Big Data Analytics for Electricity Theft Detection in Smart Grids

Inam Ullah Khan[1,*], Nadeem Javaid[2], C. James Taylor[1], Kelum A. A. Gamage[3] and Xiandong Ma[1]
[1]*Engineering Department, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK*
[2]*Department of Computer Science, COMSATS University Islamabad, Islamabad, 44000, Pakistan*
[3]*James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK*
*Correspondence: inamullah@ieee.org

***Abstract*—In Smart Grids (SG), Electricity Theft Detection (ETD) is of great importance because it makes the SM cost-efficient. Existing methods for ETD cannot efficiently handle data imbalance, missing values, variance, and non-linear data problems in the smart meter data. Therefore, an effective integrated strategy is required to address underlying issues and accurately detect electricity theft using big data. In this work, a simple yet effective approach is proposed by integrating two different modules, such as data pre-processing and classification, in a single framework. The first module involves data imputation, outliers handling, standardization, and class balancing steps to generate quality data for classifier training. The second module classifies honest and dishonest users with a Support Vector Machine (SVM) classifier. To improve the classifier's learning trend and accuracy, a Bayesian optimization algorithm is used to tune SVM's hyperparameters. Simulation results confirm that the proposed framework for ETD significantly outperforms previous machine learning approaches such as random forest, logistic regression and SVM in terms of accuracy.**

***Index Terms*—Big data, Electricity theft detection, Feature engineering, Data classification, Smart grid.**

I. INTRODUCTION

In the energy sector, power systems are electrical grids that provide electricity to homes and industries within a large geographical area. Electricity is an expensive commodity and needs to be carefully and efficiently utilized. From generation to distribution, a power network encounters two types of losses: Technical Losses (TL) and Non-Technical Losses (NTL). TL occur due to losses in cables, transmission lines and transformers during energy transfer and cannot be prevented within a distributed network. In contrast, NTL occur when there is an illegal usage of electricity with an aim to escape from utility charges. Meter tempering and bypassing, tapping on secondary voltages, and synchronously switching power circuits are one of the few examples of NTL in power network. The primary cause of NTL is electricity theft, which gives rise to approximately \$89.3 billion of revenue loss annually [1].

Electricity theft is one of the SG's leading drivers that often causes a wide range of anomalies at planning and distribution levels. To counter this, the role of Electricity Theft Detection (ETD) has become increasingly important in the SG. The advanced methods for ETD based on big data is always an essential and challenging issue. The primary purpose of ETD is to minimize NTL in the power system and balance the energy supply-demand gap. An accurate and stable ETD method brings extraordinary energy management compliance and develops a win-win situation for the generation and consumption side stakeholders [2].

Accurate ETD methods are of great importance for SGs but many intricate factors in big data would intensify the difficulty of using these methods for ETD. The big data phenomenon is dynamic and complex that involves distinctive aspects of the time series data where the variation trends over time are non-linear. Accurate ETD is essential, but it is challenging to increase scalability, robustness, and accuracy due to the widespread non-linear data. SMs continuously monitor the associated factors such as time and consumption pattern of a consumer's consumption in real-time. As a result, the amount of data available for ETD is significantly big and hence challenging to handle, especially for ETD [4].

In recent years, researchers have extensively explored different types of techniques for theft detection using big data from SG. For instance, Maddilina *et al*. [5] used Support Vector Machine (SVM) and a well-known boosting classifier, named XGBoost for identifying anomalies in usage pattern of consumers. With smart meter's (SM) data analysis, consumers are ranked based on their load profiles. Afterwards, essential features are extracted from auxiliary data. The SVM utilized the empirical risk minimization principle to improve the training process and enhance classification performance with the boosting algorithm. In the proposed strategy, the authors did not take into account the data preparation steps. However, like any other Machine Learning (ML) algorithms, SVM's performance is further improved when refined data are fed into the classifier for training.

Authors in [6] proposed a hybrid technique combining Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) methods. The MLP is used for auxiliary data, whereas time-sequenced electricity data are evaluated with LSTM. The authors achieved good prediction results; however, model's classification performance could be further improved if the class imbalance problem was solved during the data preparation stage. Besides, the model's performance is relatively high on fewer data training in terms of False Positive Rate (FPR). However, when the data input for model training is high, its performance degraded to 54.5% Precision-Recall Area Under the Curve (PR-AUC).

In [7], Shuan *et al*. used a well-known deep neural network model named as Convolutional Neural Network (CNN) for the detection of electricity theft accurately. Nevertheless, a major drawback of model's generalization arises when classification output is taken from a fully connected layer CNN. To address this problem, authors in [8] applied a Random Forest (RF) technique to get final output of the classification task. In the proposed work, the imbalanced class problem is solved with Synthetic Minority Oversampling Technique (SMOTE). Although the proposed method achieved better generalization capabilities however, SMOTE's synthetic data generation creates overfitting problem. When a model overfits, it means that model is better for training but not for testing/classification.

Most of the recent works are based on selection or classification approaches where ML and Artificial Neural Networks (ANNs) have shown improved performance for ETD strategies. However, both methods have limited abilities. For example, ML models have low detection rate and high FPR, failing to handle imbalanced class and overfitting problems. Similarly, ANN models have limited generalization capabilities, sensitivity to erroneous values, limited control over convergence/stability, and limited abilities to deal with the uncertainty. Besides, the learning-based models do not consider the big data characteristics, and their performance evaluation criterion are based only on price/load data, which is not large [8], [9]. With the consideration of big data characteristics, the classification accuracy of a model needs to be further improved.

This work investigates ETD issues like binary classification task using big data from the SG. Inspired by [6], Fig.1 shows the framework of the proposed system that is based on two modules: data preparation and classification. For this purpose, an SVM underpinned framework is proposed to solve the challenging binary classification task. To divide the given data into correct classes (honest and fraudulent), SVM is used where it finds an optimal boundary among different data points. Although SVM is an appropriate approach for classification tasks, the following challenges need to be addressed to further enhance classification accuracy.

- ***Computational Overhead***: In the work, Hu et al. [10] investigated that SVM's performance is adversely affected by unreliable information due to which model's computational overhead increases. In ETD problem, extraneous and redundant features increase computational overhead and make the classifier's training process difficult, which in return decrease the classification accuracy of the model.
- ***Hard to Tune Hyperparameters***: In SVM, three super parameters, namely kernel parameter, intensive loss function, and cost penalty control the classifier's performance. To obtain optimum results, tuning these super parameters is a relatively tricky task for higher accuracy and better efficiency. Two well-known methods, namely cross-validation and gradient descent are used to adjust SVM's super parameters [6]. However, both methods make the converging process hard and increase computational complexity.

To address above mentioned challenges, an integrated framework for ETD is proposed as shown in Fig.1. First, data preparation module performs interpolation, normalization, and balancing tasks. Precisely, data interpolation fills the missing values and brought consistency in the data set. Afterward, the data normalization (puts the values between 0–1) is performed to bring uniformity. Once the data preparation step is completed, the processed data are sent to the classifier. In the proposed framework, we chose SVM because it performs well on the classification tasks [6]. SVM is very sensitive to the value of the hyperparameter. For this purpose, a Bayesian Optimization Algorithm (BOA), is employed to detect electricity theft accurately. The main contributions of the research work to achieve higher accuracy are listed below:

1) We propose an integrated framework that is based on two modules, namely pre-processing and classification. Due to the cascading effect, SM's theft data is efficiently handled and analysed.
2) To achieve this framework, data preparation steps are proposed to accomoate data imputation, outliers handling, standardization and handling imbalanced class for refining data.
3) We also design a BSVM algorithm to tune the hyperparameters of SVM. The BSVM has higher accuracy and computational efficiency than the basic SVM and recent ML techniques in the proposed area.
4) For performance evaluation, extensive simulations on real-world data traces of grid's workload have been considered. The numerical results show that the proposed model achieve better performance statistics than benchmark methods.

The rest of the article is organised as follows. Section II describes the data preparation module of the proposed ETD framework. Similarly, Section III demonstrates the SVM classifier and its enhancement with BOA. The proposed framework for ETD is verified with multiple scenarios in Section IV. Finally, Section V concludes this work.

II. DATA PREPARATIONS

The preliminary analysis of data is a mandatory step in high dynamic time series analysis, which includes imputation, data standardization and handling imbalanced class data. The details of these methods are given below.

A. *Handling Missing Values*

The electricity consumption record of consumers is usually composed of incomplete information or missing values. The reasons behind the issue may be the failure of hardware and corruption of data. In high time-series data, the missing values can not be dropped; therefore, the imputation is performed synthetically to fill these values. In most cases, the filling of missing values is performed through averaging. In this paper,

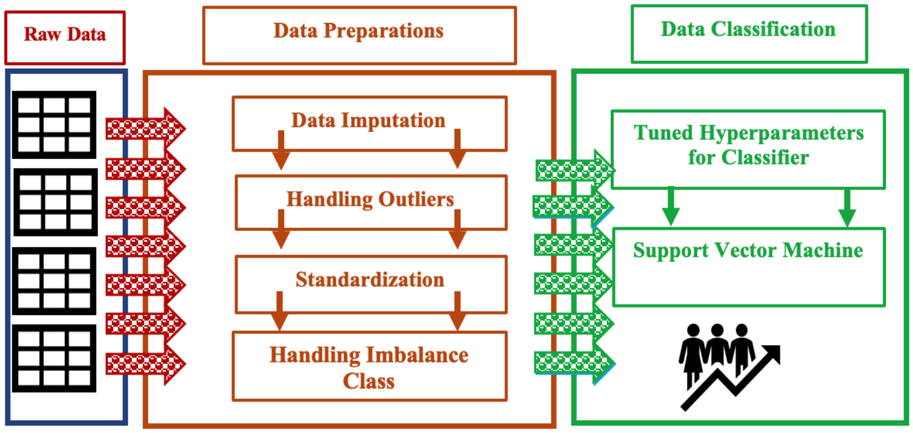

Fig. 1: Proposed Framework for Electricity Theft Detection

the missing values are recovered through interpolation method [11] calculated using Eq. (1):

$$f(x_i) = \begin{cases} \left(\frac{x_{i-1}+x_{i+1}}{2}\right), & if\ x_i \in NaN, x_{i\pm1} \notin NaN, \\ x_i, & otherwise, \end{cases} \tag{1}$$

where x_i is the recorded or missed (null) observation in the dataset. The null value is represented as NaN. If x_i is null then it is filled according to Eq. (1).

B. Handling Outliers

In the State Grid Corporation of China (SGCC) dataset, there are numerous outliers due to which data is skewed; hence training process becomes complex. These outliers must be identified and removed to avoid overfitting and time complexity problems while preparing data for training. The "three-sigma rule of thumb" proposed in [12] is utilized for detecting and recovering the outliers. Mathematically, it is expressed in Eq. (2):

$$x_{out} = \begin{cases} X, & if\ x_i > X, \\ x_i, & otherwise, \end{cases} \tag{2}$$

where X denotes $Avg(x_i + 2\sigma(x_i))$.

C. Data Standardization

The data standardization is performed by min-max normalization method [13] as given in Eq. (3):

$$x_{new} = \frac{x_i - min(x)}{max(x) - min(x)}. \tag{3}$$

D. Handling Imbalanced Data

One of the critical problems in SM's data is the majority class's domination (honest consumers) compared to the minority class (thieves). In such a scenario, the distribution is not normal and skewed towards the majority class because of an unclear decision boundary [14]. The classifier would become biased, may not learn critical features, and tend to become overfit. Traditional methods to deal with such issues are random under-sampling and random oversampling. However, these methods are not preferred because of specific problems, namely computational overhead, under and overfitting. Considering the nature of these problems, we opt for a relatively new class balancing approach that combines the properties of SMOTE and Tomek Links techniques; we name the new technique as STLU. This technique has not been utilized in ETD strategies for class balancing to the best of our knowledge. In STLU, SMOTE is an oversampling technique, which synthesizes new plausible examples in the majority classes. In contrast, Tomek Links identifies different nearest neighbors' classes in a dataset and removes majority class samples to achieve a suitable balance between both classes of honest and fraudulent customers.

III. CLASSIFICATION

This module describes the final classification task via the processed data. We chose SVM because it is one of the most adopted, robust and efficient machine learning methods to provide a higher classification accuracy. We define a matrix of electricity consumption data below in Eq. (4):

$$\text{X} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1n} \\ x_{21} & x_{22} & ... & x_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & ... & x_{mn} \end{bmatrix}, \tag{4}$$

where, rows and columns represent the time stamps and the feature index of the data, respectively. The classified component x_{mn} is $n-th$ component of m electricity consumption value that is to be classified. The diagonal matrix can also be formulated as shown in Eq. (5):

$$\text{X} = \begin{bmatrix} \overrightarrow{t_1} \\ \overrightarrow{t_2} \\ . \\ . \\ . \\ \overrightarrow{t_m} \end{bmatrix}, \tag{5}$$

where,

$$\overrightarrow{t_k} = [x_{k1}, x_{k2}, ...x_{kn}]\ k \in [1, m]. \tag{6}$$

For a given training set $\{(x_i, y_i)\}_{i=1}^{N}$ (x_i and y_i represent samples and target classes, respectively) with binary output $y_i = \pm 1$, the classification problem is investigated below using Eqs. (7–9) [6]:

$$f(x, c) = \sum_{i=1}^{N} c_i \lambda_i(x) + b, \tag{7}$$

where b depends on data distribution and $c_i^{\infty}(i = 1, 2, ...)$ are classifier parameters to be adjusted. Eq. 7 defines a hyperplane in N dimensional space. The regularized risk function is calculated using following Eq. (8):

$$w(c) = \frac{\sum_{i=1}^{N} |y_i - f(xi, c)|_{\in} + \mu c^2}{N}, \tag{8}$$

where $\in$ represents parameter for intensive loss function, μ is a constant, and y_i is the actual class. To obtain optimal values of classifier's parameters c, reguarized risk function minimzation is required for which the robust error function is calculated below in Eq. (9):

$$x = \begin{cases} 0, & if \ |y_i - f(xi, c)| <\in, \\ |y_i - f(xi, c)|, & otherwise, \end{cases} \tag{9}$$

A. *Optimal Classification with Bayesian Optimization Algorithm*

In SVM, we aim to minimize the regularized risk function. The regularized risk function has a strong relationship with three super parameters: the type of SVM kernel parameter (σ), cost penalty (c) and the intensive loss function $(\in)$. The need for parameter optimization is undeniable, and computational efficiency is achievable if optimal values for these super parameters are chosen. In the past, various methods, such as Cross-Validation (CV), Grid Search (GS), Gradient Descent (GD), and heuristic algorithms are proposed to adjust super parameters of SVM. However, these methods may cause a problematic convergence process due to high computational overhead. Furthermore, CV, GS, and RS methods are comparatively ineffective because of random search values are not updated on the previous best to choose the next hyperparameters. For this purpose, a reliable BOA is chosen to tune SVM's super parameters.

The BOA is chosen for parameter optimization because it is more directed, faster, and predictable according to the posterior probability. The BOA improves the hyperparameter selection by making use of earlier experiments. First, it constructs a probabilistic model of the function with super parameters and evaluates it on the validation test. With multiple iterations, the BOA gathers relevant information about the optimal locations with a perfect balance between exploration (super parameters likely to give uncertain outcome) and exploitation (expected optimum parameters). It provides better results in fewer iterations as compared to the RS and GS algorithms. It starts with taking a history of super parameters settings $\lambda_n = \lambda_1, \lambda_2, \dots \lambda_n$ and respective function evaluation $y_1 = y_1, y_2, \dots y_n$ to acquire a new set of super parameters λ_{n+1}. In next iteration, λ_{n+1} is used as new population for model evaluation to get new function value y_{n+1}. Both function and super parameter values are saved to the history for execution of next iteration λ_{n+1}, y_{n+1}. In this way, the objective function's optimized value is achieved with a history of function evaluation and super parameters values.

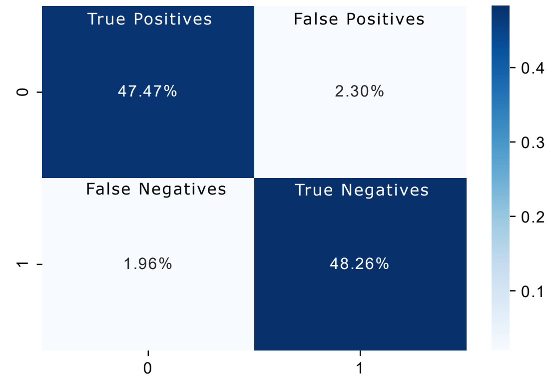

Fig. 2: Confusion Matrix for BSVM

With efficient data pre-processing and enhanced classification methodologies, the proposed framework is capable of performing ETD accurately. In the next section, a detailed analysis of real-world electricity theft data is given.

IV. NUMERICAL ANALYSIS

A. *Simulation Setup*

This section investigates the capabilities of the proposed framework and the simulator is developed with Python according to the system framework devised in Section II. The simulator runs on a platform with MAC i7, 16GB RAM, and a 256 GB hard disk. For this framework, the input data are acquired from the most extensive power providing company in China, i.e., SGCC, from 2014 to 2016. The data contain an electricity consumption profile of over 42372 consumers where 38757 consumers are identified as fair consumers and the remaining 3615 are fraudsters consumers.

B. *Performance Matrix:*

The performance metrics are determined from the confusion matrix (CM), i.e., a matrix that describes different results in classification problems, as shown in Fig. 2 which will be explained late in the paper. In a binary classification problem, the CM has four possible outcomes with two rows and two columns. These are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TN and TP score mean that honest and dishonest consumers are identified accurately by the classifier. Similarly, FP and FN score means that the numbers of honest and dishonest consumers are misclassified. Based on CM results, the following Eqs. (10–13) calculate the performance of any classifier:

$$Precision = \frac{T^+}{T^+ + F^+}, \tag{10}$$

$$Recall = \frac{T^+}{T^+ + F^-}, \tag{11}$$

$$F_1 \ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{12}$$

$$Accuracy = \frac{T^+ + T^-}{T^+T^- + F^+ + F^-}, \tag{13}$$

where, T^+, T^-, F^+ and F^- represent TP, TN, FP and FN, respectively.

C. Simulation Results

*1) **Impact of Handling Imbalance Class**:* In an imbalanced class problem, one class significantly dominates the other class; hence, it results in the suppression of the minority class. Fig. 3 shows the difference between minority and majority classes before handling imbalance class. Clearly, the majority class (red circles) customers are in a much higher ratio, and biased classification is expected. Without addressing the imbalance class issue, the value of AUC is 0.5850, Precision is 0.7021 and Recall is 0.4453. The model fails to provide promising results while calculating Recall, as many fraud instances are misclassified as fair. To solve this problem, we apply STLU, which efficiently balances minority and majority classes, and its impact is shown in Fig. 4. With balanced data, model training and generalization are improved as shown in Table 1.

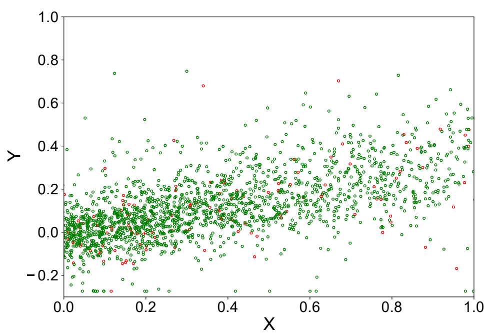

Fig. 3: Imbalnced Data Before Sampling

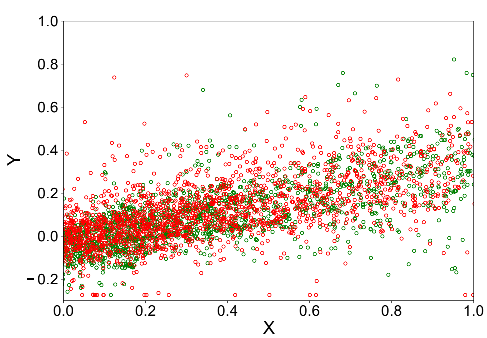

Fig. 4: Balanced Data After Sampling

*2) **The BSVM Performance on Theft Detection**:* SVM is a prevalent technique for classification tasks, and like any other ML methods, its performance is mainly based on hyperparameter values. We tuned SVM's super parameters with BOA, and the objective is to find an optimal hyperplane that distinguishes between different classes.

The Receiver Operating Characteristic (ROC) curve is the best performance metric used for detecting suspects in imbalanced class distribution [13]. It is the graphical representation of T^+ rate and F^+ rate and area under the ROC curve is called Area under the Curve (AUC). It separates the distribution of fraudulent class from fair class and is expressed as follows in Eq. (14):

$$AUC = \frac{\Sigma_i \in SR_i - \frac{1}{2}|S|(|S|+1)}{|S| \times |H|}, \tag{14}$$

where R_i denotes the rank of suspicion number of fraudulent consumers in ascending order, $|S|$ and $|H|$ are the cardinality of suspicious and honest consumers.

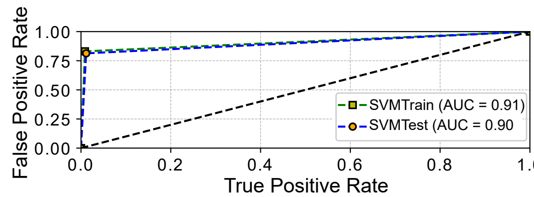

Fig. 5: ROC-AUC Curve of SVM

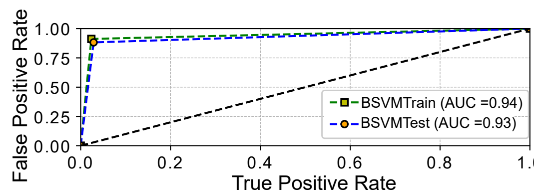

Fig. 6: ROC-AUC Curve of BSVM

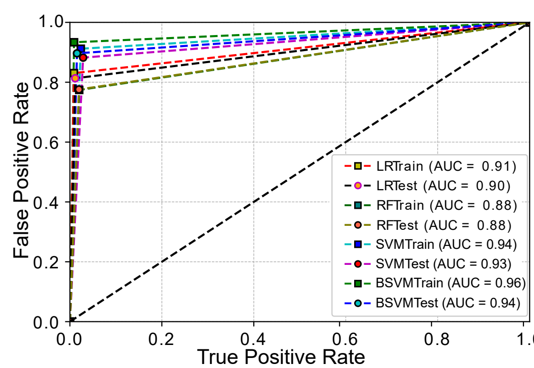

Fig. 7: ROC-AUC-based Performance Comparison

The limit of ROC curve ranges from 0 to 1. An ideal situation arises, when no curve overlaps each other. When AUC approaches 1, it demonstrates the validity of classifier while AUC less than 0.5 shows that the classifier does not have the ability to discriminate among classes. Figs. 5 and 6 show ROC-AUC curves of SVM and BSVM. The AUC of BSVM has been significantly improved both for training and testing. The BOA optimizes SVM's super parameters jointly. Therefore, the BSVM's performance is better both in training and testing. Simultaneously, the AUC of SVM for training and testing are 0.91 and 0.90, whereas, for BSVM, the values are 0.94 and 0.93, respectively. This demonstrates that the acquired results are improved if BOA is used to find the hyperparameters' values of the SVM classifier

TABLE I: Comparision among BSVM and other Benchmark Schemes

Methods	Training Ration 60%					Training Ratio 80%				
	Precision	Recall	F1	Accuracy	AUC	Precision	Recall	F1	Accuracy	AUC
LR	0.713	0.710	0.688	0.700	0.700	0.770	0.725	0.725	0.770	0.720
RF	0.688	0.677	0.687	0.687	0.755	0.751	0.753	0.747	0.757	0.755
SVM	0.680	0.689	0.683	0.682	0.690	0.680	0.689	0.683	0.682	0.690
BSVM	0.969	0.915	0.941	0.941	0.938	0.969	0.915	0.941	0.941	0.938

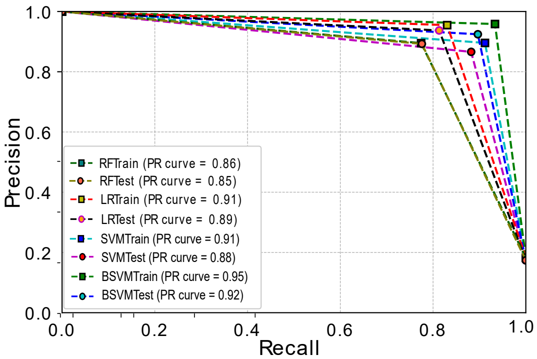

Fig. 8: Precison-Recall Curve for ALL Methods

*3) **BSVM Performance Comparision with Benchmark Schemes**:* In this case study, the performance of BSVM is compared with conventional SVM and two other benchmark classifiers, i.e., RF and LR. Figs. 7 and 8 show ROC-AUC and Precision-Recall curves for all techniques. In this case, BSVM achieves higher accuracy for training and testing, which is up to 0.95 and 0.92, respectively. Furthermore, the confusion matrix in Fig. 2 shows that the ratio of FPR for BSVM is only 2.30%, which is significantly less and is acceptable for real-world scenarios. It implies that the proposed approach is reliable and can be applied in a practical network to identify fraudulent consumers.

*4) **BSVM Performance on Different Train/Test Data**:* ML models are sensitive to the size of training data. This case study aims to confirm whether BSVM maintains its effectiveness when medium to high sizes (60% and 80%) of training data is available for classifier's training. Table I provides an overview of LR, RF, SVM, and BSVM performance for different training dataset sizes. All obtained results of traditional classifiers show an expanding trend. Following investigation of the results, it is observed that the increase in training instances enhance traditional classifiers' performance. Morever, BSVM outperforms other algorithms in terms of Precision, Recall, F1 score, Accuracy, and AUC performance metrics.

V. CONCLUSIONS

An accurate and reliable ETD method is essential for the electric power industry's planning and decision-making process. In this work, the SG's ETD problem is investigated via the combined effect of feature pre-processing and improved classification modules. Precisely, missing and inconsistent values of SM's data are adjusted with data interpolation and standardization techniques. Additionally, the class imbalance problem is resolved with a newly developed STLU technique. Finally, the BOA obtains suitable values for cost penalty, kernel function, and intensive loss function automatically and efficiently for SVM. The numerical results show that our proposed framework is more accurate than LR, DT, and SVM in terms of Precision, Accuracy, etc,. and can effectively be applied to industrial applications.

Acknowledgements: The authors acknowledge funding support from Lancaster University, UK and COMSATS University Islamabad, Pakistan to support this project.

REFERENCES

[1] Mohammadi, Mohsen, et al. "Small-scale building load forecast based on hybrid forecast engine." Neural Processing Letters 48.1 (2018): 329-351.

[2] Khan, Inam Ullah, et al. "Big Data Analytics Based Short Term Load Forecasting Model for Residential Buildings in Smart Grids." IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE(2020): 544-549.

[3] Khan, Inam Ullah, et al. "Heuristic Algorithm Based Optimal Power Flow Model Incorporating Stochastic Renewable Energy Sources."IEEE Access 8 (2020): 148622-148643.

[4] Gao, Wei, et al. "Different states of multi-block based forecast engine for price and load prediction." International Journal of Electrical Power & Energy Systems 104 (2019): 423-435.

[5] Buzau, Madalina Mihaela, et al. "Detection of non-technical losses using smart meter data and supervised learning."IEEE Transactions on Smart Grid 10.3 (2018): 2661-2670.

[6] Wang, Kun, et al. "Robust big data analytics for electricity price forecasting in the smart grid." IEEE Transactions on Big Data 5.1 (2017): 34-45.

[7] Buzau, Madalina-Mihaela, et al. "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters." IEEE Transactions on Power Systems 35.2 (2019): 1254-1263.

[8] Li, Shuan, et al. "Electricity Theft Detection in Power Grids with Deep Learning and Random Forests." Journal of Electrical and Computer Engineering (2019).

[9] Ma, J,et al. "A Review of Forecasting Algorithms and Energy Management Strategies for Microgrids." Systems Science and Control Engineering 6.1 (2018): 237-248.

[10] Buzau, Madalina Mihaela, et al. "Detection of non-technical losses using smart meter data and supervised learning." IEEE Transactions on Smart Grid 10.3 (2018): 2661-2670.

[11] Khan, Zahoor Ali, et al. "Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data." Sustainability 12.19 (2020): 8023.

[12] Han, Hu, Jianwu Dang, and Enen Ren. "Comparative study of two uncertain support vector machines." 2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI). IEEE (2012):388-390.

[13] Zheng, Zibin, et al. "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids." IEEE Transactions on Industrial Informatics 14.4 (2017): 1606-1615.

[14] Nam, Hyeonseob, and Hyo-Eun Kim. "Batch-instance normalization for adaptively style-invariant neural networks." Advances in Neural Information Processing Systems (2018): 2558-2567.