energies

MDPI

Article

Noise Reduction Power Stealing Detection Model Based on Self-Balanced Data Set

Haiqing Liu, Zhiqiao Li * and Yuancheng Li

School of Control and Computer, North China Electric Power University, Beijing 102206, China; hqliu@ncepu.edu.cn (H.L.); yuancheng_li@ncepu.edu.cn (Y.L.)

* Correspondence: li_zhiqiao@163.com; Tel.: +86-185-0072-5152

Received: 15 February 2020; Accepted: 5 March 2020; Published: 7 April 2020

Abstract: In recent years, various types of power theft incidents have occurred frequently, and the training of the power-stealing detection model is susceptible to the influence of the imbalanced data set and the data noise, which leads to errors in power-stealing detection. Therefore, a power-stealing detection model is proposed, which is based on Improved Conditional Generation Adversarial Network (CWGAN), Stacked Convolution Noise Reduction Autoencoder (SCDAE) and Lightweight Gradient Boosting Decision Machine (LightGBM). The model performs Generation- Adversarial operations on the original unbalanced power consumption data to achieve the balance of electricity data, and avoids the interference of the imbalanced data set on classifier training. In addition, the convolution method is used to stack the noise reduction auto-encoder to achieve dimension reduction of power consumption data, extract data features and reduce the impact of random noise. Finally, LightGBM is used for power theft detection. The experiments show that CWGAN can effectively balance the distribution of power consumption data. Comparing the detection indicators of the power-stealing model with various advanced power-stealing models on the same data set, it is finally proved that the proposed model is superior to other models in the detection of power stealing.

Keywords: conditional generation network; data set imbalance; stacked convolution noise reduction encoder; LightGBM; power theft detection

1. Introduction

Power safety is of great significance to social production and citizen daily life. In recent years, various types of power theft incidents have occurred frequently, causing huge economic losses to the state and power supply companies, and disrupting the power order of legal power consumers. In addition, illegal cross-connecting cables by power theft will keep the transformer at the end of the power grid overloaded for a long time, which directly affects the stability of normal power supply and reasonable power allocation by power supply companies, and also leads to great security risks. With the continuous emergence of new methods of stealing electricity [1], the methods of measuring equipment being privately modified have become more professional [2]. Along with the introduction and implementation of the national smart grid, while bringing convenience to power system control, it has also caused the amount of consumer electricity data to grow exponentially, and the annual data volume of large cities has already exceeded 10 billion. The explosion of professional power theft and power consumption data has increased the difficulty of power theft investigation and put forward higher requirements for current automatic power theft detection methods.

In order to solve the above problems, many scholars have used machine learning algorithms [3] to analyze the daily power consumption patterns of users to build classification models, including decision trees [4], random forest (RF) [5], support vector machines (SVM) [6], neural network (NN) [7], etc. R. Punmiya et al. [8] proposes a gradient boost theft detector (GBTD) based on the latest three gradient

boost classifiers (GBC): extreme gradient boost, categorical boosting and light gradient boosting method. Reference [9] propose a novel end-to-end solution to self-learn the features for detecting anomalies and frauds in smart meters using a hybrid deep neural network. Scholars such as Madalina Mihaela Buzau use all the information the smart meters record (energy consumption, alarms and electrical magnitudes) to obtain an in-depth analysis of the customer's consumption behavior [10,11] uses, with Long Short Term Memory (LSTM) units to process sequential power consumption data. Ramos, C.C.O. et al. [12] introduced the optimum-path forest classifier for a fast non-technical losses recognition. Ford, V. et al. [13] discusses a novel application of a machine learning technique for examining the energy consumption data to report energy fraud using artificial neural networks and smart meter fine-grained data. It can be seen that most power theft detection methods are trained by using pre-classified datasets containing markers. However, the amount of data of normal users in power data sets is generally large, while users of power theft only account for a very small portion. If the data set is used for training directly, a small number of power-stealing users may accumulate in the noise of normal users, which makes the classification result more biased towards normal users, resulting in a low detection rate of power stealing. Aiming at the problem of power imbalanced data sets, the power theft detection models in [1–13] either ignore the existence of the problem or use only some traditional methods, such as undersampling and oversampling. However, undersampling will lead to insufficient utilization of information on power theft data, and oversampling will cause the model to overfit. Some literatures have proposed a small sample oversampling technique (SMOTE) [14], which can alleviate the disadvantages of imbalanced data training to a certain extent, but this technique will lead to increased data noise. In other areas, such as credit card fraud imbalanced data sets [15], there are also cases of imbalanced data sets. Documents [16–20] describe various methods for dealing with imbalanced data sets. Aiming at the imbalanced power data set, this paper preprocesses the collected data and generates conditions using the conditional optimization Wasserstein criterion to generate a confrontation network (Conditional Wasserstein Generative Adversarial Network, CWGAN) to achieve a balanced power consumption data set.

Since the grid data is usually long-period data, the more data there is, the easier it is to extract features from it, but the more data, the greater the challenge to the detection model. Due to the user's random electricity consumption behavior, there is a large amount of noise in the data, and general machine learning methods are sensitive to changes in data input fluctuations [21]. Once there is wrong data in the data set, it will make it difficult for the model to obtain the expected data in the test set. Detection effect: In the face of grid data with a large number of random power consumption situations [22], the direct use of deep learning models often has insufficient generalization ability. Therefore, this paper designs a stacked convolution noise reduction auto-encoder based on power data to characterize the power consumption data. Extraction: Light Gradient Boosting Machine algorithm called LightGBM is used as the model output layer to output the theft detection results. This method uses a gradient decision machine based on a learning algorithm, which has fast training efficiency, low memory consumption, and high accuracy. Especially large-scale defective sparse data can highlight its advantages [23].

In summary, this paper proposes a noise reduction power-stealing detection model with a self-balancing data set. Firstly, an improved condition generation and confrontation network CWGAN is used to process the imbalanced data set. Generate power-stealing data through CWGAN, randomly generate the enhanced data set by mixing the generated data set with the original data set. Then, a multi-noise self-encoder convolution stacking combination model, stacked convolution noise-reduction auto-encoder SCDAE, is used to extract user electricity behavior characteristics. Finally, the LightGBM algorithm is used to detect power theft.

2. Theft Detection Model

Due to the variability of the electricity stealing behavior [24], the latest electricity consumption data is often used to update the electricity stealing model, so fast and accurate training of the model is also an integral part of the overall electricity stealing model. Therefore, in addition to the feature

extractor and classifier, the structure of the power-stealing model proposed in this paper adds a data pre-processor in the training phase. The overall structure is shown in Figure 1.

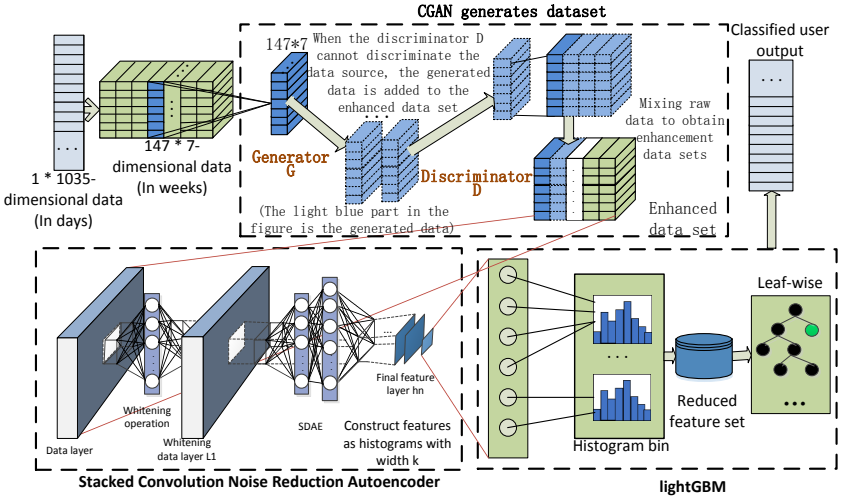


Figure 1. Electricity stealing model based on CWGAN's SCDAE-LightGBM model.

Theft detection is essentially two types of detection for user categories. To achieve two types of classification, you must first analyze the data and extract data features. However, due to the particularity of the power-stealing behavior, the number of users who steal power is usually much smaller than that of ordinary users, resulting in an extremely imbalanced data set. The extracted features will be biased towards normal user data [25], and the accuracy will be insufficient, so the imbalance of the original data must be solved first. The model's data preprocessor divides the daily power consumption vector into a weekly data matrix and uses the conditions of the Wasserstein criterion to generate an anti-network to balance power theft data. The training data preprocessor extracts the original training set data and trains the adversarial network according to a certain sampling rate. Through the generation and confrontation process of CWGAN, effective stealing data will be generated and mixed with the original training set to form an enhanced data set D. Subsequent models will be trained on the enhanced data set.

Due to the large amount and high dimensionality of power theft data, it is necessary to reduce the dimensions and feature extraction of the original data and optimize the neuron connection method to speed up model training with a limited amount of data. To this end, a convolutional stacked power feature extraction noise reduction autoencoder (SCDAE) is proposed. It takes the 147×7-dimensional enhanced data set as a training input, after normalized whitening operation and multi-layer noise reduction encoding, and finally outputs typical features that can effectively reconstruct the original data. Compared with the general SDAE, the SCDAE model has more stable and abstract feature extraction capabilities [20,26] and provides better data input for subsequent classifiers.

Finally, LightGBM is used to classify users who steal electricity. The algorithm does not need to perform one-hot feature coding [27]. Instead, it divides the power consumption data features extracted by SCDAE directly into the discrete domain to form a histogram. This not only reduces memory usage, but also improves classification efficiency. At the same time, it produces subtrees in a node expansion mode with a depth limit. When the number of splits of the decision tree is the same, it can reduce the occurrence of overfitting while obtaining higher model accuracy.

3. Processing of Imbalanced Data Sets

3.1. Stealing Data Set Balance Processing

In the theft detection, to achieve a good classification training effect, a certain amount of data is required. However, due to the particularity of electricity theft, analyzing a certain electricity

consumption data set [28] found that the sample of users who steal electricity is much smaller than the sample of normal users, and there are more than 3000 users who steal electricity from more than 40,000 users. Among them, the theft data with less sample types is called positive type, and the normal data with more sample types is called negative type samples. In addition, the randomness of normal users' electricity consumption behavior will cause data noise. As a result, a large number of negative samples are located in the category boundaries and overlapping areas of the data set, making it difficult to distinguish between power-stealing samples and normal samples.

For imbalanced data sets with noise, if a general classifier is used for training and learning, a "bias" phenomenon often occurs, that is, the classification is biased toward negative samples but it is difficult to achieve effective discrimination for positive samples. Therefore, this paper not only improves the data feature extraction ability from the algorithm level, but also optimizes the data of the classifier training process from the root source, balances the two types of data distribution and improves the detection rate of the test set.

In the data set of imbalanced electricity consumption by users, the positive type of power-stealing samples are relatively small, but to achieve effective power-stealing detection, the amount of information contained in the positive power-stealing samples is often more critical. Resampling the power-stealing sample can fully extract the data characteristics of the power-stealing user. In addition, the demand for training classifiers for data is often very large, and it is generally difficult to obtain enough positive samples for stealing electricity to constitute a data set and a test set. Therefore, it is necessary to generate positive samples in the classifier training stage to balance the data set. GAN can learn the original data distribution through the internal generation confrontation mechanism to generate samples. In this paper, the conditional generation countermeasure network proposed by Mirza [23] is used as the basis to introduce the conditional tagging quantity. The Wasserstein distance [24] is used instead of the KL (Kullback-Leibler) divergence to evaluate the conditions of the generated data distribution and the original data distribution, and an objective function of the training network that matches the characteristics of the power-stealing data is set. Finally, the two types of samples are processed differently, that is, the positive samples of power stealing are generated by the network, and the negative samples are undersampled to realize the balanced processing of the data set.

3.2. Design of CW Generation Counterattack Network for Stealing Data

The classic generative adversarial network consists of a generative model and a discriminant model. The generative model is denoted as G, and the generative model G is inputted with random data x to generate $G(x)$. Through training, the distribution function Φ_g of $G(x)$ approximately obeys the sample true distribution Φ_r, and the discriminant model D evaluates the degree of difference between the two distributions. During the model training process, the generated model G and the discriminant model D are updated alternately, making it difficult for the final discriminant model to distinguish the real data from the data generated by the generated model. The overall objective function is the following formula (Equation (1)).

$$V(G,D) = E_{x\sim\Phi_r} log(D(x)) + E_{x\sim\Phi_g} log(1-D(x)) \tag{1}$$

In the formula, E is used to represent the expectation function. The initial parameters of the discriminant function D (x) can be arbitrary values. The target discriminant model can be learned through data samples. The optimal discriminant model satisfies the formula (Equation (2)).

$$D^* = arg \max_D V(G,D) \tag{2}$$

When the optimal discriminant model exists, the objective of the optimal generative model is

$$G^* = arg \min_G D^* = arg \min_G \max_D V(G,D) \tag{3}$$

In order to generate electricity-stealing data using the generation counter network architecture, the following marks are introduced: The original data set is recorded as D_t, the original power consumption data is recorded as P_k, and the power consumption data generated by the generation model is recorded as $\widetilde{P}_k$. The loss function in the training process of the generated model is expressed in a matrix form:

$$Loss_{data} = \sum_{P_k \in D_t} \|P_k - \widetilde{P}_k\|_F^2 \tag{4}$$

Due to the discontinuous distribution of power consumption by power-stealing users, there is an optimal discriminator D^* between Φ_g and Φ_r, which can achieve target classification with a 100% probability, and the gradient is 0 [29] in the sampling data set, which causes the gradient to disappear during the neural network learning gradient approximation D^*, making it difficult to continue learning. At this time, the KL divergence used to evaluate the approximation of the two distributions tends to be infinite, and the JS divergence is a constant. Therefore, the Wasserstein distance is used instead of the KL divergence to measure $\Phi_g(x)$ and $\Phi_r(x)$ in the generation of power theft data.

$$W\left(\Phi_r, \Phi_g\right) \leq 2\left(E[\|\varepsilon\|_2^2]\right)^{1/2} + 2C\sqrt{JS(\Phi_{r+\varepsilon}\left\|\Phi_{g+\varepsilon}\right)} \tag{5}$$

where C is the minimum radius of the neighborhood containing Φ_g and Φ_r support sets, and ε represents noise. Arjovsky proved that when ϕ_θ represents the probability distribution of the function $g_\theta = (x;\theta)$, where g_θ is a generator function, and W (Φ_r, Φ_g) is also continuous when g_θ is continuous with respect to θ. K-R duality shows that W (Φ_r, Φ_g) satisfies formula

$$W\left(\Phi_r, \Phi_g\right) \leq \sup_{\|f\| \leq 1} E_{x\sim\Phi_r}[f(x)] - E_{x\sim\Phi_\theta}[f(x)] \tag{6}$$

The neural network parameter w is continuously updated through the back propagation principle, and the discriminant model objective function formula (Equation (7)) is obtained.

$$\max_{\omega \in W} E_{x\sim\Phi_r}[\varphi_w(x)] - E_{x\sim\Phi_\theta}[\varphi_w(x)] \tag{7}$$

Use $\varphi(x, w)$ approximation to approximate $f(x)$ in the objective function of the discriminant model. At the same time, in order to make Equation (7) meet Lipschitz's continuous assumption, it is necessary to make the weight of each update of the neural network generated in theft data within a certain range, generally −0.01 to 0.01. In addition, due to the high randomness and distribution uncertainty of the power-stealing data, it is often difficult to converge during the training process. To this end, user classification labels (normal users or power-stealing users) are introduced to form a condition generation adversarial network during training. The traditional classic adversarial generation network is transformed from free unsupervised learning to supervised learning that is relatively easy to converge. At this time, the objective function of the classic generative adversarial neural network is

$$\min_G \max_D V(G,D) = E_{x\sim\Phi_r} log(D(x|y)) + E_{x\sim\Phi_g} log(1 - D(x|y)) \tag{8}$$

Make the original data obey the distribution Φ_O, and generate the data obey the distribution Φ_G. Based on the above formulas X and Y, the objective function of the generation model of the anti-theft neural network is:

$$G^* = \arg\min_G \arg\max_D E_{p_k\sim\Phi_O}(D(p_k)) - E_{\widetilde{p}_k\sim\Phi_G}(D(\widetilde{p}_k)) + \sum_k \|p_k - \widetilde{p}_k\|_F^2 \tag{9}$$

Based on the above-mentioned theory, the classic GAN network is optimized and designed. The generative network and discriminative network objective function that are compatible with the

power-stealing data are constructed, and the CWGAN network power-stealing balance set algorithm can be designed by using the characteristics of user power consumption data.

4. Feature Extraction

4.1. Feature Extraction of Electricity Data

Because the collected user electricity consumption data is often measured by time, different time divisions constitute different data statistics. However, the time is too short, often the data features are vague, and it is difficult to determine the type of user. If the time is too long, it will occupy a lot of resources in the data storage and calculation process. The classifier efficiently provides the basis. Feature extraction methods represented by autoencoders [30] are more and more widely used in the field of power theft detection based on their strong generalization capabilities. In view of the large difference in power consumption between different users and the high degree of randomness, there is a lot of noise in the data of normal users and users who steal electricity. Here, convolution and denoising operations are introduced to design the stacked convolution noise reduction autoencoder (SCDAE) feature extraction of electrical data.

4.2. Electricity Data SCDAE Design

The ordinary autoencoder (AE) is a three-layer neural network structure, and the input data is reconstructed by the hidden layer h. Because of its ability to reconstruct the original data, the hidden layer h can be considered to have identifiable information. The stacked autoencoder (SAE) extracts the intrinsic data features at multiple levels while reconstructing the input data by setting multiple hidden layers h_i (i = 1, 2, ... n), and the hidden layer h_n contains all the data. The final feature, the encoder, is shown in formula (Equation (10)):

$$h_i = \sigma_e(w_i x_i + b_i) \tag{10}$$

where σ_e is the activation function of the encode, $x_i = h_{i-1}$, when i is 1; x_i is the daily electricity consumption vector of a single user divided by three weeks; and w_i' and b_i' are encoders neural network weights and biases. Due to the randomness of electricity consumption behavior of users, the noise existing in the electricity consumption data of stealing users and normal users will have a bad impact on feature extraction. To this end, a stacked noise reduction autoencoder (SDAE) is constructed, adding noise to the original power consumption data, reconstructing the original power consumption data from the noised power consumption data, and improving the generalization of the power consumption data extraction by SAE, as shown in formula (Equation (11)),

$$\widetilde{x_i} = \gamma(\widetilde{x_i}|x_i, \eta) \tag{11}$$

where is the distribution obeyed by adding noise to the original data, which is determined by the original data x_i and the parameter η.

Considering that the input is 147 × 7-dimensional power consumption, if a fully connected network is used, the training time will be too long and the training data demand will be too large. To this end, the convolution operation is introduced into SAE to form a stacked convolutional self-encoder (SCDAE), whose encoder is as shown in formula (Equation (12)):

$$h_i = \sigma_e(w_i \otimes \widetilde{x_i} + b_i) \tag{12}$$

where ⊗ is the convolution operator. In order to preserve the internal information of the power consumption data as much as possible, the pooling layer in the classic CNN is omitted in SCAE. At the same time, in order to prevent overfitting, a random neuron hiding operation is introduced in SCAE. This is to improve the network performance by blocking the neuron's joint action, specifically

introducing a Bernoulli function with probability p before the neuron to disable some neurons. Let $\zeta \sim B(p)$, then

$$\widetilde{h_i} = \zeta \cdot h_i \tag{13}$$

It should be noted that while training the encoder, the structure of a typical symmetric encoder decoder is no longer symmetrical due to the introduction of convolution operations, and it is necessary to continue undersampling in the decoder. The final feature decoder can be expressed as:

$$x_i = \sigma_d(w_i' \otimes \widetilde{h_i} + b_i') \tag{14}$$

where $\otimes$ is the undersampling operator, σ_d is the decoder activation function, w_i' and b_i' are the decoder convolution network weights and offsets, and χ_i is the original power data reconstructed by SCDAE. So the training SCDAE loss function can be defined as:

$$Loss_i = \sum \|x_i - \chi_i\|_F^2 + \Omega \tag{15}$$

where Ω is the regularization term that prevents the model from overfitting. Each layer of CDCD in SCDAE propagates features forward and gradients backward in a convolutional manner. It is difficult to converge by training the model directly in the form of SCDAE. To this end, the SCDAE is split into n-layer CDAE for stepwise training, and the feature vector obtained by the convolution of the upper CDAE will be used as the input of the lower CDAE. The training process is shown in Figure 2:

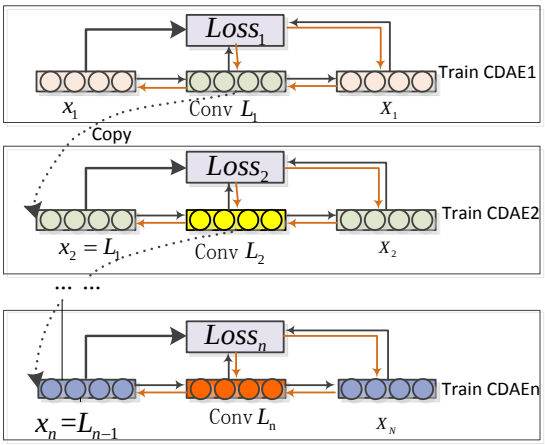


Figure 2. SCDAE model training process.

Using the above formula (Equation (15)) for the power consumption data, the SCDAE of the power data feature extraction can be obtained after the layered training and convergence according to the flow in Figure 2, which can be used directly for power feature extraction.

5. Theft Detection Based on LightGBM Classification

LightGBM training is performed on the feature data extracted by the SCDAE pair of enhanced power datasets along with the original user labels. LightGBM is a classification tool optimized based on GBDT. It uses feature contribution to train weak classifiers (decision trees) for selection. By constructing a histogram of width k to traverse the input data, the variance information gain is estimated according to Equation (16) to find the optimal segmentation point. [31]

$$\widetilde{v}_j(d) = \frac{1}{n}\left(\frac{\left(\Sigma_{v_i \in A_l} g_i + \frac{1-a}{b}\Sigma_{v_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\Sigma_{v_i \in A_r} g_i + \frac{1-a}{b}\Sigma_{v_i \in B_r} g_i\right)^2}{n_r^j(d)}\right) \tag{16}$$

In the formula, $A_l = \{x_i \in A : v_{ij} < d\}$, $A_r = \{x_i \in A : v_{ij} > d\}$, $B_l = \{x_i \in B : v_{ij} < d\}$, $B_r = \{x_i \in B : v_{ij} > d\}$, $n_l^i(d) = \sum I[v_i \in O : v_{ij} \le d]$, $n_r^i(d) = \sum I[v_i \in O : v_{ij} > d]$, A and B are the feature data sets sampled according to a certain percentage according to the gradient contribution size, O is the feature data set on the fixed node of the decision tree, and a and b are constants. Use leaf-wise growth strategies with depth limitation for acceleration. By setting $C_{a,b} = \frac{1-a}{\sqrt{b}} max_{v_i \in A^C} |g_i|$, $D = max\left(\overline{g}_l^j(d), \overline{g}_r^j(d)\right)$, we can prove the maximum value of the approximation error ε (d) of the classification model:

$$C_{a,b}^2 \cdot ln\frac{1}{\sigma} \cdot max\left\{\frac{1}{n_l^j(d)}, \frac{1}{n_r^j(d)}\right\} + 2DC_{a,b}\sqrt{\frac{ln\frac{1}{\sigma}}{n}} \tag{17}$$

Among them, n is the dimension of the data set, and σ is the probability constant. For the public third-party LightGBM library, it is necessary to set core parameters such as the learning rate of 0.1 and the number of leaves of a single decision tree of 31. Control parameters such as the minimum data amount of a single leaf of 15, and (GOSS) large and small gradient retention ratios of 0.2 and 0.1 are input and output parameters. The maximum number of features in a single cabinet is 255, and the minimum amount of data is 5, etc., to achieve the normal classification of power consumption feature data and power theft [32].

The training and detection process of the theft detection model is shown in Figure 3 below. Fifty percent, 20%, and 30% of the original data are randomly selected to form the model training set, validation set, and test set. The training process consists of three parts: the original data training set and the validation set. After training CWGAN, an enhanced data set is generated for training and increasing SCDAE and LightGBM. The test process mainly consists of two parts, the feature extractor and LightGBM classifier generated by SCDAE cropping.

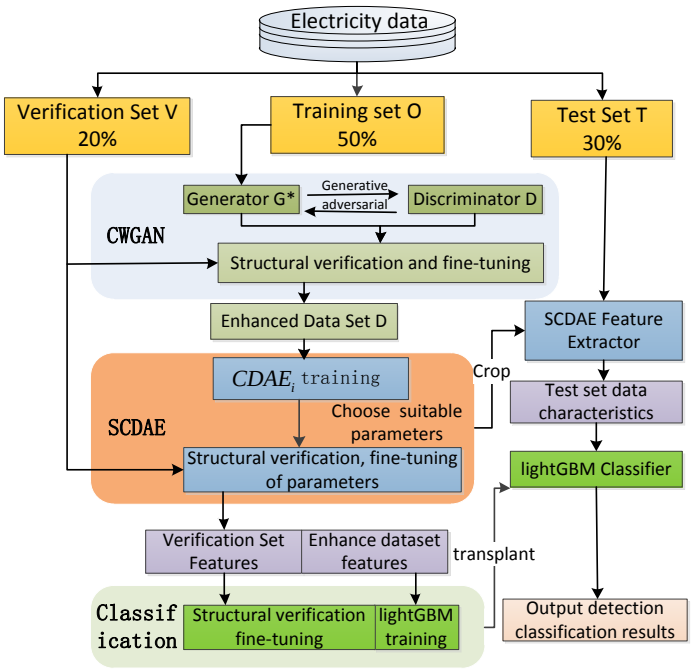


Figure 3. SCDAE-LightGBM model stealing electricity flow chart.

6. Experimental Results and Analysis

The CWGAN-SCDAE-LightGBM model (CSL model) proposed in this paper will use the public electricity consumption data set published by State Grid Corporation of China for experiments. This data set [28] contains the electricity consumption data of 42,372 electricity customers in 1035 days

(1 January 2014 to 31 October 2016). This experiment was performed under python 3.6 using the public LightGBM framework.

6.1. Evaluation Index

In classification detection, it is not possible to evaluate the performance of the classifier on an imbalanced data set based on accuracy alone, and the imbalanced data set used has certain requirements for the sensitivity and specificity of the detection, and multiple indicators need to be used. Typical evaluation indicators include recall, specificity, accuracy, F-number, and Accuracy.

6.2. Data Set Balance Verification

Based on the data analysis and the second section of the adversarial generation network design, the power-stealing balance data algorithm is as follows:

- Step 1: Count the number of stolen users n_1 and the number of normal users n_2 in the original data set D_0, and set the undersampling rate α ($0 < \alpha < 1$). Randomly undersampling normal users and mixing them with the original data set stealing users constitutes a new data set D_1 for training of the CWGAN network;
- Step 2: Train the CWGAN network;
- Step 3: Use the trained CWGAN network to generate new stealing data to form a new stealing sample set D_2. Finally, it is merged with the original data set D_0 to generate the final balanced power consumption data set. The Algorithm 1 pseudo code is as follows:

Algorithm 1 CWGAN algorithm for generating steal data

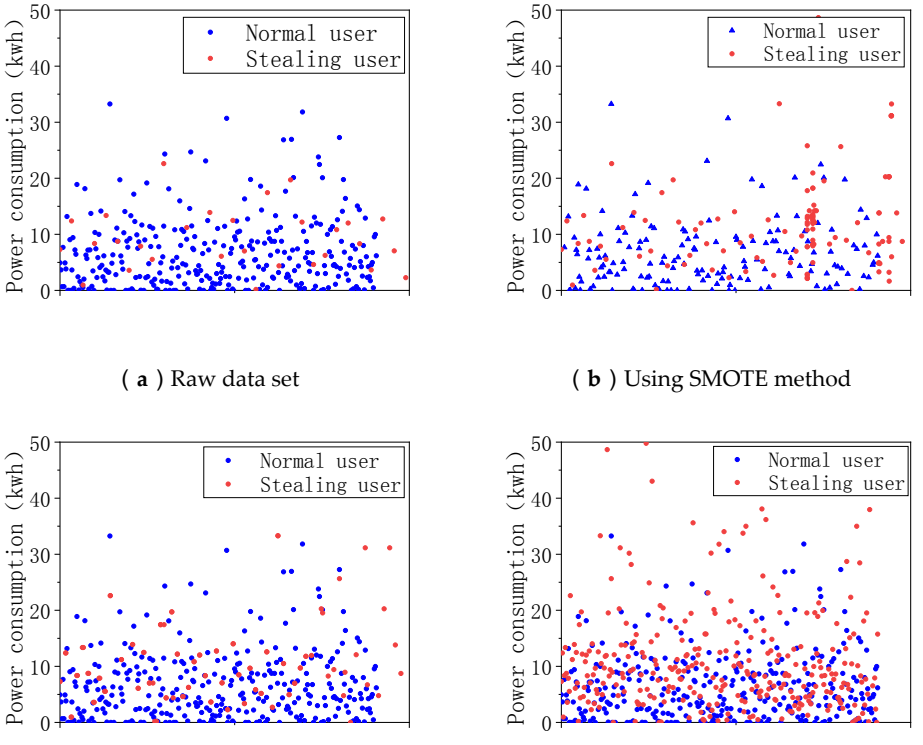
Input: unbalanced data set D0, sampling rate is $\alpha(\alpha < 1)$, number of iterations epoch1, epoch2 / * Construct training data set * /
Step 1: Calculate n1 and n2 according to D0
Step 2: Undersampling of negative samples constitutes D, the number is $n_1 + \alpha n_2$ / * construct training data set * /
Step 3: Train CWGAN model based on data set D

1. Initialization
2. For k=1,2, … ,epoch1 do
3. For j=1,2, … ,epoch2 do
4. Extract a sample of capacity m from dataset D1
5. $\{\{p_1, y_1\},\{p_2, y_2\},\{p_3, y_3\} \ldots \{p_m, y_m\}\}$
6. Random noise with a capacity of m formed from the amplified vector of U (0,1)
7. Update discriminant model $\{\{\widetilde{p}_1, \widetilde{y}_1\}),\{\widetilde{p}_2, \widetilde{y}_2\},\{\widetilde{p}_3, \widetilde{y}_3\} \ldots \{\widetilde{p}_m, \widetilde{y}_m\}\}$,
$g_w \leftarrow D_w\left[\frac{1}{m}\sum_{i=1}^{m} f_w(p_i, \widetilde{p}_i) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(\widetilde{p}_i), \widetilde{p}_i)\right]$, $w \leftarrow w + \alpha Adam(w, g_w)$
8. End for
9. / * Update generation model * / $g_\theta \leftarrow -\Delta_\theta\left[\frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(\widetilde{p}_i)) + \|g_\theta(\widetilde{p}_i) - p_i\|_F^2\right]$,
$\theta \leftarrow \theta + \alpha Adam(\theta, g)$
10. End for

/ * Generate positive samples based on generative adversarial network * /
Step 4: The random noise with a vector capacity of $n = n_2 - n_1$ extracted from U (0,1) is amplified into data $\{\{\widetilde{p}_1, y_0\}, \{\widetilde{p}_2, y_0\}, \{\widetilde{p}_3, y_0\} \ldots \{\widetilde{p}_n, y_0\}\}$. As an input to generate the network G, n positive sample sets D_2 can be generated.
Step 5: Mix the generated positive sample set D_2 with the original data set D_0 to obtain a balanced data set D_3
Step 6: Output data set D_3
Output: balanced data set

The Adam optimizer is used in the power-stealing CW confrontation generation network. One-Hot Encoding is used to encode the user type and add it to the condition variant. The dimension is two. The generation model G uses a classic convolutional neural network structure, and the discrimination model D uses a single hidden layer neural network whose activation function is Relu. Since the original data is the daily power consumption of a single user for 3 years, the input of the power-stealing generation model is a dimension of 1 × 1035. By weekly partitioning into 147 × 7-dimensional, (0,1) uniformly distributed random noise and label variable y, the final output is the same 147 × 7-dimensional power-stealing data sample as the real stealing user. The discrimination model D also inputs 147 × 7-dimensional power consumption data, outputs the probability of real data with a dimension of 1, and the output layer activation function is a sigmod function.

The results of unbalanced data set processing for the same data set using the above-mentioned power-stealing CWGAN, CGAN, and SMOTE methods are shown in Figures 4 and 5. The abscissa of Figure 4 is a sequence of data numbers, of which Figure 4a is the original power consumption data, and Figure 4b is the new power-stealing data generated by the K-nearest neighbor algorithm in the local area by the SMOTE method. It can be clearly seen that the generation and aggregation of power-stealing data are quite different from the original data distribution. Figure 4c uses classic CGAN to generate data. Because the original data is an unbalanced data set, the generated data further exacerbates the imbalance of the original data. Figure 4d is the data generated by CWGAN. The distribution of the generated data is similar to the original data distribution, and the amount of the two types of data is balanced to a certain degree. It is superior to other methods in terms of quantity and quality and guarantees the model of power theft well-trained. Figure 5 is the average daily power consumption of the original data and generated data over 365 days. It can be seen that the generated model approximates the original data well, and the generated data also effectively filters out the abnormalities caused by the default values in the original data noise.



(**a**) Raw data set

(**b**) Using SMOTE method

(**c**) CGAN generates power theft data

(**d**) CWGAN generates power theft data

Figure 4. Comparison of different methods for processing imbalanced data sets.

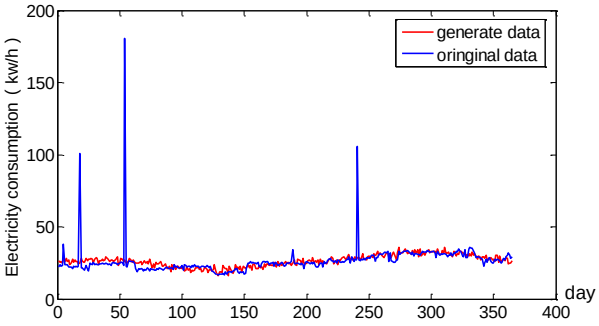


Figure 5. Comparison of CWGAN processing data set fit.

The loss value (Loss) and Matthews correlation coefficient value (MCC) of the CWGAN preprocessed data set and the original data set for direct SCDAE + LightGBM training are shown in Figure 6. The MCC value is a typical indicator used to evaluate an imbalanced data set. When it is 1, it means a completely accurate prediction. As shown in Figure 6a, when the CWGAN balanced data set is not used, as the number of trainings increases, the inflection point appears in the test loss, that is, the model appears to overfit. After using CWGAN to balance the data set, as shown in Figure 6b, the occurrence of overfitting is effectively alleviated. Similarly, the MCC values are shown in Figure 6c,d. After using the CWGAN balanced data set, there is a significant improvement.

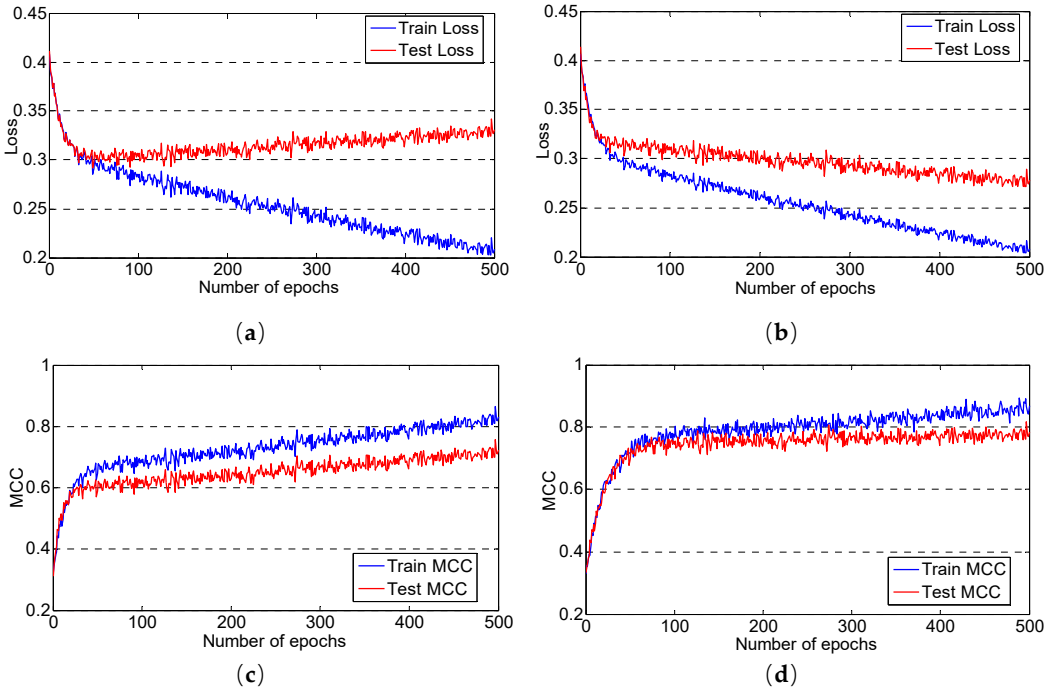


Figure 6. (**a**) and (**b**) is curve of loss, (**c**) and (**d**) is curve of MCC value.

6.3. SCDAE Feature Extraction Verification

Feature extraction uses multiple noise reduction autoencoders to be stacked in a convolutional manner. Features are extracted from user electricity data, and the stacked structure can be determined layer by layer from bottom to top. The goal of training SCDAE is to minimize the overall reconstruction loss function E_{SCDAE}, and the reconstruction loss function of the i-th CDAE is L_i, then the overall reconstruction loss function of SCDAE can be defined as $E_{SCDAE} = 1 - \prod_i (1 - L_i)$, and E_{SCDAE} can be a smaller value (25%). The number of stacked layers of SCDAE is determined by adjusting the dimension of the output features of each hidden layer. For the L_i of each CDAE in the validation set, the size of

the output feature dimension of the hidden layer is shown in Figure 7. In Figure 7a, the turning point of the slope of the loss value is (116, 0.097), that is, when the dimension of the output feature value of the first hidden layer is 116, the loss value will hardly increase with the increase of the feature value dimension. Similarly, for two to four auto-encoders, take 73, 36 and 17, respectively. Through the joint test fine-tuning, the E_{SCDAE} can be close to 25%, and the SCDAE can be determined to be a four-layer stacked structure.

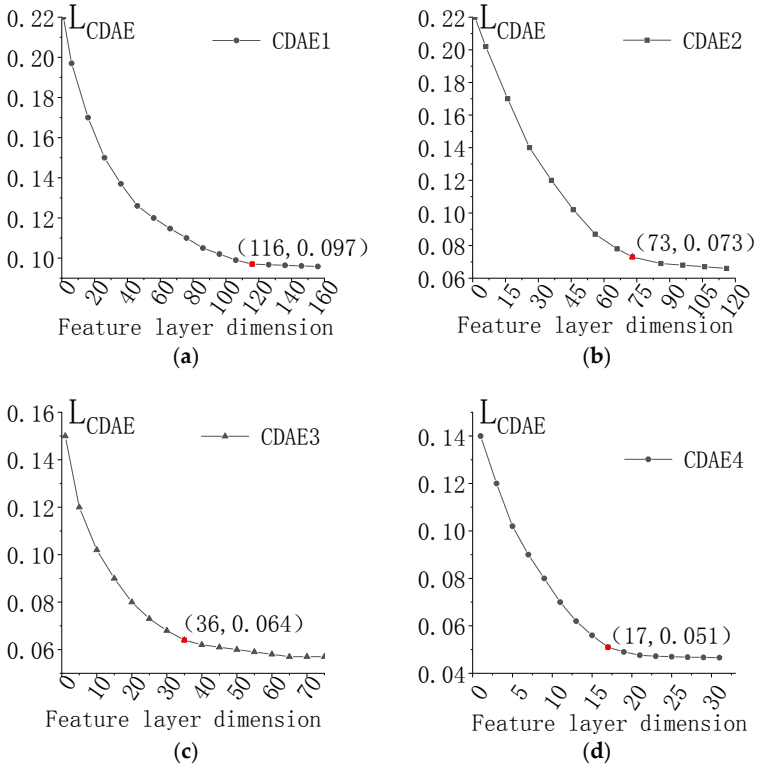


Figure 7. (**a**), (**b**), (**c**), (**d**) are the loss values of the first layer, the second layer, the third layer, and the fourth layer of the SCDAE training process.

To determine the epoch value for feature extraction training, all labeled samples need to be trained. Too small or too large an epoch values can cause underfitting or overfitting. Figure 8: after 70 epochs, both the AUC score and F1 score decreased slightly, and SCDAE overfitting occurred. The AUC score and F1 score of the 50th epoch reached 0.9738 and 0.8773, respectively, so the epoch value of the experiment was set to 50.

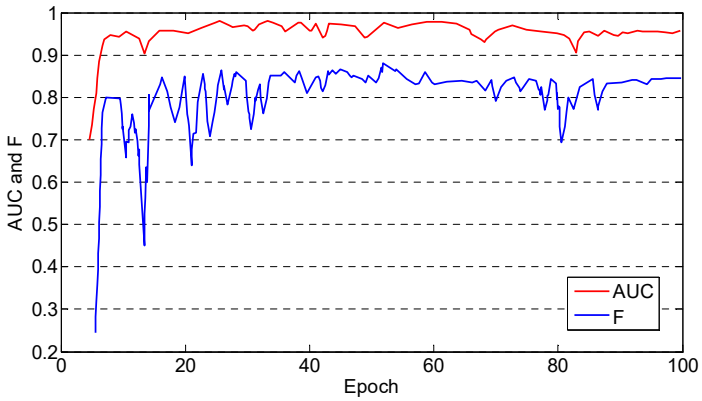


Figure 8. SCDAE convergence analysis.

Figures 9 and 10 show the comparison of the test law of the theft detection model CSL (CWGAN-SCDAE-LightGBM) designed in this paper with several existing advanced steal-the-power models in the ROC curve and during a certain number of iterations. From the comparison of the ROC curves of several methods, we can see that the ROC curve of CSL is closest to the upper left corner, the area under the curve is the largest, and the detection effect is the best.

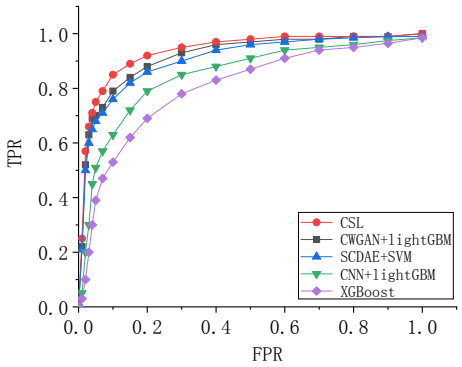


Figure 9. Comparison of ROC curves of each model.

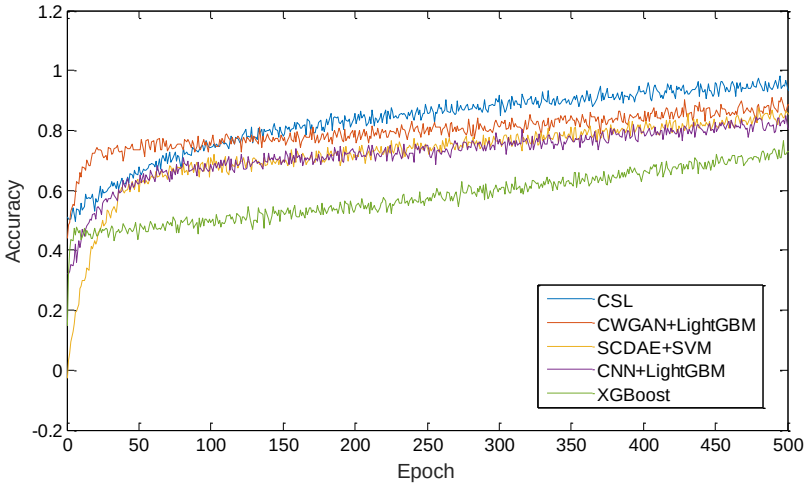


Figure 10. Comparison of iteration curves of each model.

It can be seen in Figure 10 that with the increase of training iterations, the detection algorithm of the CWGAN enhanced data set is added at an early stage, and the accuracy rate is higher than that without the enhanced data set. In addition, when the SCDAE training tends to converge, the CSL model has a clear advantage over other detection algorithms in terms of test accuracy. Most of the other methods have the same number of repetitions at about 85%, and the CSL model can finally reach a stability rate of more than 90% on the test set, up to 97.6%.

To sum up, this paper proposes two improvements in the detection of power theft: CWGAN handles unbalanced data sets and SCDAE power feature extraction. After training convergence, it can effectively improve the performance of ROC and accuracy on the test set, and increase the accuracy of power theft detection to more than 90%.

7. Conclusions

The new CSL power-stealing detection model proposed in this paper deals with unbalanced data sets through CWGAN. The generated power-stealing data is mixed with the original data to form an enhanced data set for subsequent feature extractor training. Experiments show that the model not only makes the model converge quickly, but the MCC value is higher under the same epoch and the final MCC value of the model is increased by 0.1 to 0.8 compared to the case without data balancing operation.

In addition, in view of the interference noise phenomenon in the user's electricity data set, a comprehensive convolution and encoder idea is proposed to extract the power-stealing feature extractor SCDAE. On the one hand, the noise in the data set is filtered by the noise reduction auto-encoder to avoid the adverse impact of the noise data. On the other hand, the noise reduction autoencoders are stacked by convolution to extract more typical features in the theft of electricity, laying a good foundation for subsequent classification detection. Finally, through experiments comparing the training and test results obtained on different data-stealing detection models on the same data set, it is concluded that the CSL power-stealing detection model has improved the typical indicator accuracy from about 85% to more than 90% compared to the common power-stealing detection model, which has obvious advantages. The current work of this paper still has certain limitations, which involve the need to adjust a large number of parameters when using the LightGBM library. The LightGBM model parameters play an important role in the final effect of the power-stealing model. In this paper, only manual debugging is used to implement some parameters, and LightGBM's advantages in the classification of power-stealing features have not been fully utilized. Parameter adaptive adjustment methods can be added in the future to achieve the optimal approximation of model parameters.

Author Contributions: Data curation, H.L.; Methodology, Z.L.; Supervision, Y.L.; Writing – original draft, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

G	Generative model
D	Discriminant model
$V(G,D)$	Overall objective function
Φ_g	Distribution function of G (x)
Φ_r	Sample true distribution
E	Expectation function
D^*	Optimal discriminant model
G^*	Optimal generative model
D_t	Original data set
P_k	Original power consumption data
$Loss_{data}$	Loss function
W (Φ_r, Φ_g)	Wasserstein distance
C	Minimum radius
ε	Noise
ϕ_θ	Probability distribution of g_θ
g_θ	Generator function
h_i	Multiple hidden layers
h_n	Hidden layer
σ_e	Activation function of the code
$\otimes$	Undersampling operator
σ_d	Decoder activation function
x_i	Daily electricity consumption vector
w_i'	Encoder Neural network weights
b_i'	Encoder Neural network biases
k	Histogram width
n	Dimension of the data set
σ	Probability constant
A,B	Feature data sets
v_i	Vector with dimension s in space X s
O	Feature data set
g_i	Negative gradients of the loss function

References

1. Jokar, P.; Arianpoo, N.; Leung, V.C.M. Electricity theft detection in AMI using customers' consumption patterns. *IEEE Trans. Smart Grid* **2017**, *7*, 217–227. [CrossRef]
2. Zheng, K.; Chen, Q.; Wang, Y.; Kang, C.; Xia, Q. A Novel Combined Data-Driven Approach for Electricity Theft Detection. *IEEE Trans. Ind. Inform.* **2019**, *15*, 1809–1819. [CrossRef]
3. Messinis, G.M.; Rigas, A.E.; Hatziargyriou, N.D. A Hybrid Method for Non-Technical Loss Detection in Smart Distribution Grids. *IEEE Trans. Smart Grid* **2019**, *10*, 7080–7091. [CrossRef]
4. Jindal, A.; Dua, A.; Kaur, K.; Singh, M.; Kumar, N.; Mishra, S. Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Trans. Ind. Inform.* **2017**, *12*, 1005–1017. [CrossRef]
5. Shuan, L.; Han, Y.; Yao, X.; Ying, C.S.; Wang, J.; Zhao, Q. Electricity Theft Detection in Power Grids with Deep Learning and Random Forests. *J. Electr. Comput. Eng.* **2019**, 1–12. [CrossRef]
6. Lu, F.; Ding, X.; Yin, X.; Chen, H.; Wang, Y. Support Vector Machine Stealing Identification Method Based on Sample Optimization Selection. *Comput. Meas. Control* **2018**, *26*, 223–226.
7. Zheng, Z.; Yang, Y.; Niu, X. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1707–1715. [CrossRef]
8. Punmiya, R.; Choe, S. Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing. *IEEE Trans. Smart Grid* **2019**, *10*, 2326–2329. [CrossRef]
9. Buzau, M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Hybrid Deep Neural Networks for Detection of Non-Technical Losses in Electricity Smart Meters. *IEEE Trans. Power Syst.* **2020**, *35*, 1254–1263. [CrossRef]
10. Buzau, M.M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. *IEEE Trans. Smart Grid* **2019**, *10*, 2661–2670. [CrossRef]
11. Chatterjee, S.; Archana, V.; Suresh, K.; Saha, R.; Gupta, R.; Doshi, F. Detection of non-technical losses using advanced metering infrastructure and deep recurrent neural networks. In Proceedings of the 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), Milan, Italy, 13 July 2017.
12. Ramos, C.C.O.; Souza, A.N.; Papa, J.P.; Falcao, A.X. Fast Non-Technical Losses Identification Through Optimum-Path Forest. In Proceedings of the 2009 15th International Conference on Intelligent System Applications to Power Systems, Curitiba, Brazil, 11 December 2009.
13. Ford, V.; Siraj, A.; Eberle, W. Smart grid energy fraud detection using artificial neural networks. In Proceedings of the 2014 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG), Orlando, FL, USA, 19 January 2014.
14. Pan, T.; Zhao, J.; Wu, W.; Yang, J. Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Inf. Sci.* **2020**, *512*, 1214–1233. [CrossRef]
15. Wang, Y. Identification of credit card fraud under imbalanced data. *Commun. World* **2018**, *25*, 219–220.
16. Depuru, S.S.S.R.; Wang, L.; Devabhaktuni, V.; Nelapati, P. A hybrid neural network model and encoding technique for enhanced classification of energy consumption data. In Proceedings of the 2011 IEEE Power and Energy Society General Meeting, San Diego, CA, USA, 24–29 July 2011.
17. Glauner, P.O. Large-scale detection of non-technical losses in imbalanced data sets. In Proceedings of the Seventh IEEE Conference on Innovative Smart Grid Technologies (ISGT 2016), Minneapolis, MN, USA, 6–9 September 2016.
18. Georgios, D.; Fernando, B. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst. Appl.* **2018**, *91*, 464–471.
19. Fiore, U.; Santis, A.D.; Perla, F. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. *Inf. Sci.* **2017**, *479*, 448–455. [CrossRef]
20. Ben Said, A.; Mohamed, A.; Elfouly, T. Deep learning approach for EEG compression in mHealth system. In Proceedings of the 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, Spain, 27–30 June 2017.
21. Wei, L.; Sundararajan, A.; Sarwat, A.I.; Biswas, S.; Ibrahim, E. A distributed intelligent framework for electricity theft detection using benford's law and stackelberg game. In Proceedings of the 2017 Resilience Week (RWS), Wilmington, DE, USA, 18–22 September 2017.

22. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, 3149–3157.
23. Wang, C.; Li, P.; Yu, H. Analysis and application of new forms of smart distribution network and its flexibility characteristics. *Autom. Electr. Power Syst.* **2018**, *42*, 13–21.
24. Zhao, H.; Shi, H.; Wu, J.; Chen, X. Research on Imbalanced Learning Based on Conditional Generative Adversarial Network. *Control Decis.* 1–10.
25. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
26. Buckman, J.; Roy, A.; Raffel, C.; Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 23 February 2018.
27. Available online: https://github.com/henryRDlab/ElectricityTheftDetection (accessed on 22 September 2018).
28. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *Comput. Sci.* **2014**, *2*, 2672–2680.
29. Zhao, W.; Meng, Q.; Zeng, M.; Qi, P.-F. Stacked Sparse Auto-Encoders (SSAE) Based Electronic Nose for Chinese Liquors Classification. *Sensors* **2017**, *17*, 2855. [CrossRef] [PubMed]
30. Hu, T.; Guo, Q.; Sun, H. Power theft detection based on stacked decorrelation autoencoder and support vector machine. *Autom. Electr. Power Syst.* **2019**, *43*, 119–127.
31. Zou, H.; Zhou, Y.; Yang, J.; Jiang, H.; Xie, L.; Spanos, C.J. DeepSense: Device-free Human Activity Recognition via Autoencoder Long-term Recurrent Convolutional Network. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018.
32. Ju, Y.; Sun, G.; Chen, Q.; Zhang, M.; Zhu, H.; Rehman, M.U. A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting. *IEEE Access* **2019**, *7*, 28309–28318. [CrossRef]