Pattern Recognition using Support Vector Machines as a Solution for Non-Technical Losses in Electricity Distribution Industry

Azubuike N. Aniedu, Hyacinth C. Inyiama, Augustine C. O. Azubogu, Sandra C. Nwokoye

Abstract*: *Contending with Non-Technical Losses (NTL) is a major problem for electricity utility companies. Hence providing a lasting solution to this menace motivates this and many more research work in the electricity sector in recent times. Non-technical losses are classed under losses incurred by the electricity utility companies in terms of energy used but not billed due to activities of users or malfunction of metering equipment. This paper therefore is aimed at proffering a solution to this problem by first detecting such loopholes via the analysis of consumers' consumption pattern leveraging Machine learning (ML) techniques. Support vector machine classifier was chosen and used for classifying the customers' energy consumption data, training the system and also for performing predictive analysis for the given dataset after a careful survey of a number of machine learning classifiers. A classification accuracy (and subsequently, class prediction) of 79.46% % was achieved using this technique. It has been shown, through this research work, that fraud detection in Electricity monitoring, and hence a solution to non-technical losses can be achieved using the right combinations of Machine Learning techniques in conjunction with AMI technology.

Keywords*: *Clustering, classification and association rules, Correlation and regression analysis, Machine learning

I. INTRODUCTION

Energy distribution companies have, over the years, been faced with a barrage of problems in terms of equitable distribution of energy to users, proper metering and monitoring of usage, dealing with both technical and non-technical losses, to mention but a few. Whereas technical losses are energy losses encountered due to friction, heat conversion, electromechanical or magnetic losses in the transmission/distribution equipment and cables, Non-Technical losses have been described as any consumed energy or service which is not billed because of measurement equipment failure or Ill-intentioned and fraudulent manipulation of said equipment [1].Theft and vandalism especially, are major issues that utility companies have had to contend with over the years. Energy distribution companies have had to part with billions of dollars to vandals and fraudsters due to theft, illegal generation and sale of tokens, fraudulent diversion of collected funds by staff and so on. But among all these, energy theft has been the most difficult to contend with, especially in emerging economies. About 50% of electricity consumption in developing nations are gotten through energy theft [2]. United States of America records a loss of more than 6 billion US Dollars through energy theft each year [3]. In Canada, BC Hydro reports 100 million dollars in losses every year [4]. Utility companies in India and Brazil incur losses around 4.5 billion and 5 billion dollars respectively due to electricity theft [5], [6]. This creates a lot of avoidable losses in the system which affects the quality of supply, the electricity load on the generating station, and the tariff im-posed on usage by genuine customers. Improper monitoring system also prevents the authorities from knowing exactly how much money had actually been realized from sales of energy.The architecture of the electricity metering system is being changed in recent times from off-line, semi-independent systems to a fully integrated online system described as advanced metering infrastructure (AMI) as one of the ways of surmounting this growing need in the sector. Advanced metering infrastructure (AMI) is a hierarchical structure comprising the networking of varied digital or electronic hardware and software which com-bine interval data measurement with continuously avail-able remote communication. It enables measurement of detailed, real-time information and frequent collection and transmission of such information to various parties for proper monitoring and recording. A key component of the AMI is the smart meter. Smart meters are energy meters which in addition to basic metering capabilities, are equipped with additional features enabling them to communicate to remote servers, monitor and control energy usage by home appliances with options of remote management (disconnection, reconnection, etc) and cred-it recharge. However, recent researches has shown that even AMI-based systems can be defrauded and some of its security features bypassed. Most of these AMI-based threats have been categorized into physical attacks, cyber hacks and data attacks and hijacks and these invariably undermines the efforts put in AMI in preventing NTL necessitating the need for further research towards curbing this issue. In view of this, the attention of researchers have been turned towards the investigation and analysis of energy usage information of electricity customers with a view of discovering a pattern which could be used to distinguish between a valid customer and a fraudulent one.

Manuscript received on March 09, 2021.
Revised Manuscript received on March 15, 2021.
Manuscript published on March 30, 2021.

*****Correspondence Author

Azubuike N. Aniedu*, Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria. Email: an.aniedu@unizik.edu.ng

Hyacinth C. Inyiama, Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria. Email: hc.inyiama@unizik.edu.ng

Augustine C. O. Azubogu, Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria. Email: ac.azubogu@unizik.edu.ng

Sandra C. Nwokoye, Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria. Email: sc.nwokoye@unizik.edu.ng

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

This research work therefore focused on the analysis of consumers' consumption pattern leveraging machine learning techniques.

II. RELATED LITERATURE

As is it with every human invention, AMI-based energy management has its own issues and challenges. As noted in the introductory part of this paper one major challenge is the issue of theft. Several authors agree that energy theft continues to be a major issue to utility companies, even after deployment of smart meters. Stephen McLaughlin et al in their papers titled 'Energy theft in the Advanced Metering Infrastructure' [7] and 'AMIDS: A Multi-Sensor Energy Theft Detection Framework for Advanced Metering Infrastructures' [8] insist that the single requirement of energy theft is the manipulation of the demand data. In their work they identified three ways to tamper with the demand data, which includes a) while it is recorded (via electromechanical tampering), b) while it is at rest in the meter, and c) as it is in flight across the network. Sudheer K. et al [9] categorized electricity theft under several categories including: By under voltage technology, by under current technology, Stealing electricity by phase-shifted technology, Stealing electricity by difference expansion technology.Regardless of how they occur, it is becoming increasingly obvious that detection and subsequent prevention of non-technical losses cannot be done easily without the assistance of smart grids and smart meters. Hence it is no wonder that several authors have suggested different ways of detecting these losses. Nabi Mohammad et al [10] postulated measures for controlling electricity theft in their work "A smart prepaid energy metering system to control electricity theft". According to them, measures like protection against shorting the phase line and dis-connecting the neutral line, protection against whole meter bypassing, control of electricity theft using observer meter and protection against tampering can be achieved using a smart prepaid energy metering system and AMI. Some of other suggestions on tackling this menace include a proposal to use genetic algorithm and support vector machines (SVM) to detect abnormalities by Nagi et al [11]. Monedero et al [12] proposed the application of data mining techniques including use of neural networks and statistical techniques for these detections. Thomas Hartmann and co [13] proposed the use of live machine learning using multi-profiling. Nizar and co [14] in the same vain presented a novel approach for analysing NTL using modern computational technique called extreme learning machine (ELM). AlsoDepuru S. [15] employed several data classification algorithms including Support Vector Machines (SVM), Genetic Algorithm (GA), Neural Network (NN) model, and Rule Engine Algorithm in order to classify illegal consumers based on their energy consumption patterns.In view of these scholarly works, it is clear that one of the major ways via which non-technical losses can be curbed in the electricity sector is through the application of machine learning technique towards recognition of inconsistency in usage patterns in customer's electricity usage information. One of such techniques is the Sup-port Vector Machines.

III. METHOD AND SYSTEM ANALYSIS

A. Data Acquisition, Preparation and Feature Selection

In order to monitor and correctly analyze electricity consumption of consumers, instantaneous electricity readings of each individual consumer has to be captured and transmitted to a central location for such analysis. This has been made possible by smart meters and AMI technology. Hence live electricity usage data from the UCI Machine Learning Repository [16] were sourced for use in this analysis and research. The Dataset is a real instantaneous time-series electricity consumption of 370 clients taken at 15 minutes interval between January 2011 and December 2014. That is, each client has about 96 instances of their energy consumption taken daily for a period of 4 years. The energy readings are in kilowatts.

Because using the whole 4 year data of 370 clients would be overbearing to the classifier, and also because some of the years (like 2011) have some missing data for some clients, the data for this analysis was streamlined to include only from 1st January 2012 through 31st December 2012 containing a total of 35131 attributes of the 370 clients (instances).

$$Total\ attributes, a_T = M_T \times N \times I \tag{1}$$

Where M_T is the total number of Months under con-sideration, N is the number of days in each Month and I is the number of instantaneous electricity energy reading of each client taken at 15 minutes interval.Out of the 35131 instances selected, a set containing 11808 instances were further selected as the training data. Hence

$$a_{Tt} = \sum M_{i2} \times N \times I \tag{2}$$

Where a_{Tt}represents the total training data and given that the months of the year are represented in a matrix S, in the form:

$$S = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \tag{3}$$

The M_{i2} formation hence selects the second month of the beginning of each season of the year. This was done so because electric energy consumption of users are affected by the season of the year, hence the dataset selected for training the classifier comprise of readings from January (Winter) , April (Spring), July (Summer) and October (Autumn) in order to capture variations in the different seasons of the year. The rest of the dataset (that is, $a_T - a_{Tt}$) was then used as the testing data.Since the collected data is not in the format of the classifier input, the data was further subjected to a transposition hence from (2),

$$a_{Tt} = (a_{Tt})^T \tag{4}$$

Where $(a_{Tt})^T$ is the transpose values ofa_{Tt}.

B. Selection of Classifier and Parameters

A number of research techniques exists for adoption when carrying out a classification analysis as is the case of this work. As a matter of fact, the work of Manuel Fernandez-Delgado and co [17] evaluated 179 different classifiers arising from 17 families and these are just the common ones as there are many more. Hence in choosing a suitable technique for this work, the following criteria were considered: support resource, ease of manipulation, openness of source, as well as of course performance matrix of these classifiers. From the work of Delgado et al, the Random forest (RF) version "are most likely to be the best classifiers" when "implemented in R and accessed via caret" [17]. The second best is the SVM (when implemented with LibSVM in C programming language), followed by the neural networks.

In a bid to ascertain the most appropriate classifier for the dataset used for this work, nine of the top listed classifiers were subjected to a paired corrected tester using Waikato Environment for Knowledge Analysis (WEKA) [18]. The result is given in Table 1 while a column chart is shown in Fig. 1.

Table 1: Result of accuracy test for nine data classifier.

S/N o	Classifier	Type	CV Accuracy
1	Zero Rules	Rule System	69.46
2	LibSVM	Support Vector Machines	76.97
3	Sequential Minimal Optimization (SMO)	Support Vector Machines	70.49
4	IBk (K-Nearest Neighbour)	Instance Based	76.11
5	REPTree	Decision Trees	69.46
6	Random Forest	Ensemble	79.68
7	Naïve Bayes	Bayesian	19.62
8	Multilayer Perceptron	Neural Networks	50.85
9	Simple CART	Decision Trees	69.45

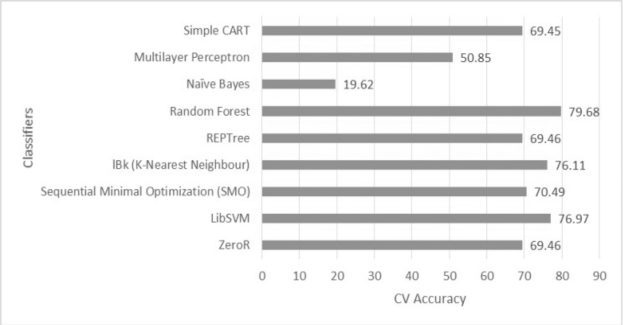


Fig. 1.Column Chart for CV Accuracy Test on Nine Classifiers

From Table 1 and Figure 1, it could be observed that classifier RandomForest performed best (79.68) followed closely by LibSVM (76.97). However the researcher in-tends to correlate with the results of other researchers in this field and compare results. So against these back-drops, the LibSVM classifier was chosen for classifying the dataset, training the system and also for performing predictive analysis for the given dataset.

Kernel functions in SVMs are selected based on the data structure and type of boundaries between the classes. The representative and widely applied kernel function based on Euclidean distance is the radial basis function (RBF) kernel, also known as the Gaussian kernel [19]:

$$K^{RBF}(x_{i,}x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \tag{5}$$

Where γ > 0 is the RBF kernel parameter. The RBF kernel induces an infinite-dimensional kernel space, in which all image vectors have the same norm, and the kernel width parameter "γ" controls the scaling of the mapping [19].

However, to ascertain the exact kernel most suited for the dataset used for this research work, the four kernel types of SVM was subjected to a paired Corrected Testing. The result of the test is given in Table 2 and shown in Fig. 2. It could be seen from that table that for the dataset, the RBF and Sigmoid kernel had the best cross validation accuracy of 79.46%. Therefore, the Radial Basis Function kernel was selected for the kernel for use in the SVM classifier.

Table 2: Paired corrected tester for four SVM kernel types

<table>
<tr><td colspan="5">Tester: weka.experiment.PairedCorrectedTTester
Analyzing: Percent_correct
Datasets: 1
Result sets: 4
Confidence: 0.05 (two tailed)</td></tr>
<tr><td>Dataset</td><td>(1)</td><td>(2)</td><td>(3)</td><td>(4)</td></tr>
<tr><td>4SeasonConsumption</td><td>79.46</td><td>72.86</td><td>71.41</td><td>79.46</td></tr>
<tr><td colspan="5">Key:
(1) functions.LibSVM 'Radial Basis Function: $exp(-\gamma\|u - v\|^2)$
(2) functions.LibSVM ' Linear: $u' \times v$
(3) functions.LibSVM ' Polynomial: $(\gamma \times u' \times v + coef0)^{degree}$
(4) functions.LibSVM ' Sigmoid: $\tanh(\gamma \times u' \times v + coef0)$</td></tr>
</table>

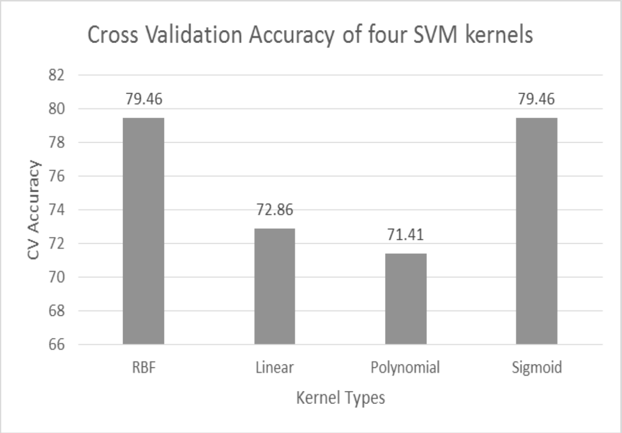


Fig. 2.Column Chart Showing Cross Validation Accuracy Testing for Four SVM Kernel Types.

Retrieval Number:100.1/ijisme.B1280037221
DOI:10.35940/ijisme.B1280.037221
Journal Website: www.ijisme.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication
© Copyright: All rights reserved.

C. Parameter Selection

To train SVM problems, specification of some parameters is key. As explained by Chang and co. [20], SVM makes available a simple tool to examine an array of parameters. For every parameter setting, SVM acquires cross-validation (CV) accuracy. Eventually, the parameters having the highest CV accuracy are selected. The selection tool for the parameters presumes that the RBF (Gaussian) kernel is used. The RBF kernel takes the form of (5) so (C,γ) are parameters to be decided. The cost parameter C is the parameter for the soft margin cost function, which determines how each individual support vector is influenced; this process entails a tradeoff of error penalty for stability. It "tunes" the algorithm between better fitting the available data or giving a larger margin.

Gamma, γ, is the free parameter of the Gaussian radial basis function. If γ is small, it implies the Gaussian has large variance implying that $x_j$will have more influence. What this means is that even when the distance between them is large, the class of the vector x_j will be determined by the class of the support vector x_i if the γ is small. If γ is large, the support vector's influence will not be widespread because the variance is small. In other words, large γ gives rise to high bias and low variance models. The reverse is also the case.

Possible intervals of C (or γ) were provided with the grid space. Thereafter, all grid points of (C,γ) were examined to determine which one giving the highest CV accuracy (see Table 3). The entire training-set were then trained using the best parameters in order to obtain the final model. From Table 3, it could be observed that (C,γ)≡(1.0,≥1) gave the highest CV accuracy. More advanced parameter selection methods were not considered because for only two parameters(C and γ), the number of grid points was not too large. For multi-class classification, a one-to-one method was adopted by LibSVM, under a specified (C,γ), to generate the CV accuracy. For all k(k-1)/2 decision functions, the same (C,γ) was proffered by the parameter selection tool.

Table 3: Determination of best (C,γ) for highest CV accuracy.

Tester: weka.experiment.PairedCorrectedTTester
Analysing: Percent_correct
Datasets: 1
Resultsets: 10
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 4/11/17 10:54 AM

Dataset	'4SeasonConsumption_abrid
(1)	39
(2)	39
(3)	39
(4)	39
(5)	50
(6)	47
(7)	50
(8)	50
(9)	50
(10)	50

Key:
(1) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(2) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 3.0 -E 0.001 -P 0.1 -seed 1' 14172
(3) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 30.0 -E 0.001 -P 0.1 -seed 1' 14172
(4) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 300.0 -E 0.001 -P 0.1 -seed 1' 14172
(5) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.01 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(6) functions.LibSVM '-S 0 -K 2 -D 3 -G 0.001 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(7) functions.LibSVM '-S 0 -K 2 -D 3 -G 1.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(8) functions.LibSVM '-S 0 -K 2 -D 3 -G 3.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(9) functions.LibSVM '-S 0 -K 2 -D 3 -G 30.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172
(10) functions.LibSVM '-S 0 -K 2 -D 3 -G 300.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1' 14172

From the result of the test (shown in Table 3) a stem plot (Fig. 3) was drawn showing the different (C,γ) combinations compared and the CV accuracy results obtained.

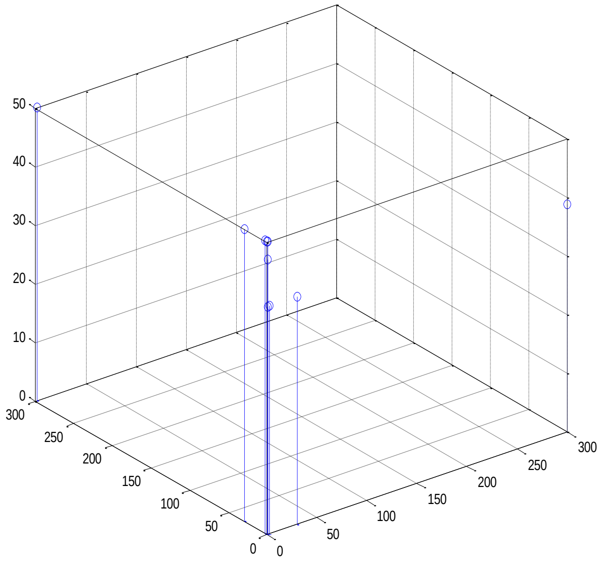


Fig. 3.Stem plot Cross Validation Accuracy Tester to Determine Highest of C,γ Values

D. Support Vector Machines (SVM)

SVM (Support Vector Machines) is a machine learning technique used extensively for pattern recognition and data analysis. Corinna Cortes and Vladimir Vapnik proposed the current standard which was adopted from the original algorithm invented by Vladimir Vapnik [21][22]. SVM seeks to create a hyperplane which would separate data sets into their classes. The main thrust is to assist the machine to discover structure from data and classify them into proper class labels [23]. The objective of SVM is to generate a model that, given only the attributes of the test data, can predict the target values (based on training data). It is to find the hyperplane (classifier) that maximizes the gap between data points on the boundaries (so called support vectors), assuming an in-finite number of such hyperplanes exists [24]. For example, given the hyperplane as shown (Fig. 4).

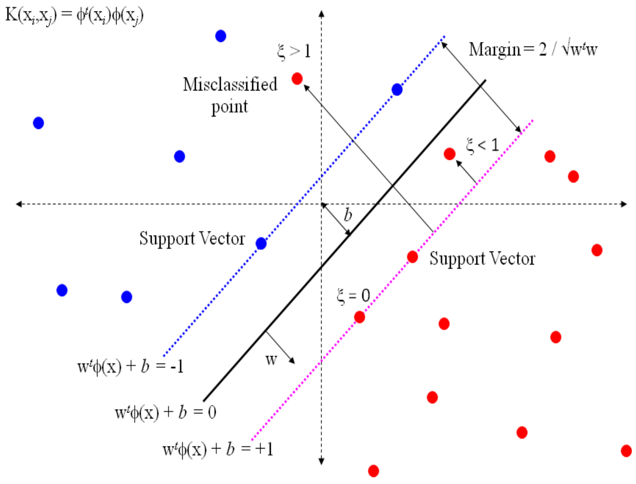


Fig. 4.Support Vector Hyperplane [24].

According to Nagi et al [11], in SVM, training is performed in a way such to obtain a quadratic programming (QP) problem. The solution to this QP problem is global and unique. For empirical data $(x_1,y_1),\ldots,(x_l,y_l) \in R^n * \{-1,+1\}$ that are mapped by φ:R^n→ F into a "feature space", the linear hyperplanes that divide them into two labeled classes can be mathematically represented as:

$$w * \emptyset(x) + b = 0 \qquad w \in R^n, b \in R \tag{6}$$

To construct an optimal hyperplane with maximum-margin and bounded error in the training data (soft margin), the following QP problem is to be solved [11]:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{7}$$

Subject to

$$y_i(w^T * \emptyset(x_i) + b) \geq 1 - \xi_i, \tag{8}$$
$$\xi \geq 0, i = 1, \ldots, l$$

where $\emptyset(x)$ maps $x_l$into a higher-dimensional space. The first term in cost function (7) makes maximum margin of separation between classes, and the second term provides an upper bound for the error in the training data. The constant C ∈ [0, ∞), called the regularization parameter, creates a tradeoff between the number of misclassified samples in the training set and separation of the rest samples with maximum margin. Due to the possible high dimensionality of the vector variable *w*, usually the following dual problem is solved

$$\min \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \tag{9}$$

Subject to

$$y^T\alpha = 0, \qquad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, l$$

Where $e = [1, \ldots, 1]^T$is the vector of all ones, Q is an l by l positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$ is the kernel function. After problem (9) is solved, using the primal-dual relationship, the optimal *w* satisfies

$$w = \sum_{i=1}^{l} y_i \alpha_i \emptyset(x_i) \tag{10}$$

And the decision function is

$$sgn\,(w^T * \emptyset(x) + b) = sgn\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b\right) \tag{11}$$

$y_i \alpha_i \forall_i, b$, label names, support vectors and other information such as kernel parameters are stored for prediction

From (6) it is seen that the optimal hyperplane in the feature space can be written as the linear combination of training samples with $\alpha_i \neq 0$.These informative samples, known as support vectors, construct the decision function of the classifier based on the kernel function:

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i k(x, x_j) + b\right) \tag{12}$$

LibSVM is a library for developing SVMs based on classification model developed by C.C. Chang and C.J. Lin [20]. it could be adapted to most machine learning knowledge environments like Matlab, R, Pyton, WEKA (Waikato Environment for Knowledge Analysis) etc. LibSVMtools uses LibSVM classifier as a wrapper class. It is used to build the SVM classifier and runs faster than other SVM tools.

SVM offers many advantages including, as listed by [24], flexibility in the choice of the form of the threshold, robustness towards small number of data points, delivery of unique solution etc.

E. Model Training

From the Primal SVM model given in (7) and the Gaussian kernel given in (5)

Utilizing the C-SVC SVM type, with the following parameters:

Cache size =40Mb,
Coefficient = 1,
Cost parameter C for C-SVC= 1.0,
Gamma = 1.0,
$K^{RBF} = \exp(-\gamma * |x_i - x_j|^2)$,
Weight for each class = (1 1 1 1)
Cross validation fold = 10

With standardized data, non-shrinking heuristic, the model was trained, subjecting it to a set of chosen attributes, cutting across the four seasons of the year.

IV. SYSTEM IMPLEMENTATION AND RESULTS

A. Data Classification Using Trained Model

Based on the model parameters given in the preceding section, the classifier haven been adequately trained, was used to classify the given data set. The summary of the classification of the 370 instances using the trained model is given in Table 4.

B. System Implementation: Class Prediction Using SVM

Out of the total dataset selected for analysis a further breakdown was done which filtered out four months representing the four seasons of the year. This is used as the training data whereas the remaining data is applied as testing data for prediction. Having successfully trained the model using the training data, the testing data was applied to it in order to ascertain the accuracy of the classifier's training and its ability to predict the class of each instance. Table 5 – Table 7 shows the output of the prediction run. The preliminary information including classifier scheme with its attributes and the number of instances and attributes associated with the supplied test data is given in Table 5.

A sample of the actual run information (20 instances) containing the predicted classes, error information and probability distribution of the classification can be seen in Table 6, whereas the summary of the run showing that 294 instances representing 79.45% of the total instances were correctly classified with a mean absolute error of 0.1027 is shown in Table 7.

Table 4: Summary of LibSVM classifier output on 370 instances.

```
=== Run information ===
Scheme:weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 1.0 -R
0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -seed 1
Relation:     4SeasonConsumption
Instances:    370
Attributes:   11808
Test mode: evaluate on training data

=== Classifier model (full training set) ===
LibSVM wrapper, original code by Yasser EL-Manzalawy (=
WLSVM)
Time taken to build model: 32.27 seconds

=== Summary ===
Correctly Classified Instances       294 (79.4595 %)
Incorrectly Classified Instances      76 (20.5405 %)
Kappa statistic                     0.442
Mean absolute error                    0.1027
Root mean squared error                   0.3205
Relative absolute error              42.2306 %
Root relative squared error               92.1927 %
Total Number of Instances                 370
```

=== Detailed Accuracy By Class ===

				Weighted Avg.
TP Rate	0	1	1	0.795
FP Rate	0	0.655	0.006	0.455
Precision	0	0.776	0.949	0.634
Recall	0	1	1	0.795
F-Measure	0	0.874	0.974	0.705
ROC Area	0.5	0.673	0.997	0.67
Class	RED	GREEN	BLACK	

```
=== Confusion Matrix ===
 ab       c   d   <-- classified as
 074          0   2   |     a = RED
 0257         0   0   |     b = GREEN
      0       0      37 |     d = BLACK
```

Table 5: Prediction Run Preliminary Information.

```
=== Model information ===
Filename:     libsvm-final run model.model
Scheme:weka.classifiers.functions.LibSVM -S 0 -K 2
      -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P
      0.1 -seed 1
Relation:4SeasonConsumption
Attributes:   11808
[list of attributes omitted]

=== Classifier model ===
LibSVM wrapper, original code by Yasser
      EL-Manzalawy (= WLSVM)
=== Re-evaluation on test set ===
User supplied test set
Relation:     prediction data
Instances:    370
Attributes:   11808
```

Table 6: Sample of Prediction Statistics

=== Predictions on test set ===

inst	actual,	predicted	error	probability distribution
1	1:RED	2:GREEN	+	0 *1 0 0
2	2:GREEN	2:GREEN		0 *1 0 0
3	2:GREEN	2:GREEN		0 *1 0 0
4	3:RED	2:GREEN	+	0 *1 0 0
5	2:GREEN	2:GREEN		0 *1 0 0
6	2:GREEN	2:GREEN		0 *1 0 0
7	2:GREEN	2:GREEN		0 *1 0 0
8	2:GREEN	2:GREEN		0 *1 0 0
9	1:RED	2:GREEN	+	0 *1 0 0
10	3:RED	2:GREEN	+	0 *1 0 0
11	2:GREEN	2:GREEN		0 *1 0 0
12	4:BLACK	4:BLACK		0 0 0 *1
13	2:GREEN	2:GREEN		0 *1 0 0
14	2:GREEN	2:GREEN		0 *1 0 0
15	4:BLACK	4:BLACK		0 0 0 *1
16	2:GREEN	2:GREEN		0 *1 0 0
17	2:GREEN	2:GREEN		0 *1 0 0
18	2:GREEN	2:GREEN		0 *1 0 0
19	2:GREEN	2:GREEN		0 *1 0 0
20	3:RED	2:GREEN	+	0 *1 0 0

Table 7: Summary Information for the Prediction Run.

```
=== Summary ===
Correctly Classified Instances    294  (79.4595 %)
Incorrectly Classified Instances  76 (20.5405 %)
Kappa statistic                  0.4608
Mean absolute error                 0.1027
Root mean squared error                0.3205
Total Number of Instances              370
```

It could be seen that 79.46% accuracy can be achieved in terms of predicting the class of users in based on their energy usage and utilization patterns (Table 7). This is a significant step in the right direction as such will go a long way in early detection of inconsistent use (characterizing fraud perhaps) of energy supplied. This could then be further inspected by utility personnel for confirmation purposes and hence prevent unnecessary wastage of generated power.

C. Result Discussion

From Table 7, the LibSVM correctly classified 294 out of the 370 considering a total of 11808 attributes which corresponds to 79.46% accuracy. It could also be observed (from the detailed accuracy by class section) that the Receiver Operating Characteristics (ROC) area (also called Area Under Curve (AUC), which is a plot of True Positive Rate (TP) against False Positive Rate (FP)) for RED class was 0.5% each while that of GREEN and BLACK classes were 0.673% and 0.997% respectively. AUC represents a probability that a positive will be ranked higher than a negative. The confusion matrix showed that the total 257 and 37 instances, belonging to the GREEN and BLACK class respectively were all correctly classified. Figure 5 – Figure 7 shows a plot of the area under ROC for the four classes.

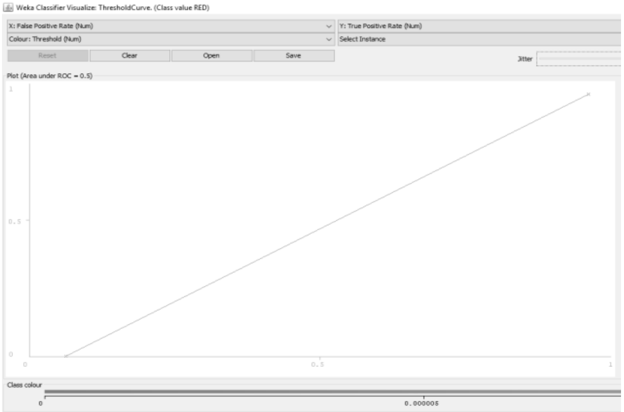
The f-measure (or f-score) is calculated using the formula

f-measure = (2 × Precision × Recall)/(Precision + Recall) (13)

Where Precision=TP/(TP+FP) and Recall=TP/(TP+FN)

TP = True Positives, FP = False Positives, FN = False Negatives

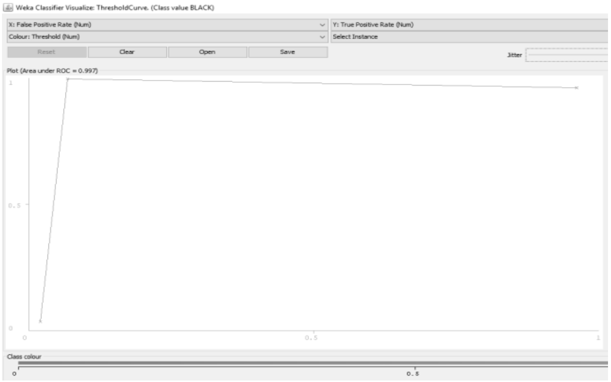
Precision is the percentage of elements correctly classified as positive out of all the elements classified as positive, while recall is the percentage of elements correctly classified as positive out of all the positive elements


Weka Classifier Visualize: ThresholdCurve. (Class value RED)

X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance

Plot (Area under ROC = 0.5)

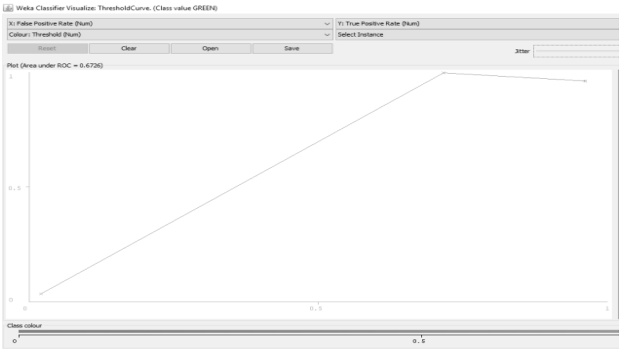
Fig. 5.Area Under ROC for Class Value RED.


Weka Classifier Visualize: ThresholdCurve. (Class value BLACK)

X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance

Plot (Area under ROC = 0.997)

Fig. 6.Area Under ROC for Class Value GREEN


Weka Classifier Visualize: ThresholdCurve. (Class value GREEN)

X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance

Plot (Area under ROC = 0.6726)

Fig. 7.Area Under ROC for Class Value BLACK

Each class *i* has a particular precision and recall in a multiclass case. Here, an element predicted to be in*i* and is truly in it is a "true positive", while an element predicted not to be in *i* and is not in it is a "true negative".The weighted f-measure is a weighted average of the classes' f-measure, and this weighting is determined by the proportion of the number of elements in each class.

V. CONCLUSION AND RECOMMENDATIONS

Monitoring and management of energy usage has be-come a trending issue among researchers and utilities chiefly due to the progressive search for the nagging is-sue of non-technical losses which continue to waste tons of funds for the utilities and the government. In this work, machine learning, particularly support vector machine (LibSVM) was employed in classifying users and also predicting user's activities in terms of energy usage, with the bid to detecting fraudulent users hence ensuring prompt disconnection of such from the grid. Using LibSVM, a library wrapper for SVM, the re-searcher was able to achieve a 79.46% accuracy in classification of users and subsequent prediction of future outcomes given similar parameters in terms of energy usage information. This research work concentrated on scenarios where users bypass their energy meters or manipulate the AMI infrastructure in such a way that consumed energy is not accurately recorded and hence not reported nor billed.

It is recommended therefore that further research be conducted taking into accounts other forms of fraudulently consuming energy not paid for like illegal generation of energy credits, partial connection of total house load to the meter and even the use of non-smart meter as is still obtained in most developing countries. It is also recommended that other aggregations of machine learning profiles be explored in Classification, regression etc. in other to improve the result achieved in this work.

REFERENCES

1. J. Parmar. (2013) Total Losses in Power Distribution & Transmission Lines-Part1. [Online]. https://electricalnotes.wordpress.com/2013/07/01/total-losses-in-power-distribution-transmission-lines-part-1/
2. P., Antmann, "Reducing technical and non-technical losses in the power sector. ," Washington, DC, USA, 2009.
3. P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," IEE Security and Privacy, vol. 7, no. 3, pp. 75-77, 2009.
4. CBC News. (2010) Electricity theft by B.C. grow-ops costs $100m a year. [Online]. http://www.cbc.ca/news/canada/britishcolumbia/electricity-theft-by-b-c-grow-ops-cost-ayear-1.969837
5. Ministry of Power, India. (2013) Overview of power distribution. [Online]. http://www.powermin.nic.in
6. Federal Court of Audit, Brazil, "Operational audit report held in national agency of electrical energy, Aneel, Brazil," Brazil, 2007.
7. S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy Theft in the Advanced Metering Infrastructure," in Critical Information Infrastructures Security, , Berlin, 2010, pp. 176-187.
8. S. McLaughlin, B. Holbert, S. Zonouz, and R., Berthier, "AMIDS: A Multi-Sensor Energy Theft Detection Framework for Advanced Metering Infrastructures.," Tainan, Institute of Electrical Electronic Engineering (IEEE), pp. 354 - 359, 2012.
9. K. Reddy Sudheer, P. Musthafa, and K. Sakthidhasan, "Equipment for Anti-Electricity Stealing with Remote Monitoring.," International Journal of Engineering Research and Applications, vol. 1, no. 2, pp. 241-245, 2013.
10. N. Mohammad, A. Barua, and M. A. Arafat, "A Smart Prepaid Energy Metering System to Control Electricity Theft," Institute of Electrical Electronic Engineering (IEEE), pp. 562-565, 2013.
11. J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and A. M. Mohammad, "Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines," IEEE Xplore, pp. 1-6., 2008.

Retrieval Number:100.1/ijisme.B1280037221
DOI:10.35940/ijisme.B1280.037221
Journal Website: www.ijisme.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication
© Copyright: All rights reserved.

12. I. Monedero, F. Biscarri, and C. Leon, "MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques.," Springer-Verilag, Berlin, pp. 725-734, 2006.
13. Thomas Hartmann et al., "Suspicous Electric Consumption Detection Based on Multi-Profiling Using Live Machine Learning," in 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2015.
14. A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis With Extreme Learning Machine Method.," IEEE Xplore, pp. 946 - 955, 2008.
15. S. S Depuru, Modeling, Detection and Prevention of Electricity Theft for Enhanced Performance and Security of Power Grid, 2012.
16. M. Linchman, UCL Machine Learning Repository, 2009.
17. M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," Journal of Machine Learning Research, vol. 15, no. 10, pp. 3133-3181, 2014.
18. The University of Waikato, Waikato Environment for Knowledge Analysis (WEKA). , 2013.
19. D. Wang, D. S. Yeung, and E. C. C. Tsang, "Weighted Mahalanobis Distance Kernels for Support Vector Machines," IEEE Transactions on Neural Networks, vol. 18, no. 5, pp. 1453-1462, September 2007.
20. C. C. Chang and C. J. Lin. (2013) LIBSVM - A Library for Support Vector Machines. [Online]. http://www.csie.ntu.edu.tw/~cjlin/libsvm
21. V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.
22. N. Cristianini and J. S. Taylor, An Introduction to Support Vector Machines. Cambridge, MA: Cambridge University Press., 2000.
23. K. Xie. (2011) Support Vector Machine - Concept and matlab build. [Online]. [Online]. www.egr.msu.edu/classes/ece480/capstone/spring11/group04/application_Kan.pdf
24. E Mokshyna. (2014) Support Vector Machines. [Online]. https://dl.acm.org/citation.cfm?id=1968168

AUTHORS PROFILE

Azubuike N. Aniedu, holds a PhD in Computer and Control Engineering from Nnamdi Azikiwe University. Awka. He is currently the Deputy Director ICT at Nnamdi Azikiwe University (NAU) and a senior researcher and lecturer in Electronic and Computer Engineering Department, NAU with major interest in computer/ control engineering, Machine learning, Data Science and computer networks. An IEEE Entrepreneurship Ambassador, a consultant with IEEE Consultancy Network, certified IBM IOT Cloud Developer and Cloud Application Developer, Certified Oracle Cloud Infrastructure Associate plus other certificates and badges. Chair and board member/committee member of various committees and organizations both within and outside the University, An editorial member of several high impact journals, a prime member of NAU Bioinformatics and Genomics Consortium, among others. He is widely published and has also won some research grants. A registered engineer (COREN), and member of IAENG, IEEE, ISOC and IACSIT.

Hyacinth ChibuezeInyiama, is a seasoned computer scientist and Engineer, with a wealth of experience in both industry and Academics. He obtained his Bachelor's degree in Computer Technology from University of Wales, University College of Swansea, U.K (1978). For his Postgraduate Studies, he went to University of Manchester, Institute of Science and Technology (UMISI UK). He also worked at University of Manchester, Institute of Science and Technology (UMIST) as a special research assistant for years where he eventually did his PhD in Microprocessor-based electronic control (1981). His research was sponsored by the British Science Research Council (BSRC) because of its novel nature. He has been a professor of Computer Engineering for more than 20 years. He is the Edi-tor-in-Chief of JICCOTECH and has over 100 Journal publications, 8 books and many other Technical reports and conference papers. He is duly registered professionally as both computer scientist and computer Engineer.

Augustine C.OAzubogu, obtained a B.Eng (1986) in Electronic Engineering from University of Nigeria Nsukka, M.Eng (2005) in Electronics & Telecommunication from University of Port Harcourt and Ph.D (2011) in Wireless Communication Engineering and Digital Signal Processing from Nnamdi Azikiwe University Awka. Azubogu, A.C.O is an experienced registered engineer and an academic of over 25 years' experience. He joined Nnamdi Azikiwe University faculty in 2005. He is currently a Professor of Digital Signal Processing and Wireless Communication in the department of Electronic & Computer Engineering and a Research Associate with the Centre for Sustainable Development, Nnamdi Azikiwe University, Awka, Nigeria. His research interests include: Applications of Adaptive Antennas in Wireless Communication, Quantifying and Exploiting TV Whitespace for Broadband Internet Access, Convergence of AI Technologies and Wireless Communication and Wireless Sensor Networks and Applications in Infrastructure Monitoring. Prof. Azubogu has published a book and more than 35 papers in refereed journals and conferences. He is a registered Engineer with COREN and is a member of IEEE.

Sandra Chioma Nwokoye, holds a B.Eng in Computer Engineering from Nnamdi Azikiwe University, Awka and a PGDE from Nnamdi Azikiwe University, Awka plus other certifications. She is a Technologist in the department of Electronic and Computer Engineering, NAU with major interest in computer/control engineering, Machine learning, Data Science, she also has published articles. A member International Association of Engineers (IAENG), Association of Professional Women Engineers in Nigeria (APWEN).