *132*, 385–395.

Sahin, Y., & Duman, E. (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. *Proceedings of the International Multiconference of Engineers and Computer Scientists 2011*, *1*, 1–6.

Xiong, X., Cheng, Z., Chen, G., Zhang, Y., Fu, M., & Liu, M. (2019). A SVM-Based Fraud Detection System Using Short-lived Electricity Consumption Data. *Proceedings of the 2nd International Conference on Information Technologies and Electrical Engineering*, 1–6. https://doi.org/10.1145/3386415.3387061

Yeckle, J., & Tang, B. (2018). Detection of Electricity Theft in Customer Consumption Using Outlier Detection Algorithms. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, 135–140. https://doi.org/10.1109/ICDIS.2018.00029