LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374E-ISSN:1855-363X
VOL.23,NO.10(2025)

ENHANCEMENT DETECTION OF ELECTRICITY THEFT UNDER PART XIV OF ELECTRICITY CRIMINAL ACT, 2003 IN DISTRIBUTION SYSTEMS USING RANDOM FOREST FED EXTREME LEARNING MACHINE

Dr. KARTHIKEYAN RAMASAMY[1], VEERAMANI SHANMUGAM[2], Dr.DURGADEVI VELUSAMY[3]

[1]Professor, Department of Electrical and Elecronics Engineering, M.Kumarasamy College of Engineering, Karur – 639113, Tamilnadu, India.
[2]Assistant Professor, Department of Electrical and Elecronics Engineering,Trichy Engineering College,Konalai, Trichy – 621132,Tamilnadu, India.
[3]Assistant Professor, Department of Information Technology, Sri SivasubramaniyaNadar College of Engineering,Kalavakkam, Chennai – 603110,Tamilnadu, India.

Corresponding Author: **Dr.KARTHIKEYAN RAMASAMY**
papkarthik@gmail.com

Abstract. Day-to-day electricity fraud and thefts in utilities have increased by distribution and consumer levels, and they are also considered non-technical losses. The primary objective of this work is to detect electrical power theft in the distribution system by using a Random Forest (RF) algorithm fed Extreme Learning Machine (ELM), Extereme Gradient Boosting (XGBoost) and Categerical Boosting (CatBoost), which is an advanced algorithm from aCatboostRF for Support Vector Machine (SVM) classifier. The paper proposes the framework of training a models and testing a models of the data by using information data set to predict the target value. It gives them more efficient when compared to the already used techniques. The two-level data processing approach is proposed in this paper since the data processed by CatboostRF are provided as input to the ELM classifier. The results show that it reduces the deviation error and increases the classification rate accuracy.

Keywords: Random Forest (RF), Electrical Theft, Support Vector Machine (SVM), Categerical Boosting (Catboost), Extreme Learning Machine (ELM), Confusion Matrix.

1. Introduction

These days, electrical utilities must address non-technical losses caused by theft committed by fraudulent consumers. Electricity theft is a criminal practice and should be punishable by confinement. It is beneath the non-technical losses. (Newswire, 2014) The annual rising Markets Smart Grid survey electricity theft and its result is the world loss INR 6160.16 billion annually. The highest losses were in India (INR 1117.39 billion), followed by Brazil (INR 724.24 billion) and Russia (INR 351.77 billion). Ben

Gardner, President of northeast group stated that india has huge money to electricity theft than other country. The state of Maharashtra has the highest power losses, including Mumbai alone INR 193.13 billion per year. Approximately 15.25% of losses are due to total transmission and distribution, and few states' losses exceed more than 50%. There are two main types of power theft: hooking lines and bypassing the energy meter. According to a study, 80% of theft occurs among domestic consumers, while 20% takes place in commercial and industrial premises.

Under Part XIV of the Electricity Act, 2003, Section 135 of Indian law clearly defines electricity theft as a punishable offence. The Act includes the tapping of overhead or underground power lines, tampering with or bypassing meters to manipulatelectricity usage readings, and using artificial means to interfere with the proper registration of consumption. It also covers situations where individuals or companies consume electricity without authorization, either by directly/indirectly connecting to the grid or by misusing sanctioned connections for purposes other than those approved.

- Direct hooking from the line: It is the most used method for stealing power, hooking the power line frontwards of the energy meter, and it causes unmeasured electricity during theft.
- Jumping line (Overlapping) the energy meter: The input connection is bypassed and directly connected to the output terminal or load.
- Additional setup into the energy meter: Meters are shaped by installing a circuit indoors due to it can be slowed down at any time. The meter works healthily until the remote is on.
- Physical barrier: It is used in electromechanical meters with a rotating disc. It creates a barrier to the disc by using an additional magnet inserted inside /above the meter.
- ESD attack on electronic meter: This type of interference happens only on the electronic meter, making it either electrically short or breaking down.

Electrical energy theft detection has many techniques and metering arrangements. In a smart grid, many smart meters are available to monitor the customer's electrical consumption. Still, fraud activities are happening in the power distribution system. These come under the non-technical losses that happen due to irregular supply. The author Carbal (Cabral,2004) proposed the Rough set to detect theft by using the classification rule, and it detects the fraud consumer with a rightness rate of 20%. Still, it has the

disadvantage of fraudulent consumer behaviour, which seems normal due to lower approximation and produces a high false positive.
Fraud on the distribution side was identified by wavelet techniques and by combining multiple classifiers proposed by Rang Jiang (Jiang et al., 2002). The multiple classifiers worked through a cross-identification and a voltage scheme. The classification accuracy reached 78% on the training set and 70% on the dataset (Costa et al., 2013). Wavelet domain analysis and wavelet transforms are relatively insensitive to these slow variations. Another strategy to discover electrical theft using an Artificial Neural Network-based Knowledge Discovery process has been used, which has an accuracy of 65.03%. But, this used historical data.

Nowadays, finding fraudulent electrical consumption by efficient metering arrangements has been an active research area. The Support Vector Machine is another essential technique used in the non-technical loss detection that fraudulent consumers cause in the distribution system. (Nagi et al., 2008) J. Nagi and A.M. Mohammand proposed an SVM technique to find theft in electricity in the field of non-technical loss analysis. The fraud detection sucess rate increased from 22% to a satisfactory 53%. SVM has the primary methodology of Data acquisition, pre-processing, and normalization. However, the drawback of this scheme was that load profile inspection was performed manually. On the other, the novel proposed SVM-based data classification for electricity theft. (Depuru et al., 2011) Which is the data set from consumer energy consumption patterns based on historical data. Then, SVM is trained based on the data set, and the resultant accuracy is 60%. This proposal labels consumers as genuine or illegal. To improve the efficiency of SVM, JawartNagi(Nagi et al., 2011) outlines the framework that increases the hit rate value of SVM by using the Fuzzy inference system. In that FIS, it is an if-then rule with 72% accuracy. However, in this approach, the achieved hit rate value is still low, which suggests a high misclassification concerning fraudulent consumer detection. To increase the fault detection accuracy by more than 90%, Anish Jindal outlines the proposal for DT and SVM-based data analysis for theft detection. In that, the output of the DT is fed to SVM inputs, and its accuracy is 92%. (Durgadevi et. Al 2020)

Out of possession of above discussed, it can be summarized that the above-discussed paper struggles to give high accuracy due to the fallowing reasons,

(i) The excising approaches have high error values and hit rates and need additional techniques for improving efficiency.

(ii) The overall inspection cost increases due to theft detection for a high false positive rate.

(iii) Most existing proposals primarily focus on detecting theft at the consumer's end.

To increase the efficiency of detecting electricity theft, this paper proposed the framework of the machine learning algorithm of RF and SVM classifiers. RF classifies the best voting from multiple decision trees, and this output is fed to the inputs of SVM, which gives the estimated fraudulent consumers. The SVM is primarily used in the area of analytic application to solve classification problems (Tasi et al., 2015) SVM has the advantage of giving more accuracy when compared to any other classification technique (Jokaret al., 2016) and (Nabilet al., 2019). It serves as an approximation of a bound on the test error rate, and a substantial body of theory suggests it effectiveness. Another one is it avoiding the over-fitting (Keskesand Braham,2015).

2. Literature review

Angelos et al. (2011) introduced a novel computational approach for analyzing and categorizing electricity consumption patterns. The proposed method involves a two-stage process. Initially, the authors apply a C-means-based fuzzy clustering algorithm to group consumers with similar usage profiles. In the subsequent stage, they employ fuzzy classification by utilizing a membership matrix and calculating the Euclidean distance from each point to the cluster center. These distances are then normalized and ranked to produce a composite index score, highlighting consumers with potential anomalies or unusual consumption behaviors. The proposed technique was rigorously tested using real-world data, demonstrating its effectiveness in detecting fraudulent activities and identifying measurement errors.

Jokar et al. (2016), introduced an innovative method for analyzing Non-Technical Losses (NTLs) in electricity distribution using an intelligence-based technique, specifically Support Vector Machines (SVM). The primary aim of their study is to support TenagaNasionalBerhad (TNB) in Malaysia in mitigating NTLs caused by electricity theft. The proposed model identifies potential fraudsters by analyzing irregularities and atypical consumption patterns. This data mining approach involves extracting

features from historical consumption data and applying SVM to detect abnormal behaviors strongly associated with NTL activities. The model categorizes consumption profiles to prioritize customers for onsite inspection based on significant deviations in their usage patterns. Simulation results demonstrate that this SVM-based approach outperforms existing measures employed by TNB in reducing NTL incidents.

Nagi et al. (2011), enhanced a Support Vector Machine (SVM)-based fraud detection model by incorporating a fuzzy inference system (FIS). This integration introduces fuzzy IF-THEN rules, which incorporate human expertise into the detection process. The FIS serves as a post-processing layer, refining the model's ability to identify customers with a higher likelihood of fraudulent activity. By applying this improved SVM-FIS approach, the fraud detection rate for TenagaNasionalBerhad Distribution increased from 60% to 72%, demonstrating a significant boost in effectiveness and cost-efficiency.

Depuru et al. (2011), addressed the challenges associated with detecting electricity theft and review existing methods for mitigating such issues. They introduce an analysis of energy consumption patterns among customers, distinguishing between typical patterns and those indicative of theft. Utilizing historical data, they create a dataset representing various theft scenarios and train Support Vector Machines (SVMs) using data from smart meters. The trained SVMs are then used to classify and group suspicious consumption profiles based on predefined rules. The paper presents the outcomes of this classification process, highlighting the effectiveness of the Catboost in identifying potentially fraudulent energy usage.

3. Contributions

The primary contribution of this paper (work) is the detectof electricity theft in consumers and fraud in the connections distribution side. These cause the framework of detecting electricity theft in consumer connections (domestic) and fraud in transmission and distribution (bulk consumers). This paper's major contribution is using the Random Forest (RF) algorithm fed Extreme Learning Machine (ELM), Extereme Gradient Boosting (XGBoost), and Categerical Boosting (CatBoost), to classify the customers based on the fraudulent. The proposal requires historical data on consumer power usage. This framework collects the parameters from the dataset based on consumer historical data for classification purposes.

4. Proposed Scheme

Robert Czechowski(Czechowski and Kosek,2016) explained the main reasons for electrical energy theft are safety, economy and society, and the techniques used to steal electrical energy are (1) housing drilling and (2) hidden connections in front of the meter. Electrical theft is detected at three levels: (i) Transmission level, (ii) Distribution level and (iii) Consumer level. This proposal has mainly focused on the consumer level. Because theft at the other two levels is easily identified using losses in the line. But, at the consumer level, many fraudulent activities are there. It is the reason only framework uses the machine learning techniques of RF for SVMclassifier to classify the typical consumers and fraudulent consumers (Karthikeyanet. Al 2022). The basic block diagram of the combination performance evaluation of given below.

Figure 1 Shows, the input parameters are given to both RF and SVM as input. RF is classified using input parameters, then the expected electricity consumption is fed to the SVM classifier, and the actual consumption is also given to Catboost. As a result, Catboost found the malevolent consumers involved in electrical energy theft. Finally, this proposal has a high accuracy rate due to this combination.

Block Diagram:

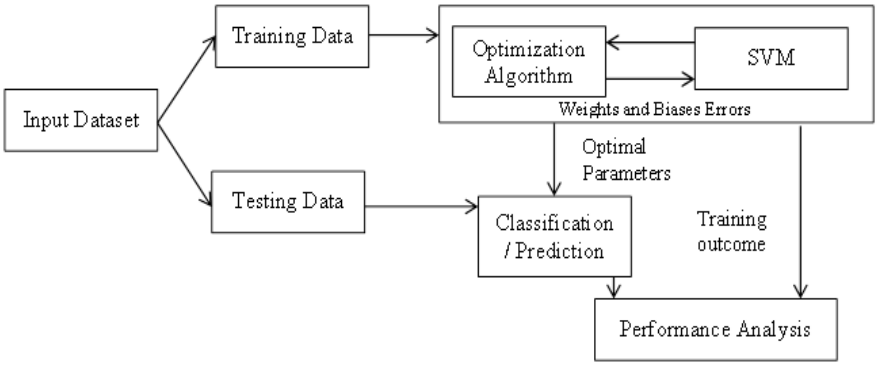

Figure1. Block Diagram of Proposed Scheme.

4.1. Random Forest

Random Forest is one of the most popular and powerful methods used today in Machine Learning Techniques in the field of computer science. The

decision tree has the problem of overfitting, which is quite large, which means it needs tree pruning. To give a solution to the discussed issue in this framework work, use the random forest to detect electricity energy theft. The RF has majorly two major processes: testing and training a model of the data set. The data set has two dimensions: row (Information value, records) and column (variable, attributes, predictors) for making a decision tree. The random forest has a forest of decision trees, and it gives majority voting to RF.

The following steps are carried out to grow RF:

- Definition of training a model and testing a model of the datasets
- Supervised training of random forest on the training of the sets.
- Classification of test a model of the data using the saved forest.
- Error found

The RF input data is called the predictors or response, and the output is called the target value. The training of the RF has two steps

(i) Boostrap
(ii) Aggregating.

Different subsets of the training model of the data are collected with replacement by ~2/3 to train each tree. The error and variable importance are calculated using the remaining training data (OOB). Each node is split with randomly selected subset variables. Two parameters are used to separate the tree; the first one is ntree, which is the number of trees to grow, called an ntree parameter. The other one is mtry, the number of variables available for splitting at each tree node, which is called amtry parameter. A smaller subset results in less correlation (reducing error) but also decreases predictive power (increasing error)

Advantages of RF:

- There is no problem with tree pruning.
- RF algorithm can handle classification and regression (Liaw and Wiener2002) types of issues.
- It gives higher accuracy than other algorithms.
- RF modules are robust to “noise” in the training data set.
- This method is fast and effective in working with large datasets.
- This model does not overfit easily.
- RF is capable of handling large missing values.

4.2. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be applied to both classification and regression problems, it is primarily used for classification. In this approach, each data point is plotted in an n-dimensional space, where n represents the number of selected features, and each feature's value corresponds to a specific coordinate. Classification is then performed by identifying the optimal hyperplane that best separates the data points. The functional block is given figure 2,

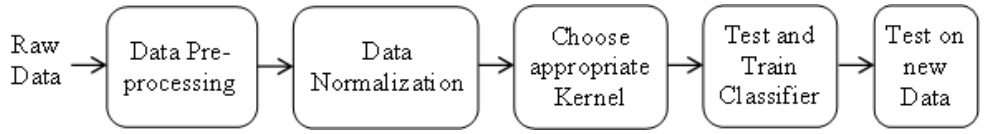

Figure 2.Functional Block Diagram of SVM

4.2.1. Data pre-processing

During data pre-processing, raw data is collected from various sensors and smart meters within the power distribution system and then processed for use in an SVM classifier. All collected attributes must be converted into a numerical format.

4.2.2. Data Normalization

This is most important step to classification in SVM. There are two type of margin to normalization. These are (i) Functional margin and (ii) Geometric margin. Training example is determined $(x^{(i)}, y^{(i)})$, and the functional margin isfound (w,b) with respect to the training example shown in equation (1),

$$\gamma^{(i)} = y^{(i)}(w^T x + b) \quad (1)$$

If $y^{(i)} = 1,$, large functional margin (i.e., it is an accurate and confidenprediction), then it need $w^T x + b$ to be a large positive number. If $y^{(i)} = -1,$ large functional margin, then large negative number $w^T x + b$. furthermore, if $y^{(i)}\ w^T x + b > 0,$ then the prediction is correct. Therefore, a large functional margin represents anaccurate prediction.

For a linear classifier tack the value {-1, 1} and it may be senseable to impose a normalization condition, such as that $||w||_2 = 1$; i.e., it could replace (w,b) with $(w/||w||_2, b/||w||_2)$, and instead consider the functional margin of $(w/||w||_2, b/||w||_2)$.

Given the training a model of the data set $S = \{(x^{(i)}, y^{(i)}); i = 1,....,m\}$ defines the function margin of (w,b) with respect to S as the smallest of the functional margins of the individual training a model. γ can be written as:

$$\gamma = \min_{i=1,....,m} \gamma^{(i)}$$

The definition of the geometric margin (w,b) with respect to training a model $(x^{(i)}, y^{(i)})$ to be equation (2)

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right) \quad (2)$$

Form the above equation, $||w||$ = 1, the margin are equals. It is provided two different notions of margin. In this proposed method, all the variables are scaled down in the range from -1 to 1.

4.2.3. Choose Appropriate Kernel

The kernel function crucial role to the working of SVM. There are two types of kernel functions available in SVM: linear and non-linear kernel functions. This paper uses the non-linear kernel for classification.

Gaussian radial basis function (RBF) is mainly used in non-linear kernel for classification. This kernel is based on the Gaussian function, which is written as equation (3),

$$K(x_i, x_j) = \exp\left(-\frac{||x_i, x_j||^2}{2\sigma^2} \right) (3)$$

Where x_i is the support vector and x_j data value of the currentposition. Sometimes the γ is used instead of $1/2\sigma^2$, that shown by following equation (4),

$$K(x_i, x_j) = \exp(-\gamma ||x_i, x_j||^2) \text{ for } \gamma > 0 \quad (4)$$

4.2.4. Test and Train Classifier

After selecting the kernel function, the training data sets are used to train the SVM classifier. Once the SVM classifier is trained, it can classify the new test value based on the trained classification model. The training is

happening in a ratio manner. The software processing and simulation results are shown in the upcoming section.

5. Simulation and Result for DT and RF based SVM

5.1 Steps Involving finding Confusion Matrix from SVM

Step 1: Import the data sheet and get y-fit values from the DT and RF.
Step 2: The standard deviation of the target is calculated.
The deviation value was taken as 0.5 or 0.4 from the ECV value.
Step 3: The class label is found from step 2. The classes are namely "0" and "1".
Step 4: Fed the input attributes and get the output attributes from the DT and RF to the SVM classifier.
Step 5: Confusion matrix calculate and produce the False Positive and True Negative rate.
Simulation was done using MATLAB software. Five attributes are given to inputs of the decision tree and random forest. The DT and RF select the predictor importance that is figure 3 and figure 4,

PerdictorImportance Estimates

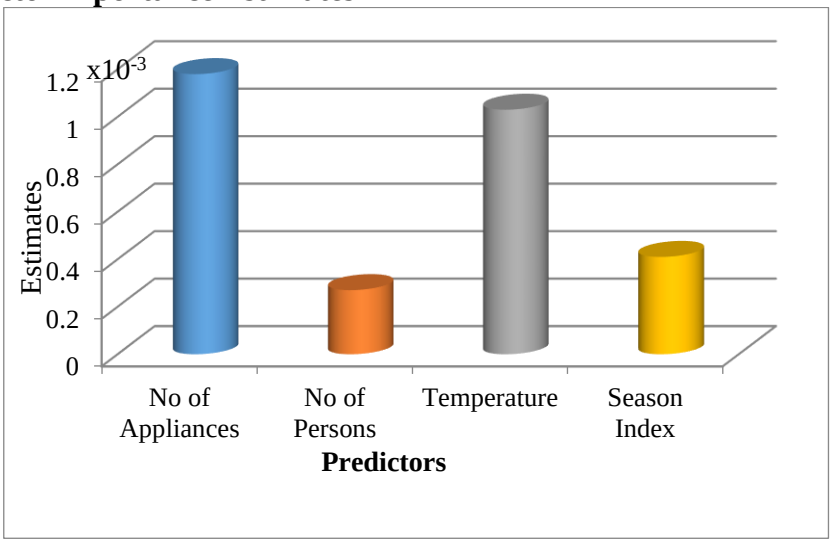

Figure 3. Variable Importance of Decision Tree

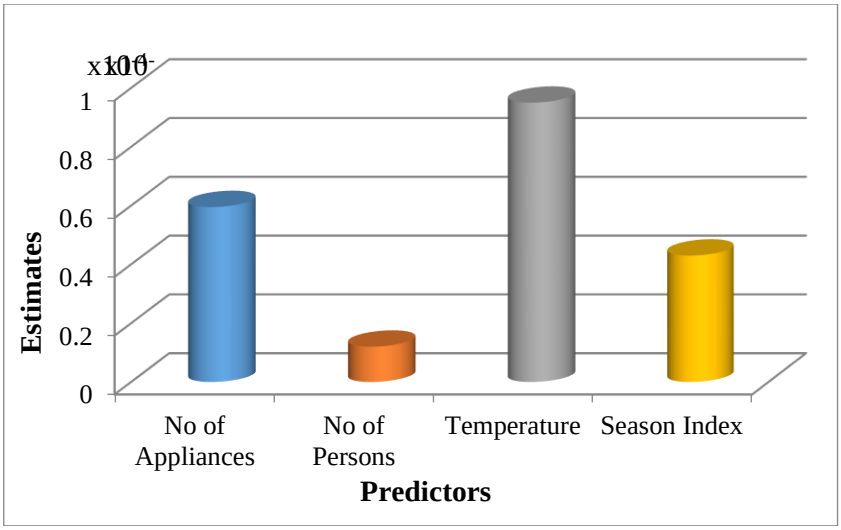

Figure 4. Variable Importance of Random Forest

The decision tree gives importance to several appliances, and the random forest gives importance to temperature. Then, these are working concerning this predictor's importance.

5.1.1. Detection of malicious consumers

The normal consumers and malicious consumers are found in the SVM classifier figure 5 and figure 6,

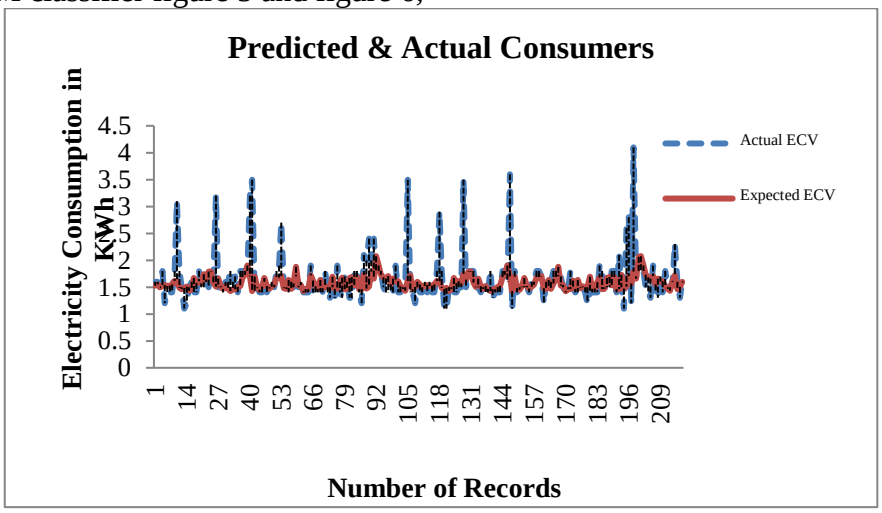

Figure 5. Normal consumers

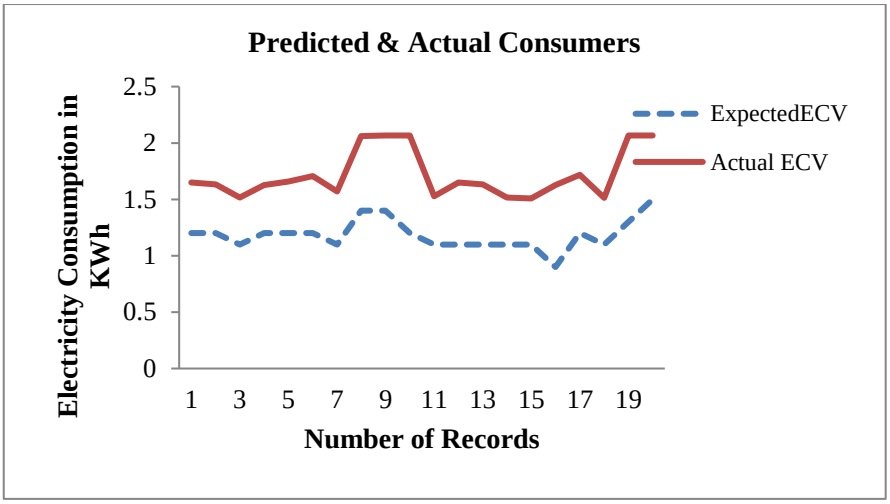

Figure 6. Malicious consumers

From the figure 5 and 6 shows the electricity consumption between actual and predictor users. The graph indicates the fraudulent users. The analysis highlights the crucial role of electricity prediction in identifying malicious consumers.

II. CLASSIFICATION ACCURACY CALCULATION

Decision treebased SVM, Random Forestbased SVM,Extereme Gradient Boosting (XGBoost),and Categerical Boosting (CatBoost)are utilized to evaluate the sensitivity and selectivity. The consumers are classified as either normal or fraudulent by using SVM. Selectivity and sensitivity can be used to calculate the accuracy of the SVM classification performance.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{Positive}}$$

$$\text{Selectivity} = \frac{\text{True Negative}}{\text{Negative}}$$

Accuracy is a function of selectivity and sensitivity and it is calculated in equation (5),

$$\text{Accuracy}=\text{Sensitivity}\left(\frac{\text{Positive}}{\text{Positive}+\text{Negative}}\right)+\text{Selectivity}\left(\frac{\text{Negative}}{\text{Positive}+\text{Negative}}\right) \quad (5)$$

According to the Decision Tree based SVM, the considered dataset consists of 215 Normal consumers and 23 Fraudulent consumers. So that positive is 23 and negative 215. The accuracy can be calculated as following equation (6)

$$\text{Accuracy} = 0.347\left(\frac{23}{238}\right) + 0.9395\left(\frac{215}{238}\right) \quad (6)$$

Accuracy = 94.52%

According to the Random Forest based SVM, the considered dataset consists of 217 Normal consumers and 21 Fraudulent consumers. So that positive is 21 and negative 217. The accuracy can be calculated as following equation (7)

$$Accuracy = 0.7619\left(\frac{21}{238}\right) + 0.8525\left(\frac{217}{238}\right) \quad (7)$$

Accuracy = 84.4%

From the equation (6) and (7) is clear that Decision Tree based SVM has more accuracy rate than Random Forest based SVM shown in table 1. Random Forest has bunch of Decision Tree so it has less generalization capability that's why we move on Extreme learning machine.

Table1. SVM Performance Table

Ratio	Training	Testing	Sensitivity	Specificity	RF+SVM Accuracy Rate
40%	96	142	0.415	0.9384	89
45%	107	131	0.415	0.915	87
50%	119	119	0.761	0.8525	84
55%	131	107	0.909	0.8144	82
60%	143	95	0.909	0.8235	83
65%	155	83	0.909	0.8219	82
70%	166	72	0.5	0.8253	80
75%	178	60	0.6	0.8235	80
80%	190	48	0.444	0.75	72
85%	202	36	0.5	0.936	89
90%	214	24	0.333	1	94
95%	226	12	1	0.9166	92

6. **ELM classifier**

Extreme learning machine are feed forward neural network that are capable of performing Classification, Regression, Clustering, Sparse approximation, compression and feature learning. It can be structured as a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes (the weights are not connecting inputs to hidden nodes) do not require tuning. These hidden nodes can be randomly assigned and remain unchanged (i.e. they are random projections but with nonlinear transforms), or inherited from previous models without modification.

In most cases, the output weights of hidden nodes are learned in a single step, effectively reducing the problem to a linear model. The name "Extreme Learning Machine" (ELM) was introduced to such models by its main inventor, Guang-Bin Huang.According to the developer, ELM produced good generalization performance and learned thousands of times faster than networks trained using backpropagation. In the literature, ELM can outperform support vector machines (SVM), which provide reliable solutions in both classification and regression applications.

The two main Machine Learning Techniques are given bellow,

- Singlehiddenlayer feed forward networks
- Multi hidden layer feed forward networks

6.1. ELM in "Generalized" Single-hidden-layer feed forward networks

It can be extended to "generalized" single-hidden-layer feed-forward networks and mathematical series/expansions is given in the below equation (8),

$$f_L(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x) \quad (8)$$

The basic ELM is designed for generalized single-layer feed-forward networks.Unlike the fully connected networks in those earlier works, ELM introduces three levels of randomness,

1. In a fully connected network, hidden node parameters are randomly generated.
2. The hidden node connections can be randomly generated; all input nodes does not connected to a particular hidden layer. Possible input nodes in within a local field are connected to any one hidden layer

A hidden node also has a sub-network, which can be formed by several nodes. The sub-network can be formed naturally with local receptive and

pooling functions, enabling the learning of local features. From this, different sections of a single ELM can contain multi-hidden layers.

6.2. Multi-hidden-layer feed forward networks

Hidden nodes of broad types of multi-hidden-layer networks, the hidden nodes do not require tuning. Multi-layer ELM concepts were provided in ELM theories in 2007. In essence:

1. Need not tune the hidden nodes, twofold:

(a) Hidden nodes can be randomly generated.

(b) Hidden nodes can be generated non-randomly, they do not require tuning. For example, the next hidden layer consist of a linear combination or nonlinear transformation, which can be generated randomly in the previous layer. In this study, a some nodes are generatedrandomly, and a some are not, but none are tuned.

2. Each ELM can capableof handling with compression, feature learning, clustering, regression, and classification. Finally, homogeneous hierarchical blocks of ELM can be constructed. For example, one ELM block can perform the feature learning, and another act as the classifier. The ELM has two hidden layers. It cannot be generatedrandomly, but it can be ordered. The hidden nodes of each layer do not require tuning.

3. ELM separates which play feature learning or clustering roles can also be used to integrate different learning models. ELM trains some layers for an entire network, while other models train some.

5. SIMULATION AND RESULT

In the Extreme learning Machine algorithm has two type of accuracy testing methods. These are given below,

1. Hold-Out Partition
2. K-Fold Cross Validation

Among the two type of accuracy testing, in this paper uses the hold out partition method. This performance with 20 hidden nodes is given in figure10.

I. Hold-Out Partition Testing Accuracy

Figure 7 compares ELM predicted accuracy rate RF+ELM and ELM with XGboost+ELMaccuracy rate. From the graph, it is clear that XGboostgives a high-efficiency rate. It identifies the fraudulent consumer in a very accurate manner. There are different ratio ranges taken to the training of the ELM classifier shown in table 2. Among all ratios, 65% and 80% of

the training ratio has the very accurate prediction of the Extreme learning machine classifier with 20 hidden nodes.

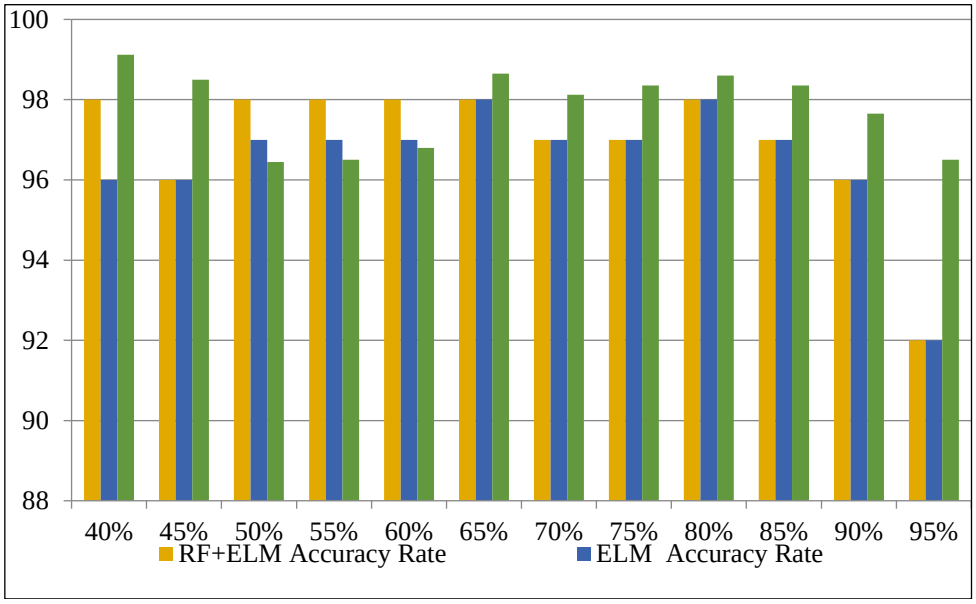

Figure 7. Comparison Graph for ELM+RF, ELM and XGboost+ELM

Table 2: Performance Comparison of ELM+RF, ELM and XGBOOST+ELMwith 20 Hidden Nodes

Ratio	Training	Testing	RF+ELM Accuracy Rate	ELM Accuracy Rate	XGBoost+ELM Accuracy Rate
40%	96	124	98	96	99.12
45%	107	131	96	96	98.5
50%	119	119	98	97	96.45
55%	131	107	98	97	96.5
60%	142	95	98	97	96.8
65%	155	83	98	98	98.65
70%	167	71	97	97	98.12
75%	178	60	97	97	98.35
80%	190	48	98	98	98.6
85%	202	36	97	97	98.35
90%	214	24	96	96	97.65
95%	226	12	92	92	96.5

II. Performance Comparison Chart for RF+ELM, RF+SVM and RF+CatboostWith 20 Hidden NodesApproach

Figure 8 compares SVM with the RF-predicted accuracy rate, ELM with the RF accuracy rate and Catboost with the RF accuracy rate. From the graph, it is clear that Catboost with RF gives a high-efficiency rate. It identifies the fraudulent consumer in a very accurate manner. There are different ratio ranges are taken to train the ELM classifier with 20 hidden nodes. From Figure 10, this paper used the Hold-Out partition for training the data set, which gives a less accurate value when compared to the Cross-validation or K-Fold validation. The performance comparison of the SVM+RF, ELM+RFandCatBoost+RFapproach with 20 hidden nodes tabulation is shown in Table 3.

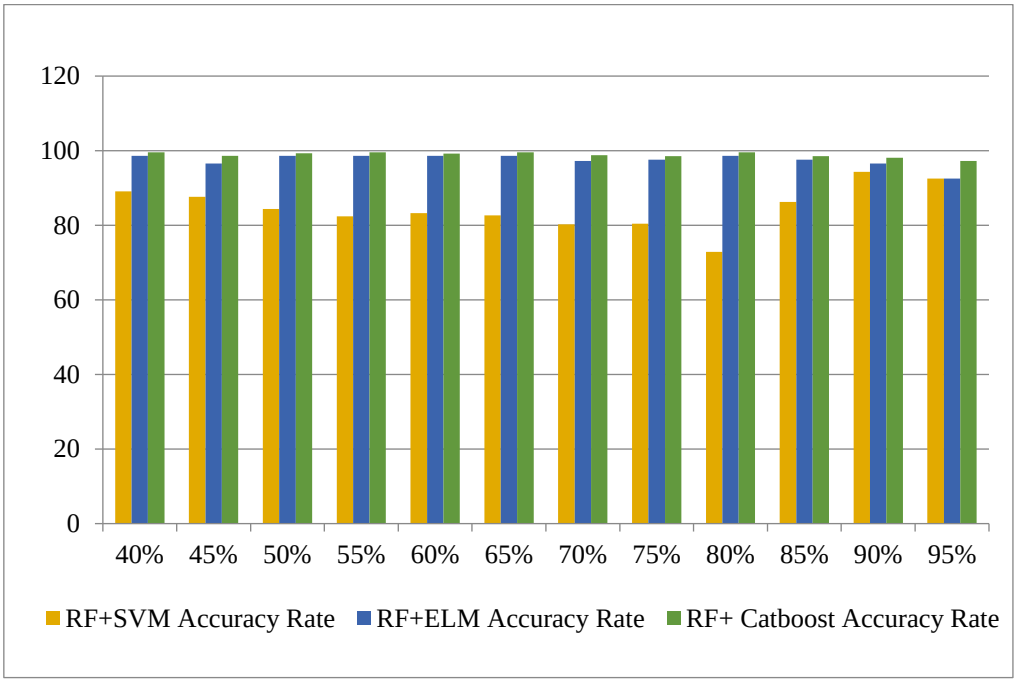

Figure 8. Comparison Graph for ELM+RF, SVM+RF and Catboost+RF

Table3: Performance Comparison of RF+SVM and RF+ELM with 20 Hidden Nodes

Ratio	Training	Testing	RF+SVM Accuracy Rate	RF+ELM Accuracy Rate	RF+ Catboost Accuracy Rate
40%	96	124	89.12	98.65	99.52
45%	107	131	87.65	96.52	98.65
50%	119	119	84.35	98.62	99.26
55%	131	107	82.35	98.64	99.56
60%	142	95	83.24	98.62	99.24
65%	155	83	82.64	98.64	99.56
70%	167	71	80.25	97.23	98.78
75%	178	60	80.45	97.58	98.54
80%	190	48	72.86	98.62	99.56
85%	202	36	86.24	97.58	98.54
90%	214	24	94.35	96.52	98.12
95%	226	12	92.52	92.5	97.26

6. CONCLUSION

A Random Forest is a collective ensemble of decision trees, making it an effective model for achieving high accuracy in data prediction. This model consists of numerous decision trees, often numbering in the tens or hundreds. In this analysis, various algorithms were compared and evaluated using MATLAB, based on a dataset that included 238 samples of electricity consumption from various states in the USA. The prediction accuracies recorded were as follows: RF+ELM achieved 98%, ELM alone also at 98%, and XGBoost+ELM reached an impressive 99.12%. It is noted that the Random Forest-based Extreme Learning Machine (RF-based ELM) requires more execution time compared to the XGBoost-based ELM. Furthermore, the XGBoost-based ELM provided more accurate predictions than both the RF-based ELM and the ELM. According to this study, the Random Forest model (with 100 trees) when used with Support Vector Machine (SVM) yielded lower classification accuracy compared to the RF-based ELM algorithm. Additionally, the RF-based SVM demonstrated less generalization capability compared to the RF-based CatBoost. To address this limitation, an advanced classification method in ELM was implemented

to predict malicious consumers using a Hold-Out partition. The results indicated that RF combined with CatBoost achieved an accuracy rate of 99.56%. Overall, RF+CatBoost exhibited superior classification performance compared to XGBoost-based ELM, RF-based ELM, Decision Tree (DT)-based SVM, and RF-based SVM.

References:

[1] Cabral, J. (2004), Fraud detection in electrical energy consumers using rough sets, in Proceedings IEEE International Conference System, Man Cybernet, 4, 3625–3629.

[2] Costa, B.C., Alberto, B.L., Portela, A.M., Maduro, W., Eler, E.O. (2013), Fraud detection in electric power distribution networks using an ANN based knowledge-discovery process, International Journal Artificial Intelligence Applications, 4(6), 17–23.

[3] Czechowski, R., Kosek, A.M. (2016), The most frequent energy theft techniques and hazards in present power energy consumption, *Joint Workshop on Cyber- Physical Security and Resilience in Smart Grids (CPSR-SG)*, Vienna, Austria, 1-7,

[4] Depuru, S.S.S.R., Wang, L., Devabhaktuni, V. (2011), Support vector machine based data classification for detection of electricity theft," in Proceedings IEEE/PES Power System Conference Expo. (PSCE), 1–8.

[5] Jiang, R., Tagaris, H., Lachsz, A., Jeffrey, M. (2002), Wavelet based feature extraction and multiple classifiers for electricity fraud detection, in IEEE/PES Transmission Distribution Conference Exhibition Asia Pacific, 3, 2251–2256.

[6] Jokar, P., Arianpoo, N., Leung, V.C.M. (2016), Electricity theft detection in AMI using customers' consumption patterns," IEEE Transactions Smart Grid, 7(1), 216–226.

[7] Kantardzic, M. (2012), Decision trees and decision rules, in Data Mining: Concepts, Models, Methods, and Algorithms, 2nd edition, pp. 169–198.

[8] Keskes, H., Braham, A. (2015), Recursive undecimated wavelet packet transform and DAG SVM for induction motor diagnosis, IEEE Transactions IndustrialInformatics, 11(5), 1059–1066.

[9] Liaw, A., Wiener, M. (2002), Classification and regression by random forest, R News, 2-3.

[10] Nabil, M., Ismail, M., Mahmoud, M., Shahin, M., Qaraqe, K., Serpedin, E. (2019), Deep learning-based detection of electricity theft

cyber-attacks in smart grid AMI networks, in: Deep Learning Applications for Cyber Security, Springer, 73–102.

[11] Nagi, J., Mohammad, A.M., Yap, K.S., Tiong, S.K., Ahmed, S.K. (2008), Non-technical loss analysis for detection of electricity theft using support vector machines, in Proceedings 2nd IEEE International Power Energy Conference(PECon.), 907–912.

[12] Nagi, J., Yap, K.S., Tiong, S.K., Ahmed, S., Nagi, F. (2011), Improving SVM based nontechnical loss detection in power utility using the fuzzy inference system," IEEE Transactions Power Delivery, 26(2), 1284–1285.

[13] Newswire PR., (2014), World Loses $89.3 Billion to Electricity Theft Annually, $58.7 Billion in Emerging Markets. [Online].

[14] Tsai, C.W., Lai, C.F., Chao, H.C., Vasilakos, A.V. (2015), *Big data analytics: A survey," Journal Big Data*, 2(1), 1–32.

[15] Karthikeyan, R., Kiruthika, B. &Durgadevi, V. Detection of cardiac arrhythmias from ECG signals using FBSE and Jaya optimized ensemble random subspace K-nearest neighbor algorithm. *Biomed. Signal Process. Control* 76, 103654.(2022).

[16] Durgadevi, V., Ganeshkumar, P. &Karthikeyan, R. A cross-layer trust evaluation protocol for secured routing in communication network of smart grid. *IEEE J. Sel. Areas Commun.* 38(1), 193–204. (2020).