Article

Detection of Electricity Theft Behavior Based on Improved Synthetic Minority Oversampling Technique and Random Forest Classifier

Zhengwei Qu ^{1,*}, Hongwen Li ¹, Yunjing Wang ¹, Jiayi Zhang ¹, Ahmed Abu-Siada ² and Yunxiao Yao ³

¹ Key Laboratory of Power Electronics for Energy Conservation and Drive Control, Yanshan University, Qinhuangdao 066004, China; lhw@stumail.ysu.edu.cn (H.L.); wyj@ysu.edu.cn (Y.W.); zjx@stumail.ysu.edu.cn (J.Z.)

² School of Electrical Engineering Computing and Mathematical Sciences, Curtin University, Perth WA 6102, Australia; A.AbuSiada@curtin.edu.au

³ State Grid Hubei DC Operation and Maintenance Company, Yichang 443008, China; yunx_yao@163.com

* Correspondence: zhengwei.qu@ysu.edu.cn; Tel.: +86-0335-807-4883

Received: 13 March 2020; Accepted: 15 April 2020; Published: 19 April 2020

Abstract: Effective detection of electricity theft is essential to maintain power system reliability. With the development of smart grids, traditional electricity theft detection technologies have become ineffective to deal with the increasingly complex data on the users' side. To improve the auditing efficiency of grid enterprises, a new electricity theft detection method based on improved synthetic minority oversampling technique (SMOTE) and improve random forest (RF) method is proposed in this paper. The data of normal and electricity theft users were classified as positive data (PD) and negative data (ND), respectively. In practice, the number of ND was far less than PD, which made the dataset composed of these two types of data become unbalanced. An improved SOMTE based on K-means clustering algorithm (K-SMOTE) was firstly presented to balance the dataset. The cluster center of ND was determined by K-means method. Then, the ND were interpolated by SMOTE on the basis of the cluster center to balance the entire data. Finally, the RF classifier was trained with the balanced dataset, and the optimal number of decision trees in RF was decided according to the convergence of out-of-bag data error (OOB error). Electricity theft behaviors on the user side were detected by the trained RF classifier.

Keywords: smart grid; nontechnical losses; electricity theft detection; synthetic minority oversampling technique; K-means cluster; random forest

1. Introduction

Power losses are usually divided into technical losses (TLs) and nontechnical losses (NTLs) [1]. NTLs refer to the power loss during the transformation, transmission, and distribution process and are mainly caused by electricity theft on the user side [2]. In most countries, electricity theft losses (ETLs) account for the predominant part of the overall electricity losses [3], and are mainly taking place in the medium- and low-voltage power grids. ETLs can cause serious problems, such as loss of revenue of power suppliers, reducing the stability, security, and reliability of power grids and increasing unnecessary resources consumption. In India, ETLs were valued at about US\$4.5 billion [4], which is still rising year by year. ETLs are reported to reach up to 40% of the total electricity losses in countries such as Brazil, Malaysia, and Lebanon [5]. ETLs of some provinces in China reached about 200 million kWh, with an overall cost of 100 million yuan. As reported in [6], the losses due to electricity theft reached about 100 million Canadian dollars every year with a power loss that could

have supplied 77,000 families for one year. The annual income loss caused by electricity theft in the United States accounted for 0.5% to 3.5% of the total income [7,8]. Therefore, the research on advancing electricity theft detection techniques has become essential due to its significance for energy saving and consumption reductions [9].

In the past an electric meter packaging, professional electric meters, and bidirectional metering conventional methods were adopted to deal with electricity theft [10,11]. Today, electricity theft detection methods rely on classifying the data collected by smart meter measurement system. Classification of the electricity theft and normal behaviors is conducted through data analysis [12].

The modern methods for electricity theft detection mainly include state-based analysis, game theory, and classification [13].

State-based detection schemes employ specific devices to provide high detection accuracy. A novel hybrid intrusion detection system framework that incorporates power information and sensor placement has been developed in [14] to detect malicious activities such as consumer attacks. In [15], an integrated intrusion detection solution (AMIDS), was presented to identify malicious energy theft attempts in advanced metering infrastructures. AMIDS makes use of different information sources to gather a sufficient amount of evidence about an ongoing attack before marking an activity as a malicious energy theft. In [16], state estimation was used to determine electricity theft users. When there was a difference between the voltage of the state estimation and the voltage of the measured node, the breadth-first search was conducted from the root node of the distribution network, and the magnitude of the difference at the same depth was compared to locate electricity theft users. In [17], in order to detect and localize the occurrence of theft in grid-tied microgrids, A Stochastic Petri Net (SPN) with a low sampling rate was used to first detect the random occurrence of theft and then localize it. The detection was based on determining the accurate line losses through (Singular Value Decomposition) SVD, which led to the recognition of theft in grid-tied MGs. State-based detection schemes will bring additional investment required for monitoring systems, including equipment costs, system implementation costs, software costs, and operation/training costs. In [18], it investigated energy theft detection in microgrids, considering a realistic model for the microgrid's power system and the protection of users' privacy. It proposed two energy theft detection algorithms capable of successfully identifying energy thieves. One algorithm, called centralized state estimation algorithm based on the Kalman filter (SEK), employed a centralized Kalman filter. However, it could not protect users' privacy and did not have very good numerical stability in large systems with high measurement errors. The other one, called privacy-preserving bias estimation algorithm (PPBE), was based on two loosely coupled filters, and could preserve users' privacy by hiding their energy measurements from the system operator, other users, and eavesdroppers. However, state-based detection schemes employ specific devices to provide high detection accuracy, which, however, come with the price of extra investment required for the monitoring system including device cost, system implementation cost, software cost, and operating/training cost.

Another approach for theft detection is based on game theory. Reference [19] formulated the problem of theft detection as a game between an illegitimate user and a distributor. The distributor wants to maximize the probability of theft detection while illegitimate users or thieves want to minimize the likelihood of being caught by changing their Probability Density Functions (PDFs) of electricity usage.

Classification-based methods include expert systems and machine learning. Expert systems are based on computer models trained by human experts to deal with complex problems and draw the same conclusions as experts [20]. The expert system of electricity theft detection based on specific decision rules was initially used. With the rapid development of artificial intelligence technology, machine learning enables computers to learn decision rules from training. Therefore, in recent years, machine learning has become the main research direction of electricity theft detection [21]. In [22], it explored the possibilities that exist in the implementation of Machine-Learning techniques for the detection of nontechnical losses in customers. The analysis was based on the work done in collaboration

with an international energy distribution company. It reported on how the success in detecting nontechnical losses can help the company to better control the energy provided to their customers, avoiding a misuse, and, hence, improving the sustainability of the service that the company provides. Reference [23] provides a novel knowledge-embedded sample model and deep semi-supervised learning algorithm to detect NTL by using the data in smart meter. It first analyzed the characteristic of realistic NTL, and designed a knowledge-embedded sample model referring to the principle of electricity measurement. Next, it proposed an autoencoder based semi-supervised learning model.

In [24], fuzzy logic and expert system were combined to integrate human expert knowledge into the decision-making process to identify electricity theft behavior. A grid-based local outliers algorithm was proposed in [25] to achieve unsupervised learning of abnormal behavior of power users. This method mapped variables features into two-dimensional plane by factor analysis (FA) and principal component analysis (PCA). The dimensionality of data and the operation cost of outlier factor algorithm were reduced by grid technology. In [26], electricity theft detection method based on probabilistic neural network was employed to detect two types of illegal consumption.

In [27], clustering analysis was carried out firstly to reduce the number of data to be analyzed, then the suspected users were found through neural network. In [28], the extreme learning machine (ELM) was used to identify the weight between the hidden and output layer, and electricity theft was detected through the measured data of the meter. In [29], a five-joint neural network was trained with power data comprising 20,000 customers and achieved considerable accuracy. SVM-FIS method was proposed in [30], which could reduce the calculation complexity and improve the detection accuracy by combining the fuzzy inference system with the SVM. In [31], a data-based method was proposed to detect sources of electricity theft and other commercial losses. Prototypes of typical consumption behavior were extracted through clustering the data collected from smart meters.

For an unbalanced dataset, intelligent algorithms tend to favor positive data (PD) in the training process and ignore the important information contained in a few negative data (ND), which may reduce the detection accuracy [32]. Therefore, optimizing the unbalance of the dataset plays an important role for improving the efficiency and accuracy of the algorithm. Data-oriented methods mostly rely on existing and validated cases of fraud either for training or validation. However, since frauds are scarce, it is difficult to obtain these samples, unless another Fraud detection methods such as unsupervised detection, or a manual inspection campaign are used [33].

The theory of unbalanced data processing has been widely used in the fields of network fraud identification, network intrusion detection, medical diagnosis, and text classification. However, it is still rarely used in electricity theft detection. Reference [34] introduced consumption pattern-based energy theft detector (CPBETD), a new algorithm for detecting energy theft in advanced metering infrastructure (AMI). CPBETD relies on the predictability of customers' normal and malicious usage patterns, and it addresses the problem of imbalanced data and zero-day attacks by generating a synthetic attack dataset, benefiting from the fact that theft patterns are predictable. In [35], a methodology was proposed to improve the performance and evaluation of supervised classification algorithms in the context of NTL detection with imbalanced data. The main contributions of our work lie in two aspects: (1) The strategies considered to counteract the effects of imbalanced classes, and (2) an extensive list of performance metrics detailed and tested in the experiments.

A comprehensive detection method for NTLs of unbalanced power data was proposed in [36], which contained three detection models (Boolean rule, fuzzy logic, and support vector machine). Reference [37] proposed two undersampling methods for the classification of unbalanced data, easy ensemble (EE) algorithm and balance cascade (BC) algorithm. The above two methods exhibited high computation and implementation complexity. In [38], a one-sided selection (OSS) method was proposed for dealing with unbalanced data. In [39], a KNN-near miss method based on the K-nearest neighbor (KNN) undersampling method was proposed. In [40], an oversampling method, called synthetic minority oversampling technique (SMOTE), was adopted, which achieved excellent results in the processing of unbalanced data and effectively solved the problem of excessive random sampling.

However, the algorithm had certain blindness in the selection of neighbors, did not consider the distribution of data when generating new data, and had strong marginality.

Reference [41] reported that, compared with single strong decision tree, weak decision tree had high computational efficiency. In addition, considering the weight sparsity of weak classifier, the recognition rate of the cluster could be further improved [42]. In [43], decision trees were used for NTL detecting and the algorithms were tested with real a database of Endesa Company. In addition, random forest classifier (RF) can save resources and computational time because the multiple decision trees run in parallel. Moreover, each decision tree can achieve random selection of data and attributes without overfitting [44].

In summary, considering the shortcomings of existing electricity theft detection methods and the unbalance of user data, a method for electricity theft behaviors detection was proposed based on improved SMOTE and random forest classifier in this paper. The main contributions of this paper can be listed as follows.

- (1) Considering the high unbalance of power user-side dataset and the shortcomings of existing methods, a new K-SMOTE method was proposed to deal with the unbalanced initial datasets. The proposed method can reduce the impact of detection accuracy caused by unbalanced data.
- (2) Combined with the unbalanced data, considering the limitation of setting decision tree in RF algorithm, the improved random forest classifier was applied to detect electricity theft behaviors. The efficiency of power theft detection could be greatly improved because multiple decision trees were running in parallel. Then, the improved RF algorithm and K-SMOTE oversampling algorithm were combined to establish the electricity theft detection system, which considered the unbalance of the users' electricity dataset.
- (3) The detection method of this paper had higher detection accuracy and better stability compared with the existing methods.

This paper is organized as follows: Section 2 proposes the K-SMOTE. Section 3 describes the proposed detection method for electricity theft behaviors. In Section 4, simulation results are presented to verify the feasibility and superiority of the proposed method. Section 5 summarizes the main conclusion of this study. In addition, the nomenclature table is shown in Appendix A.

2. Proposed Algorithm

The detection of electricity theft behaviors is a binary classification problem which calls for distinguishing of normal and electricity theft users. If the electricity data of the user side are directly used by a classifier, unbalanced data may make the classifier more prone to PD and ignore the important information contained in ND, which may degrade the performance of the classifier substantially.

As shown in Figure 1, the triangle and circle represent two kinds of datasets. Respectively, the solid box represents the actual decision boundary of the two kinds of datasets, while the dotted box represents the possible learning decision boundary of the classification algorithm. The number of triangle data in Figure 1a is less than the circular data, so they represent an unbalanced dataset. From Figure 1a,b that shows the normal dataset, it can be seen that the decision boundary of the classification algorithm may be quite different from the real decision boundary if the dataset is unbalanced.

Figure 1. The schematic diagram of the impact of unbalanced data on the classification algorithm. (a) Unbalanced data, (b) normal data.

In the actual power consumption environment, the number of users stealing electricity is far less than normal users, so the users' electricity dataset is an unbalanced dataset. Unbalanced user data will make the classification algorithm more prone to normal user samples, thereby ignoring the important information contained in a small number of electricity theft user samples, making the decision boundary of the classifier and the actual decision boundary inconsistent, resulting in serious performance degradation of the classifier. Therefore, it was necessary to use an appropriate method to balance the dataset. The traditional SMOTE method was easy to cause data marginalization problems. If there are more PD between some ND, the artificial data generated around these ND will cause the problem of blurred boundaries of PD and ND.

In the field of the detection of electricity theft, the problem about the low detection accuracy due to the unbalance of the power consumption dataset on the user side needs to be solved. Based on a kind of unbalanced data processing method based on K-means clustering and SMOTE, named K-SMOTE, the problem of low electricity theft detection accuracy caused by unbalance electricity data is solved in this paper.

2.1. SMOTE

SMOTE is a classic oversampling algorithm normally used for solving data unbalance problems [45]. Compared to the random oversampling approach, SMOTE performance is better in preventing overfitting [40] by adding ND to achieve balancing distribution with PD. The basic idea is to perform linear interpolation between the existing ND and their neighbors. Specific steps of SMOTE are as follows:

- (1) For a sample x_i in minority class samples set X , calculate the Euclidean distance from this sample to all other samples in the set, and get its k nearest neighbor, denoted as y_j ($j = 1, 2, \dots, k$).
- (2) Sampling rate is set according to the data unbalance ratio to determine the sampling magnification. For data x_i , n numbers are randomly selected from their K-nearest neighbors, and new data can be constructed as follows:

$$x_{new} = x_i + rand(0, 1) \times (y_j - x_i) \quad (1)$$

where $x_j = 1, 2, \dots, n$, $rand(0,1)$ represents a random number between 0 and 1.

New data synthesized by SMOTE is shown in Figure 2.

Figure 2. New data synthesized by SMOTE.

In Figure 2 x is the core data currently used to construct the new data: x_1, x_2, x_3, x_4 are the four nearest neighbor data of x ; r_1 , and r_2, r_3, r_4 are synthetic new data.

2.2. K-Means Clustering Algorithm

K-means clustering is a widely used algorithm that takes the distance between data points and cluster center as the optimization objective [46]. The algorithm would maximize the similarity of elements in clusters while minimizing the similarity between clusters. The K-means selects the desired cluster center, K , minimizes the variance of the whole cluster through continuous iteration and recalculation of the cluster center, and takes the relatively compact and mutually independent clusters as the ultimate goal.

The basic idea of the K-means is to determine the number of initial clusters centers, K , and randomly select K data as the center of the initial cluster in the given dataset D . Then, for each remaining data in D , calculate the Euclidean distance to each cluster center, divide it into the cluster class belonging to the nearest cluster center, and repeat the calculation to generate new cluster centers. The clustering process converges when cluster centers encountered no longer change or the number of iterations reaches the preset threshold limit.

Specific steps are as follows:

- (1) Dataset D , denoted as $\{x_1, x_2, x_3, \dots, x_n\}$, randomly select k initial cluster center as $\mu_1, \mu_2, \dots, \mu_k \in D^n$.
- (2) Calculate the Euclidean distance using Equation (2), that is, calculate the distance $d(x_i, \mu_j)$ between x_i to each cluster center, find the minimum d , and divide x_i into the same cluster as μ_i :

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2} \quad (2)$$

where $i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}\}$ and $j = \{x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn}\}$ are n -dimensional dataset.

- (3) After all data have been calculated, the new clustering center of each class can be recalculated by Equation (3):

$$\mu_{j+1} = \frac{1}{N_j} \sum_{x_i \in S_j} x_i \quad (3)$$

where N_j represents the number of data in class j .

- (4) If the cluster to which each data belongs does not change with the increase of iteration process, go to Step 2. Otherwise, go to Step 5.
- (5) Output clustering results.

2.3. K-SMOTE

This paper combined the K-means algorithm and SMOTE to balance the electricity data on the user side. The specific steps are as follows:

- (1) Let M , O , P , and N represent the unbalanced electricity dataset on the user side, the PD, and the ND. T is the training set in P , S is the majority training set, and T and S constitute the total training set O . K is the number of the initial clusters, u_i is the cluster center, and X_{new} is the corresponding new interpolated data point set.
- (2) Determine the number of initial clusters K .
- (3) For T , K-means algorithm was used to perform clustering and record cluster centers. T was divided into K clusters, and the cluster center were $\{\mu_1, \mu_2, \mu_3, \dots, \mu_k\}$.
- (4) SMOTE was used in T to achieve data interpolation based on cluster centers $\{\mu_1, \mu_2, \mu_3, \dots, \mu_k\}$, then the interpolated dataset X_{new} was obtained
- (5) T , S , and X_{new} were combined to form new training set O' .

3. Random Forest Classification Based on K-SMOTE

RF, a statistical learning algorithm proposed by Breiman in 2001 [47], is essentially a combinatorial classifier containing multiple decision trees [48]. It mainly uses bagging method to generate bootstrap training datasets and classification and regression tree (CART) to generate pruning-free decision trees. As a new machine learning classification and prediction algorithm, random forest features the following advantages.

- (1) Compared with existing classification algorithms, its average accuracy is at a high level [49].
- (2) It can process input data with high dimensional characteristics without dimensionality reduction [50].
- (3) An unbiased estimate of the internal generation error can be obtained during the generation process.
- (4) It is robust to default value problems.
- (5) Each decision tree in the random forest operates independently, realizing parallel operation of multiple decision trees and saving resources and computational time.
- (6) Randomness is reflected in the random selection of data and attributes, even if each tree is not pruned, there will be no overfitting.

The electricity data of the grid user side includes various types, such as voltage, current, power consumption, user classification, etc. So the electricity theft users need to be detected quickly and accurately, so as to promptly notify the power department or relevant stakeholders to take timely and proper action.

On the other hand, the RF classifier has poor processing ability for unbalanced datasets, so in this paper it was combined with the K-SMOTE to detect electric power theft.

3.1. Decision Tree

Random forest is a single classifier composed of several decision trees. Decision trees can be regarded as a tree model including three kinds of nodes: Root, intermediate, and leaf nodes. Each node represents the attribute of the object, the bifurcation path from each node represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root to the leaf node. The path, which starts from the root to the leaf node, represents a rule, and the whole tree represents a set of rules determined by the training dataset. The decision tree has only a single output, which starts from the root node, and only the unique leaf nodes can be reached. In other words, the rule is unique essentially. The classification idea of decision tree is a data mining process which is achieved by analyzing data with a series of generated rules.

Concept learning system (CLS), iterative dichotomiser 3 (ID3), classification 4.5 (C4.5), CART and other node-splitting algorithms can be used to generate the decision trees [51]. This paper selected the CART node-splitting algorithm because it can handle both continuous variables and discrete variables.

The principle of CART node-splitting algorithm is as follows.

Information entropy (IE) is the most commonly used indicator to measure the purity of a sample set. Assume that the proportion of the k -th samples in the set D is p_k ($k = 1, 2, \dots, r$), then the information entropy of D ($Ent(D)$) is defined as:

$$Ent(D) = -\sum_{k=1}^r p_k \log_2 p_k. \quad (4)$$

The smaller the value of $Ent(D)$, the higher is the purity of D .

CART decision tree uses the Gini-index to select the partitioning attributes. Using the same sign as in Equation (9), the purity of the dataset D can be measured using the *Gini* value, calculated as below:

$$Gini(D) = \sum_{k=1}^r \sum_{k' \neq k} p_k = 1 - \sum_{k=1}^r p_k^2. \quad (5)$$

Intuitively, $Gini(D)$ reflects the probability that two samples are randomly selected from the dataset D , and their class labels are inconsistent. Therefore, the smaller the $Gini(D)$, the higher is the purity of the dataset D .

Assume that the discrete attribute a has V possible values $\{a^1, a^1, \dots, a^V\}$. If property a is used to partition the dataset D , there will be V branch nodes, in which the v node contains all the data with a^v value on property a in D , and is denoted as D^v . We can calculate the IE of D^v according to Equation (4). Considering that the number of samples contained in different branch nodes is different, then give each branch node a weight $|D^v|/|D|$, that is, the more samples of branch nodes, the greater the influence of branch nodes. Then, the *Gini-index* of the attribute a is defined as:

$$Gini(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} Gini(D^v) \quad (6)$$

In the candidate attribute set A , select the attribute that minimizes the *Gini-index* after division as the optimal division attribute, and define the optimization attribute as a^* ; then:

$$a_* = \min_{a \in A} Gini_index(D, a) \quad (7)$$

3.2. Discretization of Continuous Variable

The continuous attribute in the decision tree needs to be discretized. The dichotomy method is used for node splitting of decision trees. The main idea of the method is to find the maximum and minimum values of a continuous variable, and set multiple equal breakpoints between them. These equal breakpoints divide the dataset into two small sets and calculate the information gain rate generated by each breakpoint. In CART decision tree, the steps of discretization of continuous variables are as follows.

- (1) Sort the values of continuous variables to find the maximum (MAX) and minimum (MIN).
- (2) If there are N values for continuous variables and each value is a breakpoint, the interval (MIN, MAX) is divided into $N-1$ intervals.
- (3) For each breakpoint A_i ($i = 1, 2, \dots, N$), *Gini-index* is calculated with A and B as intervals.
- (4) Select a breakpoint A_i with the largest *Gini-index* coefficient as the best split point of the continuous attribute.

3.3. Random Forest

3.3.1. Bootstrap Random Sampling

Bootstrap random sampling algorithm is used to obtain different training datasets for training base classifiers.

The mathematical model of bootstrap is as follows: Assuming that there are n different data $\{x_1, x_2, x_3, \dots, x_n\}$ in the dataset D , if any data is extracted from D and put back for n times to form a new set D^* , then the probability that D^* does not contain the x_i ($i = 1, 2, \dots, n$) is $(1-1/n)^n$. When $n \rightarrow \infty$, it can be launched:

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368. \quad (8)$$

Equation (8) indicates that approximately 36.8% of the original data are not extracted in each sampling. This part of the data are called out-of-bag (OOB) data.

3.3.2. OOB Error Estimate

OOB data are not fitted to the training set. However, OOB data can be used to test the generalization capabilities of the model. It has been proven that the error calculated by OOB, called OOB error, is an unbiased estimate of the true error of the random forest [52]. Therefore, OOB error can be used to evaluate the accuracy of the random forest algorithm.

The performance of the generated random forest can be tested with OOB data. The principle of OOB is shown in the Table 1, in the first column, where x_i represents the input sample and y_i represents the classification label corresponding to x_i . In the first row, T_i represents the decision tree constructed by RF. Yes “Y” indicates that the current sample participates in the classification of the corresponding decision tree, and No “N” indicates that the current sample does not participate in the classification of the corresponding decision tree. Therefore, it can be seen from Table 1 that (x_1, y_1) was not used for the construction of T_1 , T_2 , and T_3 , so (x_1, y_1) was the OOB data of the decision trees T_1 , T_2 , and T_3 . After RF model is trained, its performance can be tested by OOB dataset, and the test result is the OOB error. In addition, there is also a relationship between the number of decision trees and the OOB error; therefore, for a certain dataset, this relationship can be used to solve the optimal number of decision trees in RF.

Table 1. The schematic of OOB.

Data \ Decision Tree	T_1	T_2	T_3	$\dots$	T_n
(X_1, Y_1)	N	N	N	$\dots$	Y
(X_2, Y_2)	N	N	Y	$\dots$	Y
(X_3, Y_3)	N	N	Y	$\dots$	N
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(X_4, Y_4)	Y	Y	N	$\dots$	N

Suppose the random forest consists of k decision trees. The OOB dataset is O and the OOB data of each decision tree are O_i ($i = 1, 2, \dots, k$), bringing the OOB data into the corresponding decision tree for classification. The numbers of misclassifications of each decision tree are set to X_i ($i = 1, 2, \dots, k$), and the error size of the OOB is calculated from:

$$\text{OOB Error} = \frac{1}{k} \sum_{i=1}^k \frac{X_i}{O_i}. \quad (9)$$

3.3.3. Random Forest

Random forest is a set of tree classifiers $\{h(x, \theta_k), k = 1, 2, \dots, n\}$ and $h(x, \theta_k)$ is the meta-classifier, which is a classification regression tree composed of CART algorithm. As an independent random vector, $h(x, \theta_k)$ determines the growth of each decision tree and x is the input vector of the classifier.

A schematic diagram of the random forest algorithm is shown in Figure 3.

Figure 3. Schematic diagram of random forest algorithm.

Combined with the proposed oversampling method, the specific electricity theft detection steps are as follows:

- (1) The unbalanced user-side dataset M is oversampled by K-SMOTE to obtain dataset M' .
- (2) Divide the training set Tr and test set Te of random forest.
- (3) Set the number of initial decision tree $nTree$.
- (4) Use the bootstrap method to select training data for each decision tree. The total features in M are K . Selecting n features randomly, n is calculated using Equation (15). Then, use the CART algorithm to generate the unpruned decision trees.

$$n = \sqrt{K} \quad (10)$$

- (5) Input test set Te into each trained decision trees, and the classification result is determined according to the voting result of each decision tree. The voting classification formula is as follows:

$$f(Te_i) = MV\{h_t(Te_i)\}_{t=1}^{nTree} \quad (11)$$

where T_{ei} ($i = 1, 2, \dots, k$) represents each element in the test set, MV represents the majority vote, and $h_t(T_{ei})$ represents the classification result of element T_{ei} in decision tree T .

- (6) The current OOB error is calculated according to Equation (9). If the OOB error converges, then go to Step 7. If the error does not converge, update the decision tree number $nTree$ according to Equation (11) and return to Step 4.
- (7) Output classification result.

Based on the above theory and steps, the proposed electricity theft detection process was as shown in Figure 4.

Figure 4. The proposed process of electricity theft detection.

4. Simulation Results

In order to verify the accuracy and effectiveness of the proposed detection method, three models, back-propagation neural network (BPN), support vector machine (SVM), and RF were established. The parameters of models are as follows.

- (1) The neurons' number of input layer, hidden layer, and output layer in BPN were 20, 40 and 1, respectively. Learning rate was 0.3, momentum term was 0.3, batch volume was 100, and the maximum number of iterations was 50 [53].
- (2) The kernel function of SVM was a radial basis function (RBF), the parameter coefficient of kernel function penalty g WAS 0.07, and the penalty factor coefficient c of PD and ND were 1 and 0.01, respectively [30].

This paper selected the short-term load data of 50 urban electricity users from 15 March 2018 to 16 May 2018 in Hebei province, China. The dataset included six data types (peak, flat, and valley active

power; power factor; voltage; and current) and three user types (industrial, commercial, and residential). Data were sampled at intervals of 30 min through smart meters. Among them, the unbalance ratio was 16.47%.

4.1. Evaluation Indexes

After the classification of unbalanced data, all test sets were divided into four cases: TN (true negative), TP (true positive), FP (false positive), and FN (false negative). These indicators constituted a confusion matrix, as shown in Table 2. Confusion matrix is a way to evaluate the model performance, where the row corresponds to the category to which the object actually belongs and the column represents the category predicted by the model.

Table 2. Two-class confusion matrix.

Actual \ Predicted	Predicted Positive	Predicted Negative
	TP FP	FN TN

FP is the first type of error and FN is the second type of error. Through confusion matrix, multiple evaluation indexes can be extended.

- (1) Accuracy (ACC): ACC is the ratio of the number of correct classifications to the total number of samples. The higher the value of ACC, the better is the performance of the detection algorithm. Mathematically ACC is defined as:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}. \quad (12)$$

- (2) True Positive Rate (TPR): TPR describes the sensitivity of the detection model to PD. The higher the value of TPR, the better is the performance of the detection algorithm. TPR is defined as:

$$TPR = \frac{TP}{TP + FN}. \quad (13)$$

- (3) False Positive Rate (FPR): FPR refers to the proportion of data in ND, which actually belongs to ND, and is wrongly judged as PD by the detection algorithm. FPR is defined as:

$$FPR = \frac{FP}{FP + TN}. \quad (14)$$

- (4) True Negative Rate (TNR): TNR describes the sensitivity of the detection model to ND, which is defined as:

$$TNR = \frac{TN}{TN + FP}. \quad (15)$$

- (5) G-mean index: G-mean index is used for the evaluation of classifier performance [54]. Large G-mean index reveals better classification performance. The value of G-mean depends on the square root of the product of the accuracy of PD and ND. G-mean can reasonably evaluate the overall classification performance of unbalanced dataset, and it can be expressed as:

$$G - mean = \sqrt{TPR * TNR}. \quad (16)$$

- (6) Receiver operating characteristic (ROC) and area under the ROC curve (AUC): Receiver operator characteristic curve (ROC) was originally created to test the performance of a radar [55]. ROC curve

describes the relationship between the relative growth of FPR and TPR in the confusion matrix. For values output by the binary classification model, the closer the ROC curve is to the point (0, 1), the better the classification performance. Area under the ROC curve (AUC), is an index to evaluate the performance of the detection algorithm in the ROC curve. The AUC value of 1 corresponds to an ideal detection algorithm.

4.2. Unbalanced Processing of User-Side Data

Random oversampling, SMOTE, and K-SMOTE were used to oversample the datasets, and the results are shown in Figure 5, of which the black circle represents the normal users, the red asterisk represents the electricity theft users, and the blue box represents the data generated after oversampling.

Figure 5. Schematic diagram of samples generated by different oversampling methods. (a) Raw data, (b) random oversampling, (c) synthetic minority oversampling technique, (d) improved SOMTE based on K-means. red: electricity theft users; Blue: data generated after oversampling; GREY: normal users.

In addition, Table 3 shows the repetition rate of artificial data and original data generated by several oversampling algorithms.

Table 3. The schematic of OOB.

Algorithm	Random Oversampling	SMOTE	K-SMOTE
Data Repetition Rate/%	95.02	30.50	15.84

It can be observed from Figure 5 that a large amount of duplicated data were included in the result of random oversampling algorithm, and some data were never selected. From Table 3, the data

repetition rate of random oversampling was 95.02%, which indicates that the oversampling effect was not ideal. The data repetition rate of SMOTE was 30.5%, as can be seen from Figure 5. Data generated by SMOTE were scattered with other data and introduced noise points. The problem of data overlap still existed and could not be ignored. K-SMOTE can generate data near the center, and use representative points to limit the boundaries of the generated data to avoid introducing noise. Data generated by K-SMOTE generally follows the original distribution. Further, as shown in Table 3, the data repetition rate was only 15.84%.

4.3. Electricity Theft Detection Based on Improved RF

4.3.1. Determination of the Number of Decision Trees

The number of decision trees is relevant to the accuracy of the algorithm.

In this paper, 80% of the user data were set to form a training set and 20% to form a test set. The optimal number of decision trees can be determined by minimizing the OOB error. The relationship between the OOB error and the number of decision trees, n_{Tree} , is shown in Figure 6.

Figure 6. Value of OOB error varies with the number of decision trees.

It can be observed that when the decision tree number was larger than 368, the OOB error almost converged to a minimum level. If the number of decision trees was too small, the accuracy was low. On the other hand, too many decision trees did not improve the accuracy further and the algorithm burden was increasing. Therefore, the decision tree number was set to 368.

4.3.2. Detection Results of RF

The above-mentioned electricity users' dataset processed by K-SMOTE and not processed by K-SMOTE oversampling were, respectively, detected by RF. In order to make the simulation results more convincing and avoid randomness, three independent tests were carried out for each detection. ACC values of test data are listed in Table 4, and ROC curves are shown in Figures 7 and 8. In Figures 7 and 8, the three differently colored curves of red, green and blue represent the ROC curve of three independent tests. According to these results, when K-SMOTE was not used for unbalanced data processing, the mean value of ACC in RF was 85.53%, while the average value of ACC in RF after K-SMOTE was 94.53%. In addition, it can be concluded that ROC curve detected by RF with K-SMOTE was obviously closer to (0,1) than ROC curve detected RF algorithm without K-SMOTE. That is, the area under the former ROC curve was larger.

Moreover, the AUC index of the former was obviously better than that of the latter, which shows that it is necessary to use K-SMOTE to deal with unbalanced data before the detection of electricity theft behaviors. In addition, the detection performance of RF was also ideal.

Table 4. Accuracy value of random forest.

Simulation Times	K-SMOTE	ACC/%
1	with	96.02
	without	83.24
2	with	96.21
	without	82.01
3	with	81.32
	without	91.35
Mean	with	94.53
	without	85.53

Figure 7. Receiver operating characteristic curve detected by RF with K-SMOTE. **Figure 8.** ROC curve detected RF algorithm without K-SMOTE.

4.3.3. Comparison of Detection Performance of Different Algorithms

The electricity users' data processed by K-SMOTE were tested by BPN and SVM. Again, in order to make the simulation results more convincing, the same dataset was used and three independent tests were performed. The testing results are shown in Table 5 and Figures 9 and 10. In Figures 9 and 10, the three differently colored curves of red, green and blue represent the ROC curve of three independent tests.

Table 5. Accuracy value of three algorithms.

Simulation Times	Algorithm	ACC/%
1	SVM	70.21
	BPN	85.24
2	SVM	72.54
	BPN	83.01
3	SVM	71.02
	BPN	86.35
Mean	SVM	71.26
	BPN	84.87

Figure 9. ROC curve detected by support vector machines with K-SMOTE. **Figure 10.** ROC curve detected by back-propagation neural network with K-SMOTE.

It can be concluded from the above test results that:

- (1) Without K-SMOTE, the ACC value and AUC of RF detection method were relatively low. However, with K-SMOTE, the ACC and AUC value of three detection methods were obviously improved, which was increased about 10%. This indicates that unbalanced datasets would affect the accuracy of detection algorithm, and K-SMOTE plays an effective role in improving machine learning accuracy.

- (2) The electricity user data processed by K-SMOTE were tested by BPN and SVM. The ACC mean values of SVM and BPN were 71.26% and 84.87%, respectively, and the mean values of AUC in SVM and BPN were 0.7236 and 0.8716, respectively. These indexes were lower than the ACC and AUC of RF, which were 94.53% and 0.9513, respectively. Thus, the performance of RF was superior to SVM and BPN.

5. Conclusions

In order to better adapt to the rapid development of the power grid, aiming at the unbalanced dataset on the user side and improving the efficiency and accuracy of electricity theft detection algorithms, this paper proposed a method based on K-SMOTE and RF classifier for detecting electricity theft. The main conclusions can be summarized as below:

- (1) K-SMOTE was proposed to avoid the influence of unbalanced data on the accuracy of the classifier.
- (2) The RF classifier, which was suitable for the nature of the user-side dataset, was used to detect electricity theft. The decision trees in RF classifier could work in parallel, which improved the detection efficiency and reduced the computational time.
- (3) Compared with the conventional detection methods, the proposed method featured higher accuracy and stronger stability.

The method proposed in this paper can provide reliable targets for manual inspection, thereby reducing nontechnical losses in power systems and, hence, improving system reliability and security.

Author Contributions: Conceptualization, Z.Q.; Methodology, H.L.; Data Curation, Y.W.; Writing-Original Draft Preparation, J.Z.; Writing-Review & Editing, A.A.-S. Resources, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Project of China [grant number 2016YFF0200105], and the National Natural Science Foundation of China [grant number 51777199].

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

Appendix A

Table A1. Nomenclature.

Acronyms	Full Name
SMOTE	synthetic minority oversampling technique
RF	random forest
PD	positive data
ND	negative data
OOB Error	out-of-bag data error
TL	Technical loss
NTL	non-technical losses
PCA	principal component analysis
FA	factor analysis
ELM	extreme learning machine
BC	balance cascade

Table A1. Cont.

Acronyms	Full Name
KNN	k-nearest neighbor
TN	true negative
TP	true positive
FP	false positive
FN	false negative
ACC	Accuracy
TPR	true positive rate
FPR	false positive rate
TNR	true negative rate
IE	information entropy
OOB	out of bag
SVM	support vector machines
BPN	back-propagation neural network

References

- Lo, Y.L.; Huang, S.C.; Lu, C.N. Non-technical loss detection using smart distribution network measurement data. In Proceedings of the IEEE PES Innovative Smart Grid Technologies, Tianjin, China, 21–24 May 2012; pp. 1–5.
- Smith, T.B. Electricity theft: A comparative analysis. *Energy Policy* **2004**, *32*, 2067–2076. [\[CrossRef\]](#)
- Huang, S.C.; Lo, Y.L.C.N. Non-Technical Loss Detection Using State Estimation and Analysis of Variance. *IEEE Trans. Power Syst.* **2013**, *28*, 2959–2966. [\[CrossRef\]](#)
- Bhavna, B.; Mohinder, G. Reforming the Power Sector, Controlling Electricity Theft and Improving Revenue. Public Policy for the Private Sector. 2004. Available online: <http://rru.worldbank.org/PublicPolicyJournal> (accessed on 16 December 2019).
- Soma, S.S.R.D.; Wang, L.; Vijay, D.; Robert, C.G. High performance computing for detection of electricity theft. *Int. J. Electr. Power Energy Syst.* **2013**, *47*, 21–30.
- Smart Meters Help Reduce Electricity Theft, Increase Safety. BCHydro. Available online: https://www.bchydro.com/news/conservation/2011/smart_meters_energy_theft.html (accessed on 16 December 2019).
- Dzung, D.; Naedele, M.; Von Hoff, T.P.; Crevatin, M. Security for Industrial Communication Systems. *Proc. IEEE Secur. Ind. Commun. Syst.* **2005**, *93*, 1152–1177. [\[CrossRef\]](#)
- Krebs, B. FBI: Smart Meter Hacks Likely to Spread. Available online: <http://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread> (accessed on 16 December 2019).
- Carlos, L.; Félix, B.; Iñigo, M.; Juan, I.; Guerrero, J.B.; Rocío, M. Integrated expert system applied to the analysis of non-technical losses in power utilities. *Expert Syst. Appl.* **2011**, *38*, 10274–10285.
- Han, W.; Xiao, Y. NFD: A practical scheme to detect non-technical loss fraud in smart grid. In Proceedings of 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, Australia, 10–14 June 2014; pp. 605–609.
- Grochowski, D.; Huh, J.H.; Berthier, R. AMI threats, intrusion detection requirements and deployment recommendations. In Proceedings of the IEEE Third International Conference on Smart Grid Communications, Tainan, Taiwan, China, 5–8 November 2012; pp. 395–400.
- Hao, R.; Ai, Q.; Xiao, F. Architecture based on multivariate big data platform for analyzing electricity consumption behavior. *Electr. Power Autom. Equip.* **2017**, *37*, 20–27.
- Jiang, R.; Lu, R.; Wang, Y.; Luo, J.; Shen, C.; Shen, X. Energy-theft detection issues for advanced metering infrastructure in smart grid. *Tsinghua Sci. Technol.* **2014**, *19*, 105–120. [\[CrossRef\]](#)
- Lo, C.-H.; Ansari, N. CONSUMER: A Novel Hybrid Intrusion Detection System for Distribution Networks in Smart Grid. *IEEE Trans. Emerg. Top. Comput.* **2013**, *1*, 33–44. [\[CrossRef\]](#)

15. McLaughlin, S.; Holbert, B.; Fawaz, A.; Berthier, R.; Zonouz, S. A multi-sensor energy theft detection framework for advanced metering infrastructures. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 1319–1330. [\[CrossRef\]](#)
16. Leite, J.B.; Sanches mantovani, J.R. Detecting and locating non-technical losses in modern distribution networks. *IEEE Trans. Smart Grid* **2018**, *9*, 1023–1032. [\[CrossRef\]](#)
17. Tariq, M.; Poor, H. Electricity Theft Detection and Localization in Grid-Tied Microgrids. *IEEE Trans. Smart Grid* **2018**, *9*, 1920–1929. [\[CrossRef\]](#)
18. Salinas, S.; Li, P. Privacy-Preserving Energy Theft Detection in Microgrids: A State Estimation Approach. *IEEE Trans. Power Syst.* **2016**, *31*, 883–894. [\[CrossRef\]](#)
19. Cárdenas, A.A.; Amin, S.; Schwartz, G.; Dong, R.; Sastry, S. A game theory model for electricity theft detection and privacy-aware control in AMI systems. In Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 1–5 October 2012; pp. 1830–1837.
20. O’Leary, D.E. Summary of Previous Papers in Expert Systems Review. *Intell. Syst. Account. Financ. Manag.* **2016**, *1*, 3–7. [\[CrossRef\]](#)
21. 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. [\[CrossRef\]](#)
22. Coma-Puig, B.; Carmona, J. Bridging the Gap between Energy Consumption and Distribution through Non-Technical Loss Detection. *Energies* **2019**, *12*, 1748. [\[CrossRef\]](#)
23. Lu, X.; Zhou, Y.; Wang, Z.; Yi, Y.; Feng, L.; Wang, F. Knowledge Embedded Semi-Supervised Deep Learning for Detecting Non-Technical Losses in the Smart Grid. *Energies* **2019**, *12*, 3452. [\[CrossRef\]](#)
24. Nagi, J.; Yap, K.S.; Tiong, S.K.; Ahmed, S.K.; Mohamad, M. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. *IEEE Trans. Power Deliv.* **2009**, *25*, 1162–1171. [\[CrossRef\]](#)
25. Zhuang, C.J.; Zhang, B.; Hu, J.; Li, Q.; Zeng, R. Anomaly detection for power consumption patterns based on unsupervised learning. *Proc. CSEE* **2016**, *36*, 379–387. (In Chinese)
26. Ghasemi, A.A.; Gitizadeh, M. Detection of illegal consumers using pattern classification approach combined with Levenberg-Marquardt method in smart grid. *Int. J. Electr. Power Energy Syst.* **2018**, *99*, 363–375. [\[CrossRef\]](#)
27. Monedero, I.; Biscarri, F.; Leon, C.; Biscarri, J.; Millan, R. MIDAS: Detection of Non-technical Losses in Electrical Consumption Using Neural Networks and Statistical Techniques. In Proceedings of Computational Science and Its Applications-ICCSA 2006, International Conference, Glasgow, UK, 8–11 May 2006; Proceedings, Part V; Springer: Berlin/Heidelberg, Germany, 8 May 2006; pp. 725–734.
28. Nizar, A.H.; Dong, Z.Y.; Wang, Y. Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method. *IEEE Trans. Power Syst.* **2008**, *23*, 946–955. [\[CrossRef\]](#)
29. Muniz, C.; Figueiredo, K.; Vellasco, M.; Chavez, G.; Pacheco, M. Irregularity detection on low tension electric installations by neural network ensembles. In Proceedings of the International Joint Conference on Neural Networks, Atlanta, GA, USA, 14–19 June 2009; pp. 2176–2182.
30. Nagi, G.; Yap, K.S.; Tiong, S.K.; Ahmed, S.K.; Nagi, F. Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System. *IEEE Trans. Power Deliv.* **2011**, *26*, 1284–1285. [\[CrossRef\]](#)
31. Viegas, J.L.; Esteves, P.R.; Vieira, S.M. Clustering-based novelty detection for identification of non-technical losses. *Int. J. Electr. Power Energy Syst.* **2018**, *101*, 301–310. [\[CrossRef\]](#)
32. 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. [\[CrossRef\]](#)
33. Messinis, G.M.; Hatziaargyriou, N.D. Review of non-technical loss detection methods. *Electr. Power Syst. Res.* **2018**, *158*, 250–266. [\[CrossRef\]](#)
34. Jokar, P.; Arianpoo, N.; Leung, V.C.M. Electricity Theft Detection in AMI Using Customers’ Consumption Patterns. *IEEE Trans. Smart Grid* **2015**, *7*, 216–226. [\[CrossRef\]](#)
35. Figueroa, G.; Chen, Y.; Avila, N.; Chu, C. Improved practices in machine learning algorithms for NTL detection with imbalanced data. In Proceedings of the 2017 IEEE Power & Energy Society General Meeting, Chicago, IL, USA, 16–20 July 2017; pp. 1–5.
36. Glauner, P.; Boechat, A.; Lautaro, D.; Radu, S.; Franck, B.; Yves, R.; Diogo, D. Large-scale detection of non-technical losses in unbalanced datasets. In Proceedings of the 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Minneapolis, MN, USA, 6–9 September 2016; pp. 1–5.

37. Liu, X.; Wu, J.; Zhou, Z. Exploratory Under sampling for Class-Unbalance Learning. *IEEE Trans. Cybern.* **2009**, *39*, 539–550.
38. Victoria, L.; Alberto, F.; Salvador, G.; Palade, V.; Herrera, F. An insight into classification with unbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141.
39. Del Río, S.; López, V.; Benítez, J.M.; Herrera, F. On the use of MapReduce for unbalanced big data using Random Forest. *Inf. Sci.* **2014**, *285*, 112–137. [[CrossRef](#)]
40. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
41. Feilong, C.; Yuanpeng, T.; Miaomiaom, C. Sparse algorithms of Random Weight Networks and applications. *Expert Syst. Appl.* **2014**, *41*, 2457–2462.
42. Song, X.; Guo, Z.; Guo, H.; Wu, S.; Wu, C. A new forecasting model based on forest for photovoltaic power generation. *Power Syst. Prot. Control* **2015**, *43*, 13–18.
43. Monedero, I.; Félix, B.; Carlos, L.; Guerrero, J.I.; Jesús, B.; Rocío, M. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. *Int. J. Electr. Power Energy Syst.* **2012**, *34*, 90–98. [[CrossRef](#)]
44. Gao, D.; Zhang, Y.X.; Zhao, Y.H. Random forest algorithm for classification of multiwavelength data. *Res. Astron. Astrophys.* **2009**, *9*, 220–226. [[CrossRef](#)]
45. Fernandez, A.; García, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [[CrossRef](#)]
46. Bradley, R.; Fayyad, U. Refining Initial Points for K-Means Clustering. In Proceedings of the Fifteenth International Conference on Machine Learning; 1998; pp. 91–99. Available online: <https://dl.acm.org/doi/10.5555/645527.657466> (accessed on 20 December 2019).
47. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
48. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [[CrossRef](#)]
49. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
50. Alessia, S.; Antonio, C.; Aldo, Q. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review. *Front. Aging Neurosci.* **2017**, *9*, 329–336.
51. Quinlan, J.R. Decision trees and decision-making. *IEEE Trans. Syst. Man Cybern.* **1990**, *20*, 339–346. [[CrossRef](#)]
52. Wolpert, D.H.; Macready, W.G. An Efficient Method to Estimate Bagging’s Generalization Error. *Mach. Learn.* **1999**, *35*, 41–55. [[CrossRef](#)]
53. Xu, G.; Tan, Y.P.; Dai, T.H. Sparse Random Forest Based Abnormal Behavior Pattern Detection of Electric Power User Side. *Power Syst. Technol.* **2017**, *41*, 1965–1973.
54. Rao, R.B.; Krishnan, S.; Niculescu, R.S. Data mining for improved cardiac care. *ACM SIGKDD Explor. Newsl.* **2006**, *1*, 3–10. [[CrossRef](#)]
55. Huang, Y.A.; You, Z.H.; Gao, X.; Wong, L.; Wang, L. Using Weighted Sparse Representation Model 561 Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *BioMed Res. Int.* **2015**, *2015*, 902198. [[CrossRef](#)] [[PubMed](#)]