EnsembleNTLDetect: An Intelligent Framework for Electricity Theft Detection in Smart Grid

Yogesh Kulkarni*, Sayf Hussain Z†, Krithi Ramamritham‡, Nivethitha Somu§
*Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
†Department of Computer Science & Engineering, College of Engineering, Guindy, Anna University, Chennai, India
‡Robert Bosch Centre for Data Science & Artificial Intelligence, IIT Madras, Chennai, India
§Department of Electrical Engineering, IIT Bombay, Mumbai, India
{yogeshpict, sayfhussain.10, ramamrithamk, nivethithasomu}@gmail.com

Abstract—Artificial intelligence-based techniques applied to the electricity consumption data generated from the smart grid prove to be an effective solution in reducing Non Technical Loses (NTLs), thereby ensures safety, reliability, and security of the smart energy systems. However, imbalanced data, consecutive missing values, large training times, and complex architectures hinder the real time application of electricity theft detection models. In this paper, we present EnsembleNTLDetect, a robust and scalable electricity theft detection framework that employs a set of efficient data pre-processing techniques and machine learning models to accurately detect electricity theft by analysing consumers' electricity consumption patterns. This framework utilises an enhanced Dynamic Time Warping Based Imputation (eDTWBI) algorithm to impute missing values in the time series data and leverages the Near-miss undersampling technique to generate balanced data.

Further, stacked autoencoder is introduced for dimensionality reduction and to improve training efficiency. A Conditional Generative Adversarial Network (CTGAN) is used to augment the dataset to ensure robust training and a soft voting ensemble classifier is designed to detect the consumers with aberrant consumption patterns. Furthermore, experiments were conducted on the real-time electricity consumption data provided by the State Grid Corporation of China (SGCC) to validate the reliability and efficiency of EnsembleNTLDetect over the state-of-the-art electricity theft detection models in terms of various quality metrics.

Index Terms—Smart grids, Electricity theft, Time series classification, Ensemble learning, Imbalanced data, Dimensionality reduction.

I. INTRODUCTION

With the apparent increase in the global electricity demand, setting up new generation plants is often a difficult and tedious process due to several constraints enforced by the pollution control and environmental conservation policies [1]. The electricity loss during the generation, transmission, and distribution of electricity in the power grid is a critical challenge faced by the power utilities across the globe. Such electricity losses can be classified as [2], [3]: (1) **Technical Losses (TLs):** occurs during transmission. e.g., dissipation

of power in resistors, transmission lines, transformers, etc. and (2) **Non-Technical Losses (NTLs):** the clear difference between the total loss and the TLs. e.g., meter tampering, electricity theft, faulty meters, billing errors, and other irregularities to evade payment to the utility company by the consumers. Among these, NTLs affect the utilities' revenue and the nation's economy with their drastic impact on the quality of power supply, increased load on the power stations, and high tariffs on genuine consumers. Developed countries like U.S and U.K experience NTLs but are not as large as developing countries in Asia, and Africa [4]. In particular, electricity theft, defined as the illegal use of electricity with an intention to avoid billing charges, forms a major part of the NTLs [5]. Electricity theft is a complex research problem with several influential parameters like socio-economic, regional, infrastructure, corruption, managerial, etc. [6]. In general, electricity theft occurs at (i) **Consumers:** energy tapping and meter tampering, (ii) **Utility:** billing inaccuracies, and (iii) **Grid:** bypass meters. The electricity theft at the grid and consumer level results in serious implications for the utilities since it affects their profit and economic wellness of the nation through reduced investments in the power sector, high financial loss (around $4.5-25 billion per YEAR), electrocution deaths, and frequent power outages with overloaded generation units [6]. Moreover, it is tough for the utilities to detect and confirm electricity theft in domestic, commercial and industrial establishments, rural areas and large cities through on-site inspections, an inefficient and expensive manual process.

The advent of Advanced Metering Infrastructures (AMIs) in smart grids accompanied with low-cost smart meters enables two-way communication between the customer and the utility provider for the accounted metering and billing process through fine-grained electricity consumption data & periodic information flow on energy supply and demand. Such advancements accompanied by the massive electricity consumption data have instigated the researchers and the utilities to apply IoT, Big Data, and Artificial Intelligence techniques for the design of efficient and intelligent electricity theft detection mechanisms and accurate utility operations [7]. The design of a reliable and efficient electricity theft detection mechanism

aids the utilities to enforce legal actions on illegal consumers, achieve expected profit and future investments in the power sector for reliable & secure power services. In this way, several machine learning and deep learning techniques have been profoundly applied to energy research problems such as energy trading, virtual power plant, energy consumption monitoring and control, and electricity theft for the design of future intelligent energy networks [8]. The state-of-the-art electricity theft detection approaches can be widely classified into three, namely (i) **State based detection approaches:** monitors the state of the grid and smart meters through RFIDs and sensors; high cost of deployment and maintenance, (ii) **Game theory-based detection approaches:** provides a low-cost solution through modelling a game between the consumers and utility provider; determining utility functions of the participants (consumers, distributors, regulators, etc.) is complex, and (iii) **Artificial intelligence-based detection approaches:** clustering and classification approaches proves to be an cost-effective and reliable solution with the inherent ability of massive electricity consumption data provided by the tamper-proof smart meters to understand the consumer electricity consumption profiles. Owing to the massive electricity consumption data and advanced artificial intelligence approaches, the literature analysis of this paper is confined to the artificial intelligence based electricity theft detection models.

Support Vector Machine (SVM) is the most commonly used technique for electricity theft detection to achieve a high detection rate and fewer false alarms. Certain aspects of the electricity consumption data such as historical consumption data (location, seasonality, and category), load profile information, identification of consumers with a high probability of abnormal behaviour, and high dimensional data have been explored well using SVMs [9], Genetic algorithm-based SVM [5], fuzzy-based SVMs [10], and PCA based SVMs [11]. Electricity thieves have also been identified by analyzing their load profiles at different hierarchies of the power grid (transmission, distribution, and consumer) using hybrid SVM models such as decision tree-based SVMs [12], decision trees-k-nearest neighbour SVMs [13], Extreme learning machine (ELM), online sequential ELMs [14], and even multi-class SVMs [15]. Studies in [16], [17] have carried out a detailed comparative analysis of machine learning models to detect NTLs. Regression and distance-based models like AutoRegressive Moving Average (ARMA) [18], Nonlinear AutoRegressive with eXogenous input (NARX) [19], linear regression [20], k-means (KM) clustering-based ANNs [16], fuzzy C-means clustering [10], Extreme Gradient Boosting [21] and Optimum Path Forest (OPF) [22] were employed to detect NTLs with the detection accuracy between 77%-97%. The inherent ability of deep learning architectures to handle real-time high dimensional smart meter data and automated feature extraction capabilities have led to the development of various single and hybrid deep learning-based electricity theft detection models using Convolutional Neural Networks (CNN) [23], [24], Long Short Term Memory (LSTM) [25], Self-organizing Map (SOM) and

Multilayer Perceptron Artificial Neural Network (MP-ANN) [26], and Particle Swarm Optimization based Stacked Sparse Denoising Auto Encoder (SSDAE) [27]. Notable contributions on electricity theft detection have used Kullback–Leibler divergence [28], a combination of state estimation, multivariate control charts and A* path search algorithm [29], applied self-organizing maps [30], and undersampling boosting algorithms [31].

The challenges in the state-of-the-art electricity theft detection models such as imbalanced nature of the data, consecutive missing values in the time series data, capturing the seasonal trends while imputing missing values, complex architectures, high training time needs to be taken care of for the design of an efficient and robust electricity theft detection model. This paper formulates the identified challenges as following research questions: (i) How to handle large gaps, i.e., consecutive missing values, in time series data with high seasonal trends effectively? (ii) What is the impact of undersampling techniques on generating a balanced electricity consumption dataset without information loss? (iii) How to handle the high dimensional electricity consumption data with appropriate dimensionality reduction technique such that it captures the relations present in the data without loss of important information and low training time? Moreover, (iv) How to leverage the power of generative models for building a robust electricity theft detection model that can provide a high detection rate and less false alarm rate, especially for unseen data? The solution to the above research questions highlighted as the significant contributions are:

- EnsembleNTLDetect, a robust and scalable framework for detecting NTLs in smart grids through analysing consumption patterns from the real-time energy consumption data, is presented.
- The efficiency of enhanced Dynamic Time Warping based Imputation (eDTWBI) is improved by introducing a $Search\_Size$ parameter to reduce the search space of eDTWBI and thereby provides an effective way to handle the large missing gaps in the time series data.
- A customised stacked autoencoder is designed to handle the high dimensional electricity consumption data. The 1,034 dimensions in the original dataset were reduced to 128 dimensions while retaining 99.87% of the original data with reduced training time.
- Conditional GAN is fine-tuned to aid the robust training of classifiers. During the training phase, the classifiers are exposed to real and synthetic data so that the classifiers can model different types of energy consumption values accurately with high confidence scores.
- A soft voting ensemble classifier was designed to leverage the combined efficiency of the bagging-boosting technique based on Random Forest and XGBoost algorithms to achieve a high detection rate and low false alarm rate.
- EnsembleNTLDetect is validated using the real-time electricity consumption data obtained from State Grid Corporation of China (SGCC) using various quality metrics.
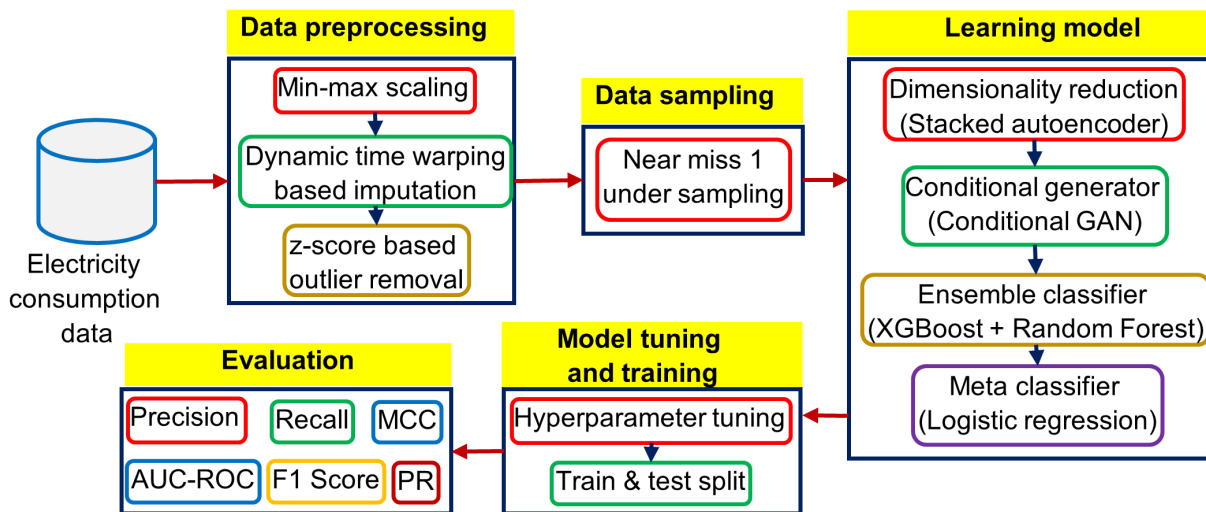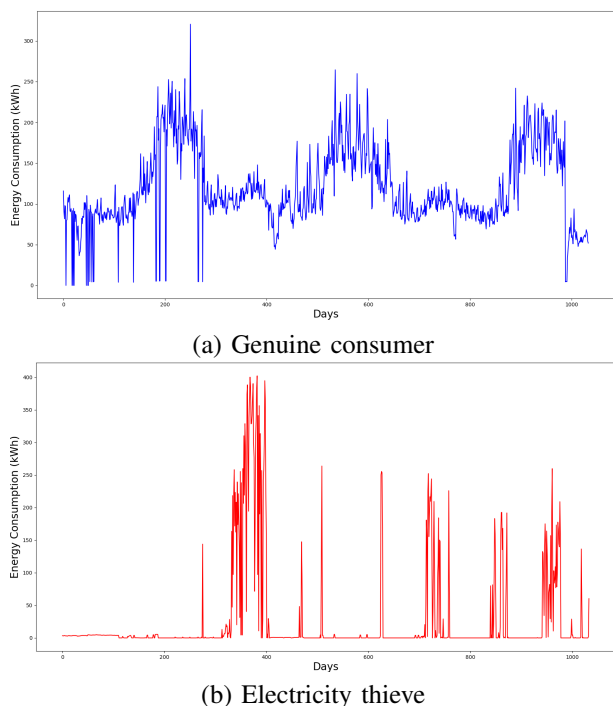
Fig. 1: Architecture of EnsembleNTLDetect

(a) Genuine consumer

(b) Electricity thieve

Fig. 2: Electricity consumption patterns of genuine user and electricity thieve

The paper is structured as follows: Section II provides a detailed insight into the architecture and workflow of EnsembleNTLDetect. Section III highlights the experimental analysis carried out to demonstrate the performance of EnsembleNTLDetect over the state-of-the-art electricity theft detection models in terms of various quality metrics, and Section IV concludes the paper with the scope for further research.

II. METHODOLOGY

Figure 1 presents the overall architecture of the EnsembleNTLDetect, the proposed electricity theft detection model. The complete working methodology of EnsembleNTLDetect with five stages, namely (i) data acquisition and preprocessing, (ii) data sampling, (iii) learning model, (iv) model tuning and training, and (v) evaluation, for efficient and reliable energy theft detection, is detailed below.

With the scarcity of open-source electricity consumption data, this study uses a real-time electricity consumption dataset released by State Grid Corporation of China (SGCC) [32]. The SGCC dataset comprises of daily electricity consumption of 42,372 consumers with 38,757 genuine consumers (*class 0*) and 3615 electricity thieves (*class 1*) recorded over a period of 2 years (1st January 2014 to 31st October 2016). A closer observation to Figure 2 states that the electricity consumption pattern of electricity thieves is aberrant (with more spikes and low) than the genuine consumers. In general, electricity consumption data recorded from the smart meters is aggregated and transferred over data channels to a central location for storage and processing. However, as a result of sensor failures, transmission errors, and server issues, the major challenges in the application of the SGCC time series dataset for electricity theft detection is three-fold (i) 11,233,528 missing values, (ii) imbalanced data in the ratio of 10:1, and (iii) outliers.

A. Data Preprocessing

1) Missing value imputation: The SGCC dataset contains about 11,233,528 missing values which approximates about 25% of the dataset. Ignoring such missing values might lead to downsizing the dataset, which poses a significant challenge in carrying out reliable analysis. Previous works [11], [23]–[25] have used linear interpolation, mean of previous and following day consumption's, filling with mean or median of a complete column, and dropping rows which have missing values beyond a certain threshold. Such methods perform well for isolated
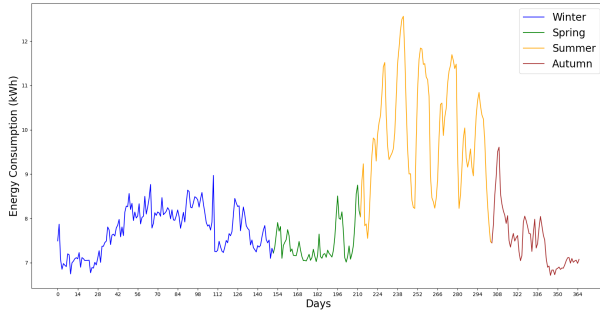
Fig. 3: Variations in electricity consumption over different seasons

data, i.e., one to three missing values but fail miserably for realistic imputations in data with consecutive missing values, correlations, seasonality trends and complex distribution.

RQ1: *How to handle large gaps, i.e., consecutive missing values, in time series data with high seasonal trends in an effective way?*

This work employs an enhanced version of Dynamic Time Warping (DTW) based Imputation (eDTWBI) [33], an algorithm for the generation of optimal time series data, i.e., to fill the large gaps (consecutive missing values) in the SGCC dataset. eDTWBI uses DTW to find two reference window that lies before and after the gap, which is also similar to the considered large gap such that the distance between them is minimal. The reference window is represented as grids for quadratic time complexity. Owing to the large size of the SGCC dataset, this work introduces *Search_Size* parameter to reduce the search space. Further, the seasonality trends in the dataset are taken care of by *Search_Size* to ensure that the imputation for gaps in a particular season is bounded by the similar sequences obtained from the same season. For example, gaps in the summer season ($Search\_Size$ = 1) are imputed using the similar sub-sequence obtained within the year's summer season (3 months). Figure 3 shows the seasonality trend of the dataset for the year 2015. This significant improvement in eDTWBI helps enhance the learning base, prediction ability, data dynamics and reduces the temporal constraints between the reference window. Algorithm 1 presents the pseudo-code of the eDTWBI algorithm for missing value imputation.

For a time-series x with large number of consecutive missing values, a gap of size T at position t is defined as the portion between two points x_t and x_{t+T-1} that has x_i *NaN* values, where $i = t : t + T - 1$. Further, Q forms the temporal window before missing values, R is a reference window for imputation that should lie within the same season, and lp is an array of location pointers pointing to reference windows with a minimum DTW cost. The workflow of eDTWBI is highlighted below:

Step 1: Create reference window: For a gap of T size at position t, create two reference windows (R_{Before} & R_{After}) containing data points that lie before and

---

Algorithm 1: Enhanced Dynamic Time Warping with Reduced Search Space

Input: $x = \{x_1, x_2, \ldots, x_N\}$, t, T, $Search\_Size = 0$, $Q = D[t - T : t - 1]$, $lp = [\ ]$

Output: Imputed DataFrame

1 Construct a DTW_Matrix D consisting of n rows and m columns where $(m, n) \in len(sequence)$ and $D_{ij} = distance(x_i, x_j)$

2 Create a $Search\_Space$ $S = D[1 : t - 2T]$

3 Set $Derivative\_Cost\_Measure$ for DTW algorithm using the following formula:

$$D_x[a] = \frac{(x_a - x_{a-1}) + ((x_{a+1} - x_{a-1})/2)}{2} \quad (1)$$
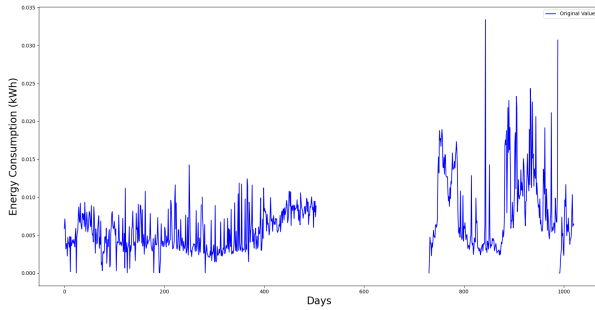
$$Derivative\_Cost = (D_x[i] - D_x[j])^2 \quad (2)$$

4 $i \leftarrow 1$ & $Search\_Size \leftarrow 3$

5 **while** $i < len(S)$ **do**

6     $k \leftarrow i + T - 1$

7   Save a reference window $R_{Before}(i) = S[i : k]$

8   **if** $R_{Before}(i)$ in $Search\_Size$ **then**

9         $dtw\_cost = DTW(Q, R_{Before}(i))$

10   **if** $dtw\_cost < Derivative\_Cost$ **then**

11             $i \leftarrow i + 1$

12   **else**

13   Save position of $R_{Before}(i)$ to lp

14   **end**

15   **else**

16   break

17   **end**

18 **end**

19 Replace all missing values at position t by an array of values after the Q's window having minimum DTW cost using the lp list.

20 **return** *Imputed Dataframe*
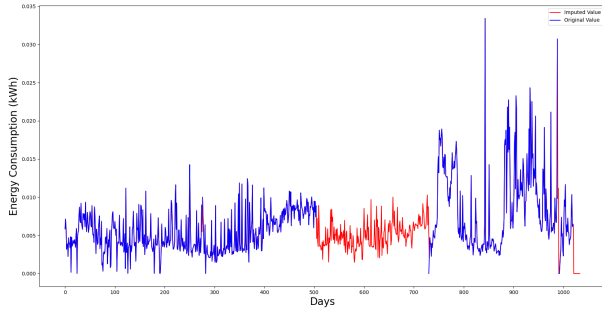
---

after the gap of length T.

Step 2: Find highly similar windows: Create sliding windows of length T for the data points that lie before and after the gap T. Identify the most similar window to reference windows (R_{Before} & R_{After}) by calculating the DTW cost and Derivative Dynamic Time Warping ($DDTW$) cost [34]. Since $DDTW$ cost is robust to outliers, save the windows whose $DDTW$ cost is lesser than the DTW cost.

Step 3: Imputation: For unbiased results, impute the large gap of length T) with the average value of the most similar windows.

Figure 4 provides the electricity consumption data with large gaps and imputed values. The complete process of imputation for the whole dataset took less than *30 minutes*. The novelty introduced in the eDTWBI algorithm in terms of restricting the search space has improved the efficiency of the EnsembleNTLDetect in terms of performance measures and execution time.

(a) Consumption patterns with large gap

(b) Consumption patterns with imputed values

Fig. 4: Imputation using Algorithm 1

2) Outliers: The imputed dataset is subjected to outlier detection and removal using *Z-score*, a computationally inexpensive outlier removal technique given by equation 3.

$$Z = \frac{\mathrm{X} - \mu}{\sigma} \tag{3}$$

Where X refers to the data point, μ is the mean, σ is the standard deviation, and Z is the *Z-score*. All data points which have $Z > 3$ or $Z < -3$ were dropped.

B. Handling Imbalanced Data

RQ 2: *What is the impact of undersampling techniques on the generation of balanced electricity consumption dataset without information loss?*
The SGCC dataset is imbalanced in the ratio of 10:1 with class 0 (genuine consumers) as the majority and class 1 (electricity thieves) as the minority. Although SMOTE has been widely used in the literature to handle data imbalance issues in the SGCC dataset, this work prefers to use the under-sampling technique due to the following reasons.

- From figure 2, it is clear that the electricity consumption pattern of electricity thieves is aberrant when compared with the genuine consumer's consumption trend. In general, the application of SMOTE generates such unusual patterns with unrealistic consumption values for the minority class (electricity thief) and are highly susceptible to overfitting. Such unusual patterns pose difficulty for the classifiers in extracting meaningful information and accurate classification; furthermore, increasing the number of samples in the minority class results in low accuracy and a minute increase in recall score. In such cases,

TABLE I: Performance comparison of Random Forest using SMOTE & Near Miss

Output	Parameter	SMOTE + RF	Near–miss + RF
0 (genuine)	Precision	0.94	0.98
1 (theft)	Precision	0.39	0.62
0	Recall	0.96	0.99
1	Recall	0.29	0.57
0	F1-score	0.95	0.98
1	F1-score	0.33	0.59

the application of under-sampling approaches ensures the classifier to establish a fine boundary between genuine consumers with usual trend and electricity thieve with unusual consumption trends. Refer Section III for more details.
- The computational efficiency of under-sampling techniques is another reason for its consideration over SMOTE.

This work employs *Near-Miss* (version 1) [35], a simple and effective under-sampling technique to handle the data imbalance in the SGCC dataset. About 40,488 samples obtained after applying the z-score outlier removal technique were reduced to 6,300 with an equal split of 3,150 samples for class 0 and class 1. An interesting point to note is that the *meaningful* information lost, if any, during under-sampling was handled by the CTGAN.

We chose random forest classifier since it is part of our proposed soft voting ensemble for performing this experiment. From Table I we can infer that for the SGCC dataset, near-miss undersampling works better in comparison with SMOTE oversampling. The recall score for SMOTE (theft class) is 0.29 since it augments the aberrant minority samples making the task more challenging for the classifier. In contrast, it is 0.57 for Near Miss since it downsamples the majority class having certain periodicity.

C. Learning Model

1) Stacked AutoEncoder: With 1,034 timestamps in the SGCC dataset, it is evident that these features carry some intrinsic relation between them. Therefore, applying the dimensionality reduction technique to identify a set of informative features is the ideal step towards the design of an efficient and reliable electricity theft detection model. Unfortunately, principal component analysis [36], [37], the most commonly used dimensionality reduction technique, fails to capture the convoluted low-dimensional manifold structure and model the intrinsic relations in the time series data [38].

RQ 3: *How to handle the high dimensional electricity consumption data with appropriate dimensionality reduction technique such that it captures the relations present in the data without loss of important information and low training time?*
In such cases, autoencoder architectures have been successfully established as an efficient dimensionality reduction tool for fault diagnosis [39], high-content screening data [40] and intrusion detection systems [41]. Autoencoders are a

TABLE II: Architecture of Stacked Auto-Encoder consisting of three Auto-Encoder's

Stacked Auto-Encoder		Auto-Encoder 1		Auto-Encoder 2		Auto-Encoder 3	
Layers	Parameters	Layers	Parameters	Layers	Parameters	Layers	Parameters
Input	(1034,)	Input	(1034,)	Input	(512,)	Input	(256,)
Dense	512, ReLU, param = 529,920	Dense	512, ReLU, param = 529,920	Dense	256, ReLU, param = 131,328	Dense	128, ReLU, param = 32,896
Batch-Norm	param = 2,048	Batch-Norm	param = 2,048	Batch-Norm	param = 1,024	Batch-Norm	param = 512
Dense	256, ReLU, param = 131,328	Dense	1034, Sigmoid, param = 530,442	Dense	512, Sigmoid, param = 131,584	Dense	256, Sigmoid, param = 33,024
Batch-Norm	param = 1,024			Batch-Norm	param = 2,048	Batch-Norm	param = 1,024
Dense	128, ReLU, param = 32,896						
Batch-Norm	param = 512						
Dense	256, Sigmoid, param = 33,024						
Batch-Norm	param = 1,024						
Dense	512, Sigmoid, param = 131,584						
Batch-Norm	param = 2,048						
Dense	1034, Sigmoid, param = 530,442						
Total Parameters	1,395,850	Total Parameters	1,062,410	Total Parameters	265,984	Total Parameters	67,456

special kind of neural networks which maps the input of a specific dimension to a latent space of reduced dimension and then decode the latent representation to a reconstructed input by minimizing the reconstruction error. This work presents a stacked autoencoder with three autoencoders specifically designed to perform an unsupervised learning based dimensionality reduction on the feature space. Table II shows the model architecture of the stacked autoencoder with three autoencoders. The weights of the hidden layers are set by training each autoencoder individually. Figure 5 presents the training procedure of the autoencoder for dimensionality reduction, and figure 6 presents the model loss (during training) for the three autoencoders, we can see that our model converges quickly within 100 epochs. The proposed stacked autoencoder has reduced 1,034 dimensions in the SGCC dataset to 128 dimensions, wherein 99.87% of the original data was captured with no loss of information. The application of autoncoder based dimensionality reduction technique has boosted the efficiency of this framework resulting in faster training and inference.

RQ 4: *How to leverage the power of generative models for building a robust electricity theft detection model that can provide a high detection rate and less false alarm rate, especially for unseen data?*

2) Handling Corner Cases: To ensure the efficiency and reliability of EnsembleNTLDetect in real-time environments, it is highly essential to look on to the critical aspects such as loss of critical information due to undersampling and ability to handle various input types. To handle such issues, this work uses Conditional Tabular Generative Adversarial Network (CTGAN) [42] to create more samples in such a
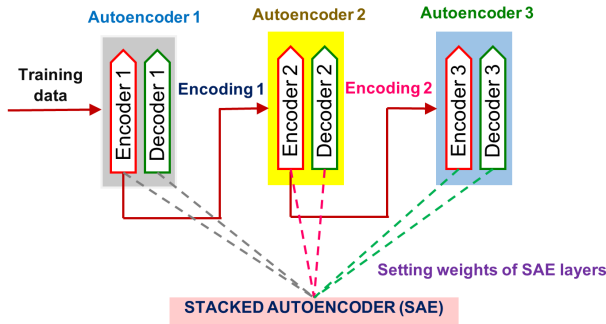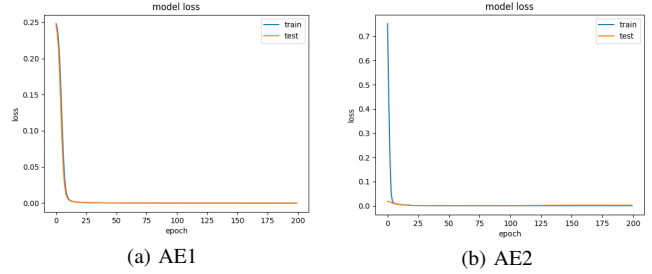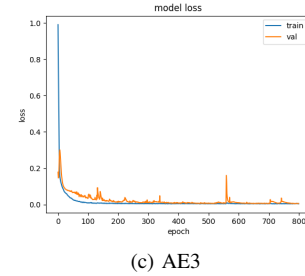

(a) AE1

(b) AE2

(c) AE3

Fig. 6: Model Loss during training for the AE's

way that the learning model is exposed to a wide range of data samples from both the classes (genuine and electricity thieves). Figure 7 shows the architecture of CTGAN.

CTGAN improves the tabular data generation through (i) mode-specific normalization: improves modelling multi-modal distributions in numeric columns and (ii) conditional training by sampling: ensures that the rare categorical data are evenly sampled. Since the SGCC dataset comprises of daily electricity consumption values (numerical), mode-specific normalization was preferred over the conditional training. The number of
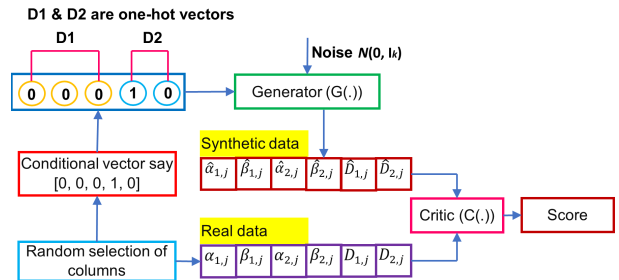

Fig. 5: Autoencoder training pipeline

Fig. 7: Architecture of CTGAN

TABLE III: Performance analysis of EnsembleNTLDetect and basic machine learning architectures for NTL detection

Classifier	Precision	Recall	F1-Score	AUC-ROC	PR-AUC	MCC
Naive Bayes	0.65	0.53	0.54	0.54	0.12	0.14
ExtraTrees	0.56	0.56	0.56	0.58	0.11	0.12
K–Neighbors	0.65	0.52	0.53	0.64	0.16	0.12
Linear Support Vector Machine	0.78	0.50	0.47	0.67	0.22	0.03
Logistic Regression	0.75	0.50	0.49	0.68	0.21	0.09
Multi–layer Perceptron	0.73	0.56	0.58	0.77	0.32	0.24
Random Forest	0.79	0.53	0.54	0.80	0.34	0.21
Gradient Boosting	0.76	0.54	0.55	0.79	0.33	0.21
EnsembleNTLDetect	**1.00**	**0.98**	**0.99**	**0.99**	**0.99**	**0.98**

modes for each column determined by the Variational Gaussian Mixture (VGM) [43] was used for normalization. These normalized values were used during the training phase and are transformed to their original scales after obtaining the generated data. Due to the complexity involved in training GANs, Wasserstein GAN with gradient penalty [44] and PacGAN [45] were used to ensure robust learning stability (prevents mode collapse), and the generator network provides diverse samples, respectively. For the SGCC dataset, 10,000 samples in the ratio of 2:1 between genuine consumers and electricity thieves were generated using CTGAN.

3) Soft voting ensemble classifier: A simple soft voting classifier with Random Forest [46], and XGBoost [47] classifiers as base learners were designed for accurate classification of genuine consumers and electricity thieves with high detection rates and less false alarm rate. Further, logistic regression was used as a meta learner to create a linear relationship between the input and output variables, i.e., a fine boundary between genuine consumers and electricity thieves using the maximum-likelihood estimation based on coefficients obtained from the training data. A soft voting mechanism is used so that the output class has the highest average probability. The output label $\hat{y}$ of a soft voting ensemble model with m classifiers of p probability is given in equation 4 where, w_j is the uniform weight of the jth classifier, $i \in \{0, 1\}$.

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j p_{ij} \qquad (4)$$

The optimal hyperparameters of the ensemble classifier were obtained through rigorous 10-fold cross-validation using GridSearchCV with the best validation accuracy. Table IV presents the optimal hyperparameters of the ensemble classifier.

TABLE IV: Optimal Hyper-Parameters

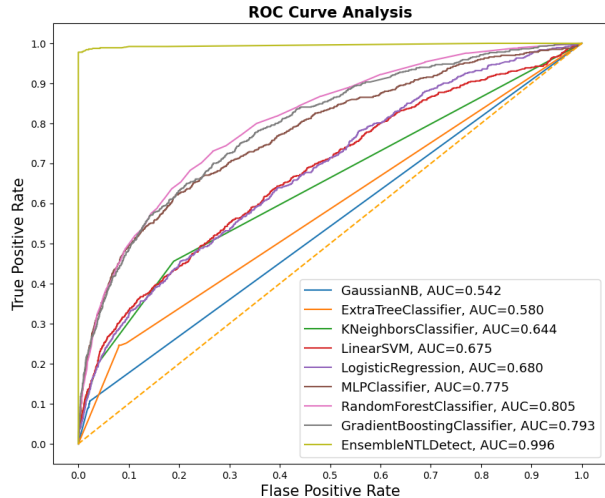| Algorithm	Parameters
Random Forest	n_estimators=300, max_features="sqrt", criterion="gini", min_samples_leaf=5, class_weight="balanced"
XGBoost	objective="binary:logistic", learning_rate=0.03, n_estimators=500, max_depth=1, subsample=0.4
Logistic Regression	penalty="l2",C=100

Fig. 8: ROC Curve of EnsembleNTLDetect and basic ML architectures

III. RESULTS & DISCUSSIONS

The experimental design and analysis of EnsembleNTLDetect was carried out in a working environment with an Intel i5 (10th Gen) processor running Windows 10 operating system with 8 GB RAM. For faster implementations, EnsembleNTLDetect and the comparative models were designed and executed in Google Colaboratory. Moreover, the implementation of the EnsembleNTLDetect and the comparative models was done using Python 3.7 with necessary packages and libraries such as scikit-learn, TensorFlow, Keras.

A. Performance Metrics

Due to the imbalanced nature of the SGCC electricity consumption dataset, the essential quality metrics derived from the confusion matrix were chosen over accuracy to assess the performance of EnsembleNTLDetect over the state-of-the-art electricity theft detection models. Here, the primary measures of the confusion matrix represent (i) True Positive (TP): correctly classified as electricity thieve, (ii) True Negative (TN): correctly classified as genuine consumer, (iii) False Positive (FP): misclassified as electricity thieve and (iv) False Negative (FN): misclassified as a genuine consumer.

TABLE V: Performance analysis of EnsembleNTLDetect and state-of-the-art electricity theft detection models

Classifier	Precision	Recall	F1-Score	AUC-ROC	PR-AUC	MCC
SVM [11]	0.75	0.71	0.72	0.60	0.78	0.67
XGBOOST [48]	0.95	0.82	0.86	0.88	0.87	0.81
Bi-directional Gated Recurrent Unit [49]	0.82	0.82	0.84	0.84	0.78	0.68
CNN + RF [23]	0.80	0.89	0.85	0.90	0.87	0.84
Wide CNN [24]	0.84	0.88	0.86	0.86	0.81	0.73
CNN + LSTM [25]	0.94	0.82	0.88	0.88	0.87	0.78
LSTM + MLP [50]	0.90	0.87	0.85	0.90	0.90	0.80
Semi-Supervised AutoEncoder [51]	0.86	0.80	0.83	0.84	0.81	0.82
EnsembleNTLDetect	**1.00**	**0.98**	**0.99**	**0.99**	**0.99**	**0.98**

All the comparison with the previous work was done keeping test size = 0.2.

1) Precision: The ratio of consumers (thieves) correctly classified as electricity thieves to the total positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

2) Recall / True Positive Rate (TPR): The ratio of consumers (thieves) correctly classified as thieves to all the predictions of actual class.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

3) F1-score: The harmonic mean of precision and recall.

$$F1 - \text{ Score} = 2 * \frac{\text{Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}} \quad (7)$$

4) AUC-ROC: A probability curve that plots the TPR against FPR.

$$\text{FPR} = \frac{FP}{TN + FP} \quad (8)$$

5) PR-AUC: Represents the precision against recall score over varying thresholds. A high score indicates that a classifier can accurately achieve TPs with very less number of FPs & FNs.

6) Matthews Correlation Coefficient (MCC): The most reliable statistical measure for imbalanced data. A high score represents that the classifier performed well for all categories of the confusion matrix (TPs, FPs, TNs, FNs).

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (9)$$

B. Analysis and discussions

At the initial phase, the performance of the EnsembleNTLDetect was evaluated over the basic machine learning architectures like in terms of the quality metrics mentioned in Section III-A. The metric values provided in this section are the average values obtained after 25 consecutive and iterative runs. Table III presents a comparative analysis of EnsembleNTLDetect over the basic machine learning architectures for NTL detection. Even though it is evident that EnsembleNTLDetect demonstrates its performance with better quality metrics, the contrast models can be categorized into two groups, (i) Type 1: Naive Bayes, Extra Trees, K-nearest neighbours, Linear SVM and logistic regression (AUC-ROC < 0.75 and MCC < 0.15) and (ii) Type 2: MLP, random forest and gradient boosting (AUC-ROC < 0.85 and MCC < 0.25) based on the AUC-ROC and MCC scores. Figure
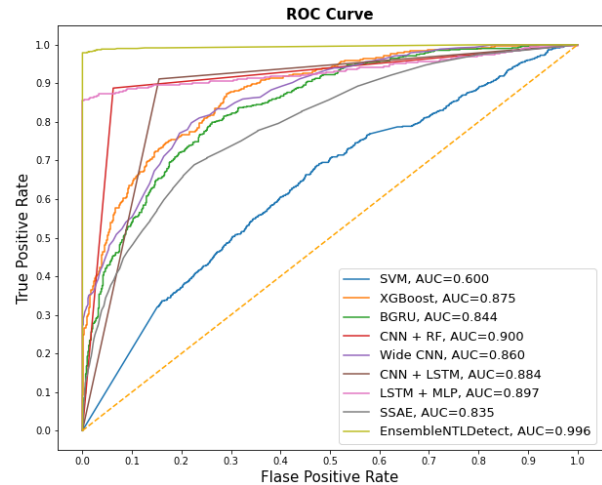

Fig. 9: ROC Curve of EnsembleNTLDetect Vs. state-of-the-art electricity theft detection models

8 provides the AUC-ROC curve of EnsembleNTLDetect and basic machine learning-based NTL detection models.

Further, the complete set of experiments were repeated to validate the performance of EnsembleNTLDetect for different sizes of the train and test dataset (Table VI). The train-test split size of 50:50, 60:40, and 70:30 was used to demonstrate the impact of CTGAN on the overall performance of EnsembleNTLDetect. In all the cases, EnsembleNTLDetect provides a quality metric score of above 0.97, especially when the training data size is reduced to 60% and 70%, the application of CTGAN for synthetic sample generation nullifies the imbalanced nature of the dataset and provides AUC-ROC and MCC values of above 0.97.

Table V provides a detailed comparative analysis of EnsembleNTLDetect over the state-of-the-art electricity theft detection models in terms of the considered quality metrics. Figure 9 presents the AUC-ROC analysis of EnsembleDetect

TABLE VI: Performance analysis of EnsembleNTLDetect for different train-test splits

Split Size	Precision	Recall	F1-Score	AUC-ROC	MCC
70 : 30	0.995	0.978	0.987	0.994	0.974
60 : 40	0.995	0.979	0.987	0.993	0.975
50 : 50	0.998	0.972	0.985	0.989	0.971

and the state-of-the-art electricity theft detection models. Even though SVMs are widely explored and applied in various forms for NTL detection, they provide marginal performance due to overfitting and high susceptibility to noise. Despite its benefits, such as memorization and generalization from the deep CNN architecture and wide components, Wide and deep CNN is ranked as an average model due to its inability to model long time series. CNN-LSTM architecture with CNNs as feature extractors and LSTMs to model long sequences in the time series data proves to be the best in the state-of-the-art electricity theft detection models. However, the complex architectures, high training time, and overfitting issues create a major impact while deploying in a real-time environment. In such cases, EnsembleNTLDetect provides optimal performance with minimal training time. Moreover, state-of-the-art NTL detection methods in [23], [48], [50] have not been verified on SGCC dataset, which partly explains their subpar performance on this dataset and lack of generalization ability. Table VII shows the execution time (in Mins) for each component of EnsembleNTLDetect. It takes about an hour to deploy and execute the entire framework from scratch, while the tradeoff between efficiency and accuracy is perfectly managed by the time taken to generating predictions within few milliseconds. The overall computation time can be further reduced with high-end computing infrastructures at the smart grid control stations. The extensive experiments carried out using the SGCC dataset have resulted in the following observations,

1. eDTWBI works exceptionally well in all scenarios, except when the nearby consumption values are very low, resulting in 0 as consumption values. Further, the introduction of $Search\_Size$ parameter reduces the overall execution time of the DTW algorithm through limiting the search space of the search window.

2. Near-miss undersampling does not result in any loss of information which is indicated by a high recall score.

3. The training pipeline of stacked autoencoders was highly efficient and effective such that there was no loss of critical information even after reducing the dimensions by approximately 87%. It also enables faster training and inference of the ML classifiers.

4. Fine-tuning CTGAN aids the classifier to model all possible types of original and synthetic data, thereby enhancing the robustness of the model to completely unseen or aberrant data.

IV. Conclusions & Future Work

This paper presents EnsembleNTLDetect, a robust and scalable framework to detect electricity theft by analyzing the electricity consumption patterns of consumers. Specific contributions attributed to address the limitations in the state-of-the-art electricity theft detection models are, (i) **Consecutive missing values:** enhanced version of dynamic time warping algorithm imputes the large gaps in the time series data and seasonality trends were preserved by introducing $Search\_Size$ parameter to restrict the search space of the reference points within the

TABLE VII: Execution time analysis of EnsembleNTLDetect

Components	Execution time (in Mins)
eDTWBI based imputation	28.6[*]
Z-Score for outlier removal	1.2
Near miss undersampling (Version 1)	6.7
Stacked Autoencoder: (Training + generating latent vectors)	5.4
CTGAN: (Training + generating 10,000 Samples)	13.4
Soft Voting Ensemble: (Hyperparameter Tuning + training + prediction)	16.8
Total execution time	72.1

* Imputation was done locally

season range of the missing values, (ii) **Imbalanced dataset:** near-miss undersampling technique generates the balanced dataset without information loss, (iii) **High dimensional data:** stacked autoencoders with three autoencoders performs an unsupervised learning based dimensionality reduction on the feature space, (iv) **Efficient training:** a fine-tuned conditional GAN provides effective training for the classifiers through exposing them to real and synthetic data with different energy consumption patterns, and (v) **Effective classification:** a soft voting ensemble classification model that uses random forest and XGBoost learns the complex high dimensional electricity consumption patterns to detect the consumers with aberrant consumption patterns with high detection rate and less false alarm rate. Extensive experimental analysis on the SGCC real-time electricity consumption dataset demonstrates that EnsembleNTLDetect outperforms the state-of-the-art electricity theft detection models by accurately classifying genuine consumers and electricity thieves with a recall and MCC score of 0.98. Further, the application of stacked autoencoders based dimensionality reduction technique has reduced the total computational cost of EnsembleNTLDetect such that it ensures simple and effective deployment for large scale real-time electricity theft detection. Experiments and in-depth analysis of EnsembleNTLDetect on different open-source electricity consumption datasets for detecting NTLs in smart grids are planned as a future directive of this work. In addition, detailed analysis on the effects of consumer metadata on the consumption patterns requires more attention to understand the various social-psychological factors that impact the consumers' electricity consumption patterns.

V. Acknowledgments

References

[1] T. Ahmad, H. Chen, J. Wang, Y. Guo, "Review of various modeling techniques for the detection of electricity theft in smart grid environment," in *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 2916–2933, 2018.

[2] Patrick GLAUNER, "Artificial intelligence for the detection of electricity theft and irregular power usage in emerging markets," Ph.D. Dissertation, Karlsruhe University of Applied Sciences, 2019.

[3] Rouzbeh Razavi and Martin Fleury, "Socio-economic predictors of electricity theft in developing countries: An indian case study," in *Energy for Sustainable Development*, vol. 49, pp. 1–10, 2019.

[4] Osman Yakubu, Narendra Babu C., and Osei Adjei, "Electricity theft: Analysis of the underlying contributory factors in Ghana," in *Energy Policy 123*, pp. 611–618, 2018.

[5] 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," in *Proceedings of IEEE Region 10 Annual International Conference (TENCON)*, pp. 1–6, 2008.

[6] Soma Shekara Sreenadh Reddy Depuru, Lingfeng Wang, and Vijay Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," in *Energy Policy 39*, vol. 2, pp. 1007–1015, 2011.

[7] D. Roverso, "Using machine learning and smart meter data for fraud detection," in https://blogs.esmartsystems.com/using-machine-learning-and-smart-meter-data-for-fraud-detection

[8] Next Kraftwerke, "What is Artificial Intelligence in the Energy Industry?"

[9] Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Syed Khaleel Ahmed, and Malik Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," in *IEEE Transactions on Power Delivery*, vol. 25, pp. 1162–1171, 2010.

[10] Eduardo Werley S. Dos Angelos, Osvaldo R. Saavedra, Omar A.Carmona Cortés, and André Nunes De Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," in *IEEE Transactions on Power Delivery*, vol. 26, pp. 2436–2442, 2011.

[11] M. Anwar, N. Javaid, A. Khalid, M. Imran and M. Shoaib, "Electricity theft detection using pipeline in machine learning," in *International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 2138-2142.

[12] Anish Jindal, Amit Dua, Kuljeet Kaur, Mukesh Singh, Neeraj Kumar, et al, "Decision tree and svm-based data analytics for theft detection in smart grid," in *IEEE Transactions on Industrial Informatics*, vol. 12, pp. 1005–1016, 2016.

[13] Xiangyu Kong, Xin Zhao, Chao Liu, Qiushuo Li, DeLong Dong, et al, "Electricity theft detection in low-voltage stations based on similarity measure and dt-kvsvm," in *International Journal of Electrical Power & Energy Systems*, vol. 125, 106544, 2021.

[14] A. H. Nizar, Z. Y. Dong, and Y. Wang., "Power utility nontechnical loss analysis with extreme learning machine method," in *IEEE Transactions on Power Systems*, vol. 23, pp. 946–955, 2008.

[15] Paria Jokar, Nasim Arianpoo, and Victor C.M. Leung., "Electricity theft detection in ami using customers' consumption patterns," in *IEEE Transactions on Smart Grid*, vol. 7, pp. 216–226, 2016.

[16] J. Jeyaranjani and D. Devaraj., "Machine learning algorithm for efficient power theft detection using smart meter data," in *International Journal of Engineering and Technology(UAE)*, vol. 7, pp. 900–904, 2018.

[17] Khawaja Moyeezullah Ghori, Rabeeh Ayaz Abbasi, Muhammad Awais, Muhammad Imran, Ata Ullah, et al, "Performance analysis of different types of machine learning classifiers for non-technical loss detection," in *IEEE Access*, vol. 8, pp. 16033–16048, 2019.

[18] Ayodele Isqeel Abdullateef, Momoh-jimoh Eyiomika Salami, Mohamud Ahmed Musse, and Mobolaji Agbolade Onasanya, "Consumer Load Prediction and Theft Detection on Distribution Network Using Autoregressive Model," in *International Journal of Scientific & Engineering Research*, vol. 4, pp. 1609–1615, 2013.

[19] Abdullateef Ayodele Isqeel, Salami Momoh Jimoh Eyiomika, and Tijani Bayo Ismaeel, "Consumer Load Prediction Based on NARX for Electricity Theft Detection," in *Proceedings of 6th International Conference on Computer and Communication Engineering:Innovative Technologies to Serve Humanity (ICCCE)*, pp. 294–299, 2016.

[20] Sook Chin Yip, Kok Sheik Wong, Wooi Ping Hew, Ming Tao Gan, Raphael C.W. Phan, et al, "Detection of energy theft and defective smart meters in smart grids using linear regression," in *International Journal of Electrical Power and Energy Systems*, vol. 91, pp. 230–240, 2017.

[21] Z. Yan and H. Wen, "Electricity theft detection base on extreme gradient boosting in AMI," in *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1-6, 2020.

[22] Leandro Aparecido Passos Júnior, Caio César Oba Ramos, Douglas Rodrigues, Danillo Roberto Pereira, André Nunes de souza, et al, "Unsupervised non-technical losses identification through optimum-path forest," in *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016.

[23] Shuan Li, Yinghua Han, Xu Yao, Song Yingchen, Jinkuan Wang, et al, "Electricity theft detection in power grids with deep learning and random forests," in *Journal of Electrical and Computer Engineering*, 2019.

[24] Zibin Zheng, Yatao Yang, Xiangdong Niu, Hong Ning Dai, and Yuren Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," in *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 1606–1615, 2018.

[25] Hasan Md. N., Rafia N. Toma, Abdullah-Al Nahid, M M.M. Islam, and Jong-Myon Kim., "Electricity theft detection in smart grid systems: a cnn-lstm based approach," in *Energies* vol. 12, no. 17: 3310, 2019.

[26] Matheus Alberto de Souza, Jose LR Pereira, Guilherme de O Alves, Braulio C de Oliveira, Igor D Melo, et al, "Detection and identification of energy theft in advanced metering infrastructures," in *Electric Power Systems Research*, vol. 182, 106258, 2020.

[27] Yifan Huang and Qifeng Xu, "Electricity theft detection based on stacked sparse denoising autoencoder," in *International Journal of Electrical Power & Energy Systems*, vol. 125, 106448, 2021.

[28] V. B. Krishna, K. Lee, G. A. Weaver, R. K. Iyer and W. H. Sanders, "F-DETA: A framework for detecting electricity theft attacks in smart grids," in *46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 407-418, 2016.

[29] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," in *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1023-1032, 2018.

[30] J. E. Cabral, J. O. P. Pinto, E. M. Martins and A. M. A. C. Pinto, "Fraud detection in high voltage electricity consumers using data mining," in *IEEE/PES Transmission and Distribution Conference and Exposition*, pp. 1-5, 2008.

[31] N. F. Avila, G. Figueroa and C. -C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," in *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7171-7180, 2018

[32] State Grid Corporation of China Dataset: http://www.sgcc.com.cn

[33] Phan TTH., Poisson Caillault É., Bigand A., "eDTWBI: effective imputation method for univariate time series," in *International Conference on Computer Science, Applied Mathematics and Applications*, pp. 121-132, 2020.

[34] Eamonn J. Keogh and Michael J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining (SDM)*, 2001.

[35] Mani, Inderjeet, and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, USA, ICML, 2003.

[36] K. Pearson, "On lines and planes of closest fit to systems of points in space," in *Philosophical Magazine*, 1901.

[37] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945-1959, 2005.

[38] W. Wang, Y. Huang, Y. Wang and L. Wang, "Generalized autoencoder: a neural network framework for dimensionality reduction," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 496-503, 2014.

[39] M. Cui, Y. Wang, X. Lin and M. Zhong, "Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine," in *IEEE Sensors Journal*, vol. 21, no. 4, pp. 4927-4937, 2021.

[40] Zamparo, Lee, and Zhaolei Zhang, "Deep autoencoders for dimensionality reduction of high-content screening data," *arXiv preprint arXiv:1501.01348*, 2015.

[41] G. Muhammad, M. S. Hossain and S. Garg, "Stacked autoencoder-based intrusion detection system to combat financial fraudulent," in *IEEE Internet of Things Journal*, 2020.

[42] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019.

[43] Christopher M Bishop, Pattern recognition and machine learning, springer, 2006.

[44] Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courvle, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017.

[45] Lin, Zinan, Ashish Khetan, Giulia Fanti, and Sewoong Oh, "Pacgan: The power of two samples in generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2018.

[46] L. Breiman, "Random forests," in *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[47] Tianqi Chen and Carlos Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, NY, USA, 785–794, 2016.

[48] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," in *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661-2670, 2019.

[49] Gul, Hira, Nadeem Javaid, Ibrar Ullah, Ali M. Qamar, Muhammad K. Afzal, et al, "Detection of non-technical losses using sostlink and bidirectional gated recurrent unit to secure smart meters," in *Applied Sciences*, 10, no. 9: 3151, 2020.

[50] M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero and A. Gómez-Expósito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," in *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1254-1263, 2020.

[51] Lu, Xiaoquan, Yu Zhou, Zhongdong Wang, Yongxian Yi, Longji Feng, et al, "Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the Smart Grid," in *Energies* 12, no. 18: 3452, 2019.