energies

MDPI

Article

Calculation Method of Theoretical Line Loss in Low-Voltage Grids Based on Improved Random Forest Algorithm

Li Huang [1,*], Gan Zhou [1], Jian Zhang [2], Ying Zeng [2] and Lei Li [1]

[1] School of Electrical Engineering, Southeast University, Nanjing 211189, China
[2] Guangdong Power Grid Co., Guangzhou 510600, China
* Correspondence: 230198242@seu.edu.cn

Abstract: Theoretical line loss rate is the basic reference value of the line loss management of low-voltage grids, but it is difficult to calculate accurately because of the incomplete or abnormal line impedance and measurement parameters. The traditional algorithm will greatly reduce the number of samples that can be used for model training by discarding problematic samples, which will restrict the accuracy of model training. Therefore, an improved random forest method is proposed to calculate and analyze the theoretical line loss of low-voltage grids. According to the Influence mechanism and data samples analysis, the electrical characteristic indicator system of the theoretical line loss can be constructed, and the concept of power supply torque was proposed for the first time. Based on this, the attribute division process of decision tree model is optimized, which can improve the limitation of the high requirement of random forest on the integrity of feature data. Finally, the improved effect of the proposed method is verified by 23,754 low-voltage grids, and it has a better accuracy under the condition of missing a large number of samples.

Keywords: low-voltage grids; theoretical line loss rate; improved random forest; decision tree optimization

Citation: Huang, L.; Zhou, G.; Zhang, J.; Zeng, Y.; Li, L. Calculation Method of Theoretical Line Loss in Low-Voltage Grids Based on Improved Random Forest Algorithm. *Energies* **2023**, *16*, 2971. https://doi.org/10.3390/en16072971

Academic Editor: Alan Brent

Received: 26 February 2023
Revised: 15 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

1. Introduction

According to the data from latest analysis report on China's electric power development, line loss of the power system is gradually getting lower—-from 5.62% in 2020 to 5.26% in 2021, which is due to technological advancement. However, there is still a relatively large gap from developed countries with an average line loss rate of 4%. With consideration of the line loss from low-voltage grids constituting about 40% of the overall power system, there is great practical significance to studying how to manage power systems scientifically and save energy by accurately and effectively digging out energy-loss and thereby reducing the potential for low-voltage grids.

The theoretical line loss in low-voltage grids is defined and determined by the technical conditions of power grid equipment, including the loss of overhead and cable line, the loss of capacitors, reactors, cameras and other auxiliary equipment, plus the loss of voltage, current transducers, energy meters, etc. Compared to some human management factors, theoretical line loss is relatively stable and the energy-loss reducing potential can be detected by the difference between the theoretical and the actual line loss. Due to complex wiring and the huge difference of line length and consumption load in low-voltage grids, traditional methods including equivalent resistance method and power flow algorithm cannot be satisfactorily applied, and the proposal of some improvements are required. For example, [1] improved the calculation of load curve shape coefficient, copper loss, small (many) power supply, and branch power, whilst [2] introduced the average time of current loss and the weakening of hypothesis conditions to improve the accuracy. A network loss calculation method based on user meter power has been proposed, but it is highly dependent on accurate information such as network structure, line type, and

length [3,4]. A method using the variable structure dissipative network theory to divide the feeder into feeder segments has also been proposed, and the detailed line loss of each piece of equipment makes the line loss distribution in the system clear at a glance with a mathematical model [5]. A new CIM-based data-sharing scheme for an online calculation of the theoretical line loss has also been presented, which can be calculated automatically by reading data from other applications that are being used in electric power company, such as electrical SCADA, GIS, etc. [6]. A technical loss estimation approach in power distribution systems using a load model in the frequency domain has also been researched, decomposing the load profile by discrete Fourier transformation, and has been applied to a model adapted to compute the spectral analysis of the losses, showing that the load model in the frequency domain calculation is robust [7].

With continuous construction of state smart grid and the rapid development of 5G technology, the ability of data collection in low-voltage grids is being improved gradually, and more data can be applied to the calculation of theoretical line loss in low-voltage grids. Meanwhile, the new generation artificial intelligence technology like big data, machine learning, and deep learning are employed to conduct theoretical line loss calculation. According to the physical distribution mechanism of the network loss for a power system [8–13], the main characteristic parameters of line loss can be extracted and the theoretical line model can be established by a different algorithm. Influencing factors of the line loss rate are analyzed by an a priori algorithm and by association rules with the data from AMI, and calculation results of the a priori algorithm and the interpretation structure model are combined to make up for the shortcomings and draw the influence path diagram of the regional distribution network line loss rate [14]. A comprehensive analysis in the time domain and the space domain of a large-scale line loss rate is carried out through K-means clustering and a temperature response model, including the relationship between line loss rate and seasons, level area, energy consumption, and so on. The research shows that it is a downward trend, and its significant change cycle is similar to the per capita GDP [15]. The strong self-learning ability of neural network to fit the relationship between line loss rate and characteristic parameters has been used [16,17]. To overcome the disadvantages of a k-means algorithm, an adjacent propagation algorithm is used for data clustering, and, on this basis, an optimized BP neural network model is built based on an LM algorithm that can improve the accuracy of theoretical line loss calculation and the convergence speed of model [18]. A method based on deep belief network (DBN), which used the greedy algorithm to carry out unsupervised layer-by-layer pre-training of neural network layer in DBN first and then implemented the supervised global fine-tuning training, has also been proposed [19]. Furthermore, ref. [20] proposed a method based on hierarchical clustering, a decision tree, and a random forest algorithm, which overcomes the shortcomings of an artificial neural network algorithm with a slow convergence speed and a greater difficulty of dealing with discrete variables directly. The affinity propagation (AP) algorithm to cluster and group data, which surpassed the shortcoming of k-means algorithm, has been used [21]. On this basis, the random forest regression model is used to analyze and calculate the regression. A reasonable interval calculation model of line loss based on the convolutional neural network is established, in which the collected data are processed in the image format first, and the reasonable line loss interval will then be calculated according to the operation data of different transformers [22].

From the reference investigation mentioned above, one can see that most of the research and calculation techniques focus on the algorithm. However, in practical application and calculation, obtainable low voltage grid characteristics are different due to the different management of basic data in different areas. Meanwhile, data missing and abnormality also exist to some extent. If the models are always trained by traditional techniques, featuring abandoning the sample sets with abnormal data, lots of sample sets cannot be applied in model training. Finally, the accuracy of the model calculation is deteriorated.

To address the above problems, the author tries out the back propagation neural network (BPNN), support vector machine (SVM), k nearest neighbour (KNN), and random

forest (RF). The experimental results show that random forest has a better generalization ability than other methods, and the overall accuracy is better than other methods. However, there is still large room for improvement, so this paper proposes a theoretical line loss calculation and analysis method for low-voltage grids based on an improved random forest. The contributions of this paper are as follows:

- The reasonable characteristic factors of the low-voltage grids are constructed according to the physical and operational characteristics. The concept of power supply torque is proposed for the first time.
- The random forest algorithm is improved by modifying the property classifying process of decision tree and optimizing the weight factor allocation method when data is missing. The problems of the high characteristic data integrity requirement is solved and the accuracy of the model is improved when a large number of samples are missing.

Table 1 shows the comparison of the proposed method with existing methods. The detailed algorithms will be given in the following sections.

Table 1. Comparison of proposed method with existing methods to calculate theoretical line loss.

	No Additional Measurement Equipment Required	No Complete Grid and Line Parameters Required	Factors Can Be Interpreted by the Circuitous Philosophy	Feature Integrity Requirement	Complexity of the Model	Accuracy of the Model
Ref. [5]	×	×	√	High	Low	High
Ref. [14]	×	√	×	High	High	High
Refs. [16,17]	√	√	×	Low	Moderate	Low
Ref. [19]	√	√	√	Moderate	Moderate	Moderate
Refs. [20,21]	√	√	×	Moderate	High	Moderate
Proposed method	√	√	√	Low	Low	High

The remainder of this paper is organized as follows. The reasonable characteristic factors of the low-voltage grids are constructed in Section 2. Section 3 presents the algorithm to calculate the theoretical line loss with random forest and improved the adaptability of the model by modifying the property classifying process of decision tree. Section 4 demonstrates the test results for the proposed method validation. Section 5 provides the conclusion of this paper.

2. Analysis of Characteristic Factors

2.1. Influence Mechanism of Theoretical Line Loss

In order to improve the performance of model training, characteristic factors should be constructed according to the influence mechanism of theoretical line loss in low-voltage grids. In this paper, equivalent resistance method with relatively high calculation accuracy is selected to study and analyze the influence mechanism of theoretical line loss in the low-voltage grids. Its basic principle is shown in Figure 1. Through the simplified circuit model, the grids are assumed to be an equivalent resistance R_{eql}, and the electric energy loss generated when the total current of the line I_{av} flowing through this resistance is equal to the sum of the power loss generated by the resistance of each branch line R_i (i = 1,2,...n).

According to the equivalent resistance method, the calculation equation of theoretical line loss in the low-voltage grids can be expressed as:

$$\Delta A = N(kI_{\mathrm{av}})^2 R_{\mathrm{eql}} K_{\mathrm{b}} t \times 10^{-3} + (\frac{t}{24D}) \sum (\Delta A_{\mathrm{dbi}} m_{\mathrm{i}}) + \sum \Delta A_{\mathrm{C}} \quad (1)$$

where ΔA (kWh) is the energy loss in a low voltage grid, N is the structure parameters of the grids and varies according to the wiring mode, k is the load shape factor, I_{av} is the average current at the secondary side of transformer, R_{eql} (Ω) is the equivalent resistance of

low-voltage grids, K_b is the three phase unbalanced coefficient, t (h) is the time of operation, D is the annual calendar days, m_i is the number of electric meters, ΔA_{dbi} (kWh) is the monthly energy loss of electric meters, and ΔA_C (kWh) is the energy loss of reactive-load compensation equipment.

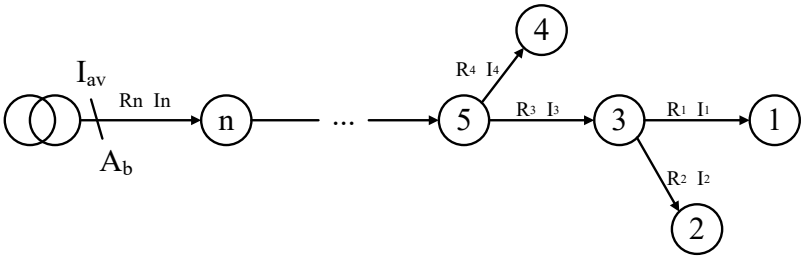

Figure 1. Equivalent resistance method calculation structure diagram.

As can be seen from Equation (1), the theoretical line loss in the low-voltage grids is composed of the loss of the line, the loss of the electric meter, and the loss of the reactive power compensation equipment. The electric meter loss and the reactive power compensation equipment loss are relatively stable, and the line loss in the station area plays an important role in the theoretical line loss of the grids. The main factors affecting the line loss of the station can be subdivided into two categories: static line factors and dynamic operation factors.

2.2. Qualitative Influence Analysis of the Factors

(1) Static line factors

Static line factors mainly include line length and type that affect the equivalent impedance R_{eql} of the low-voltage grids. When the line length becomes longer and the electrical resistivity of the selected line increases, the equivalent resistance R_{eql} of the grids will enlarge and the line loss will increase relatively. Figure 2 shows the investigation of the relationship between line length and line loss in a certain place and there is a relatively positive correlation between the two parameters.

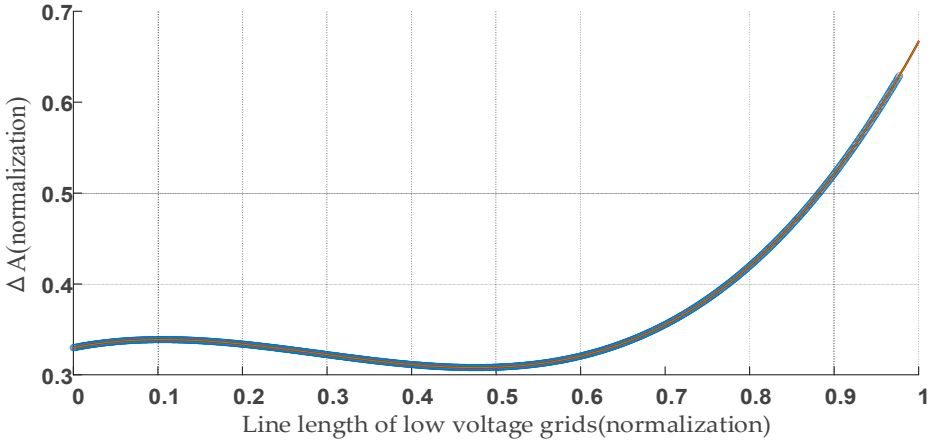

Figure 2. Correlation between line length and ΔA.

(2) Dynamic operation factors

Dynamic operation factors mainly include level of electric load, fluctuation characteristics of electric load, and degree of three phase unbalance. With the increase in the level of electric load, fluctuation characteristics of electric load and degree of three phase unbalance, the line loss will increase correspondingly. Figure 3 shows the investigation of the relationship between

the average current at the secondary side of the transformer and line loss in a certain place, and there is a obviously a positive correlation between the two parameters.

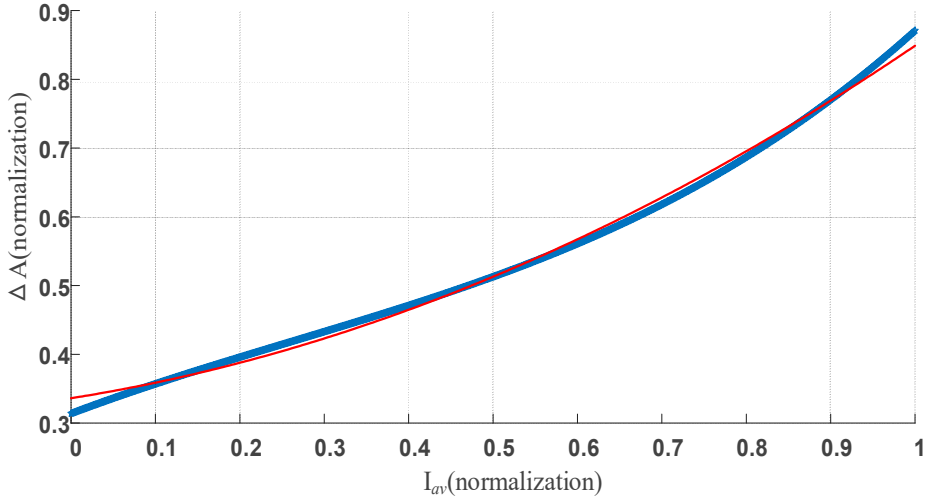

Figure 3. Correlation between the I_{av} and ΔA.

(3) Fusion factors

Since the static line factors are relatively separated from the dynamic operation factors, in order to characterize and analyze the influence mechanism more accurately, Figure 4 shows the concept of power supply torque, which is proposed for the first time in this paper by coupling the line factors and operation factors in the low-voltage grids and referring to the concept of torque in mechanics. The specific definition is shown in Equation (2), which is the average product of the power supply distance and the average daily electricity consumption of all users in the low-voltage grids:

$$M_e = \frac{\sum_{i=1}^{n} (P_i \cdot D_i)}{n} \tag{2}$$

where M_e is the supply torque, P_i is the power of consumer i, n is the amount of consumers in one low voltage grid, and D_i is the distance between consumer i and the transformer of the grid.

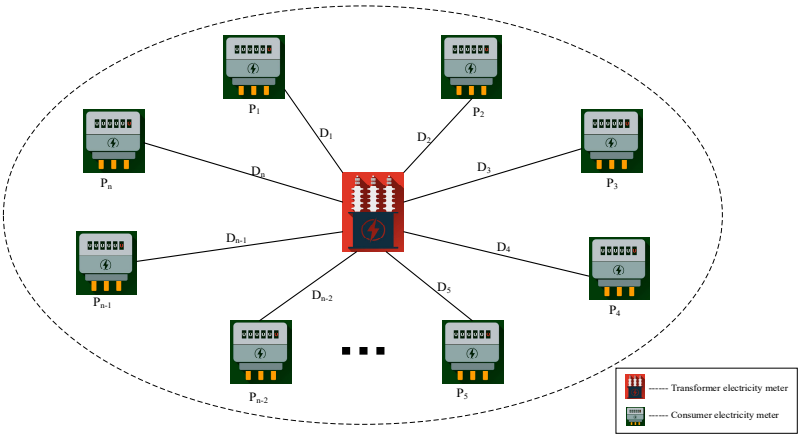

Figure 4. Schematic diagram of power supply torque.

2.3. Construction of Characteristic Factors

Considering all the factors affecting the theoretical line loss of low-voltage grids in Section 2.2, the following eight characteristic factors are selected in this paper, including power supply radius X_1, total line length of low-voltage grids X_2, amount of consumers in a low voltage grid X_3, load rate of low-voltage grids X_4, three-phase unbalance degree X_5, load shape factor X_6, power factor X_7, and power supply torque X_8. Detailed definitions of the various factors mentioned above are listed in Table 2.

Table 2. Characteristic factors for theoretical line loss in low-voltage grids.

Characteristic Factor	Characteristic Definition
Power supply radius X_1	Physical distance from furthest load point to distribution transformer
Total line length X_2	Sum of total low-voltage line length in low-voltage grids
User numbers X_3	Total user numbers in low-voltage grids, including single-phase users and three-phase users
Load rate X_4	Ratio of power consumption capacity to rating capacity of distribution transformer
Three-phase unbalance degree X_5	Unbalance degree of three-phase current in three-phase power system, which is the relative deviation between the maximum and the mean value of the three-phase current in a transformer
Load shape factor X_6	Ratio of daily current RMS value to average value on distribution transformer side in low-voltage grids
Transformer daily average power factor X_7	Ratio of active power to apparent power in low-voltage grids
power supply torque X_8	Multiplication of average power supply distance of low-voltage grids load and user average power consumption capacity

3. Construction of the Method

3.1. Model Based on Traditional Random Forest

Random forest [23–27] is an integrated learning algorithm based on a decision tree. As shown in Figure 5, multiple samples are extracted from original sample sets by Bootstrap sampling method. The decision tree model is established according to each sample, and predictions from multiple decision trees are combined to get the final results by voting mechanism. Instead of setting aside an additional portion of the test set for model evaluation, Out-of-Bag (OOB) data can be used to do the evaluation, and the In-Bag data is used to do the model training.

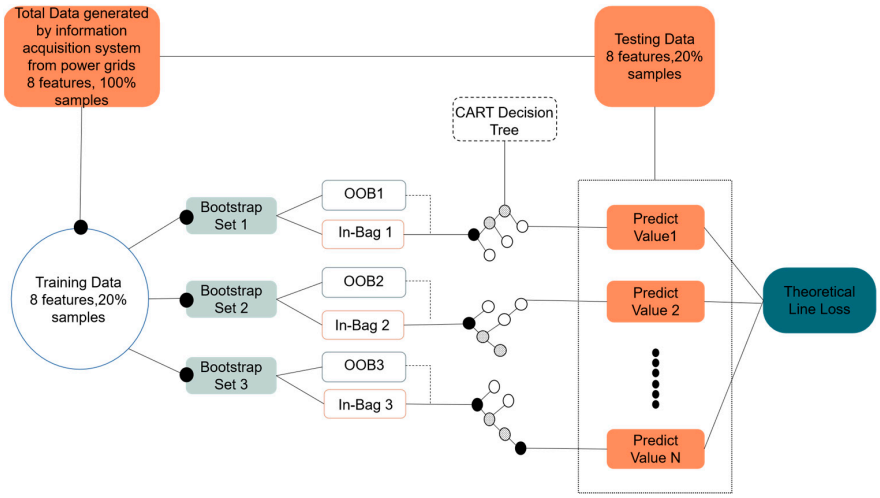

Figure 5. The schematic diagram of the topological structure of the RF algorithm.

Assume original theoretical line loss dataset from low-voltage grids and A corresponding characteristic factors dataset is $S = \{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_m, y_m)\}$. Assume the electrical characteristic index data set is $x = \{X_1, X_2, \ldots\ldots, X_n\}$. Assume theoretical line loss set is y and sample quantity is m, while sample property dimension is n and individual learner quantity of decision tree is k. The process of random forest algorithm implementation is as below:

(1) Extract m data samples randomly from original datasets and repeat the process for k times, then k sets of training datasets are obtained.

(2) Input corresponding data set into each decision tree and select classifying property for each node of the decision tree. Randomly select a subset including d properties from n data properties, and then choose the optimum classifying property from the subset. Normally, d equals to the integer closest to $\log_2 n$. Considering the decision tree algorithm is adopted for an individual learner in the random forest algorithm, the learning capability of the random forest algorithm is contingent on the performance of the decision tree. The implementation steps are described below:

   (a) If label values of all the data in S are the same, the decision tree including only one node is generated and the node value is the same as the label value.

   (b) If A is empty or all the data in S have the same value in A, the decision tree including only one node is then generated, and the node value is the same as the label value belonging to most of the data samples in S.

   (c) Select optimum classifying subset A_i from A.

   (d) Traverse all the values of A_i, and form dataset S_v including all the data with value of A_i^v from property subset A_i in S.

   (e) If S_v is empty, mark S_v as node, and the node value is the same as the label value belonging to most of the data samples in S.

   (f) If S_v is not empty, treat S_v as input dataset and $A \backslash \{A_i\}$ as property set. Repeat steps (a)~(e) until a decision tree is generated.

(3) Average strategy can be applied for regression. All the output values from the decision tree are averaged as final output value. Voting strategy can be applied for classification. Compare all the output classified values from the decision tree and take the one with most votes as the final output value.

(4) Based on historical documentations and measurement data of various low-voltage grids from the consumption data collection system, marketing system, production management system, and geographic information system, the characteristic factor data can be calculated for each low-voltage grid using definition and the calculation principle of various characteristic factors. Meanwhile, abnormal characteristic data should be cleaned. Feed the cleaned sample data into a random forest algorithm for training and establish a theoretical line loss model of low-voltage grids. Finally, finish detailed theoretical line loss calculation with the established model. A detailed algorithm flow chart is given in Figure 6.

3.2. Improved Random Forest Algorithm

ID 3, ID 4.5, and CART decision trees are the most commonly used decision tree algorithms at present. Considering that ID3 is easy to overfit and ID 4.5 is relatively complex, it is difficult to realize model training with a large number of samples. Finally, the CART decision tree is selected based on the calculation principle and calculation requirement of theoretical line loss. The equation below is utilized to select the optimum classify property:

$$Gini\_ratio(S, A_i) = Gini(S) - \sum_{v=1}^{V} \frac{|S_v|}{|S|} Gini(S_v) \quad (3)$$

$$Gini(S) = 1 - \sum_{k=1}^{C} p_k^2 \tag{4}$$

$Gini(S)$ is *Gini* index of S and $Gini(S, A_i)$ represents the change of *Gini* index before and after S being classified by A_i. C means there are C types of data samples in datasets, and p_k is the ratio of the k type of samples. According to the maximum criteria of $Gini(S, A_i)$, the CART decision tree selects the corresponding A_i as classify property. However, when there is a missing value in D, Equations (3) and (4) cannot be applied directly. If forced cleaning of missing samples is implemented, massive sample data loss might occur and model accuracy will be influenced.

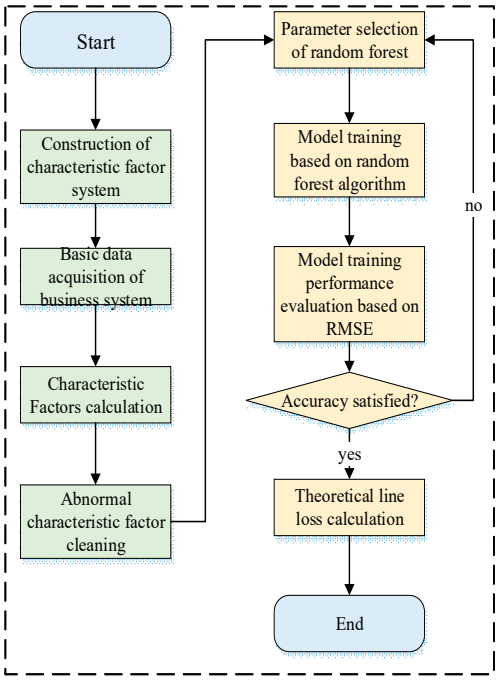

Figure 6. Theoretical line loss calculation of low-voltage grids based on random forest.

Therefore, optimization and modification for the basic CART decision tree are considered. Assume S_i is a subset of S organized by evaluation samples without missing values in A_i. There are K types of data samples and $S_{ic}^k (k = 1, 2, ..., K)$ is the k type of subset in S_i. S_i has V values in A_i and they are $A_i^1, A_i^2, ..., A_i^V$. According to these values, S_i is classified into V subsets as $S_i^1, S_i^2, ..., S_i^V$. Define weight W_j, $j = 1, 2, \ldots\ldots, m$ for each evaluation data sample X_j and define

$$p_i^k = \frac{\sum X_j \in S_{ic}^k w_j}{\sum X_j \in S_i w_j} (k = 1, 2, ..., K) \tag{5}$$

$$r_i^v = \frac{\sum X_j \in S_i^v w_j}{\sum X_j \in S_i w_j} (v = 1, 2, ..., V) \tag{6}$$

p_i^k is the ratio of the kth type of data sample in S_i and r_i^v is the ratio of A_i with value of A_i^k in S_i. Based on above definitions, Equations (3) and (4) can be modified to:

$$Gini\_ratio(S_i, A_i) = Gini(S_i) - \sum_{v=1}^{V} r_i^v Gini(S_i^v) \tag{7}$$

$$Gini(S_i) = 1 - \sum_{k=1}^{K} (p_i^k)^2 \tag{8}$$

With consideration of the sample missing, define the influence factor of weight W_i:

$$W_i = \frac{\sum X_j \in S_i w_j}{\sum X_j \in S w_j} (k = 1, 2, ..., d) \tag{9}$$

W_i represents the weight of samples without missing values in total samples.

In summary, when there are missing values in the evaluation data, the selection equation of the decision tree classify property can be modified to

$$Gini\_ratio(S, A_i) = W_i * (Gini(S_i) - \sum_{v=1}^{V} r_i^v Gini(S_i^v)) \tag{10}$$

With this modification, the previous equation for optimum classify property selection can also be used when there is a missing value in the evaluation data. Meanwhile, the influence of optimum classify property selection can be comprehensively considered with data samples having missing values or not. The topological structure of improved random forest by modifying decision tree is shown in Figure 7.

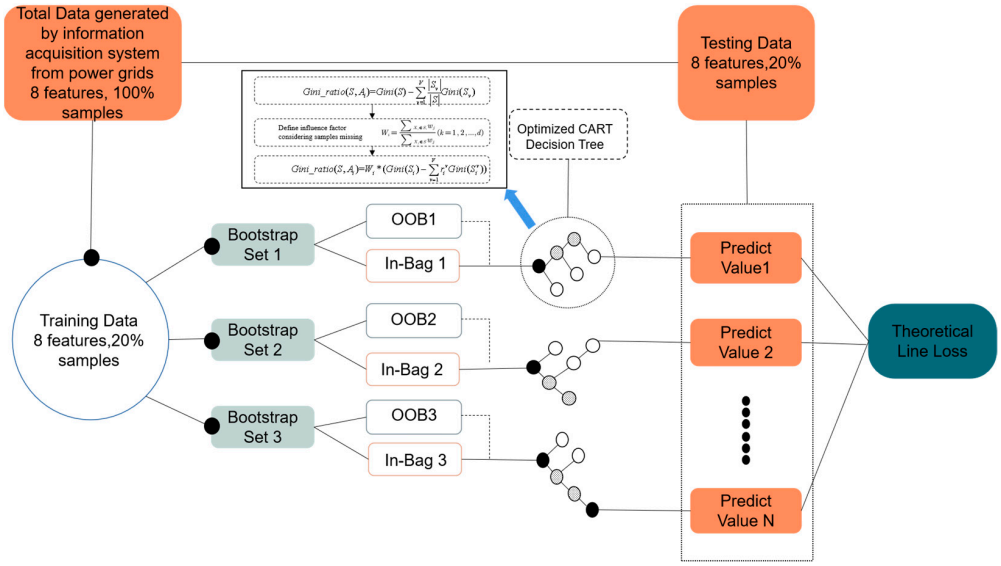

Figure 7. Topological structure of improved RF by modifying decision tree.

3.3. Evaluation of the Algorithm

For the evaluation of the algorithm, the training samples are firstly put into the algorithm, and the corresponding model parameters are trained, including the number of leaf nodes and the number of decision trees in random forest. Furthermore, the root mean square error (*RMSE*) of test samples is used to measure the accuracy of the model.

RMSE is the residual sum of squares of all calculated values, followed by the square root, which is used to indicate the accuracy of the calculated values. It can be expressed by Equation (11):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{c_i} - y_{o_i})^2} \tag{11}$$

Where n is number of test samples, y_{c_i} is the calculated value of the model, and y_{o_i} is observed value. The closer the *RMSE* is to 0, the higher the accuracy of the model.

Meanwhile, the overall calculation can be measured by the distribution of calculated values and the observed values of test samples, fitting the linear relationship between them and measuring the correlation coefficient between calculated values and observed values. Theoretically, the coefficient is between -1 and 1. The closer the coefficient is to 1, the better

the linear relationship between the calculated value and the observed value, the smaller the overall difference between the two values, and the better the calculation accuracy of the model.

4. Results and Discussion

4.1. Data Preparation

Take the line loss calculation of 23,754 low-voltage grids in a specific area as an example. Characteristic factors of each low-voltage grid are calculated once a day and the derived results are treated as one sample record. In the end, 166,283 samples are accumulated from 7 continuous days. After cleaning, these samples will be divided into two parts: one part is the training sample, accounting for 80%, and the other part is the test sample, accounting for 20%. The accuracy of the model was evaluated by the RMSE of the test sample.

During the process of abnormal data cleaning, it is found that the calculated data of characteristic factors present lots of abnormalities, due to data collection issues and non-uniform data quality. Furthermore, factors like power supply radius and low voltage line length are missing in some low-voltage grids due to different documentation management levels in different grids. Therefore, a reasonable range of various characteristic factors is formed with considerations of electrical characteristic calculation principles and low-voltage grid design regulations. Meanwhile, abnormal sample data out of range are cleaned. The cleaning and screening conditions of each characteristic factor are shown in Table 3.

Table 3. Cleaning rules of characteristic factors.

Characteristic Factors	Cleaning Rules
Power supply radius X_1	[200, 800]
Total line length X_2	[500, 10,000]
User numbers X_3	[50, 500]
Load rate X_4	[5, 60]
Three-phase unbalance degree X_5	[0, 200]
Load shape factor X_6	(0, 20]
Power factor X_7	[0.8, 1]
Power supply torque X_8	[0, 16,000]

After cleaning, 6708 data samples are retained, which only constitute 4.03% of original low-voltage grid samples. This is due to massive sample abnormalities and missing data caused by data management level issues in the original sample sets.

4.2. Analysis Based on Traditional Random Forest in High Cleaning Rate

Feed the cleaned sample data as original data into the random forest algorithm model. Select the numbers of leaf nodes and decision trees in the random forest model with RMSE as the evaluation criteria. Seen from the figure, when the number of decision trees increases, the root mean square error of the model becomes smaller, and the decreasing trend gradually slows down, but the complexity of the model does so only gradually. Meanwhile, when the number of decision trees are the same, the RMSE of the model will become smaller with fewer leaf nodes. Finally, considering the accuracy of the model and the complexity of the model, the optimal number of leaf nodes and decision trees are selected. In this situation, the final leaf nodes' number is 5 and the decision tree's number is 85, while optimum RMSE is 1.3239, as shown in Figure 8.

Figure 9 shows the distribution of the model calculated value and the observed value of the test samples, with the relatively obvious linear correlation between the model calculated value and the observed value shown. The correlation coefficient is only 0.4522. The experimental results demonstrate that the prediction accuracy is acceptable. Excellent calculation accuracy is obtained in a high distribution density range of [1, 5], while calculation accuracy is low in a low distribution density range of [5, 8]. This results from massive samples being removed in the low distribution density range during the process of force cleaning.

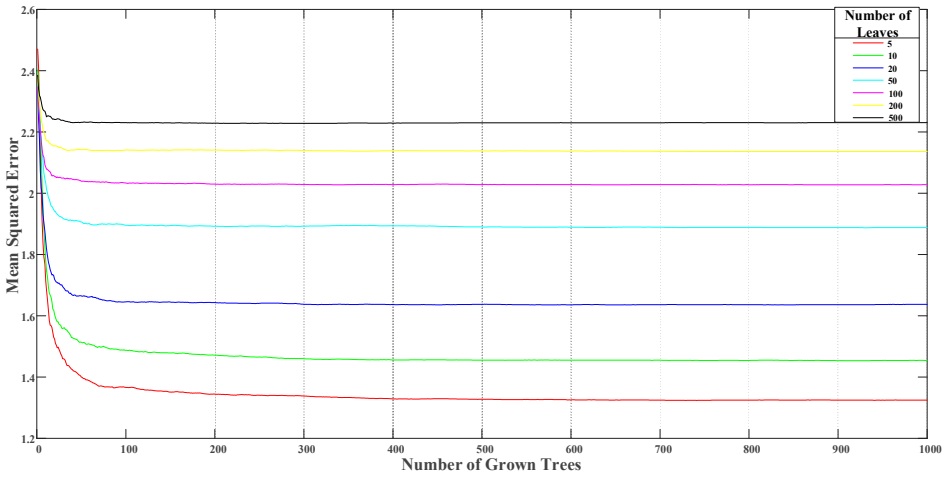

Figure 8. Parameter selection of random forest (cleaning rate 95.97%).

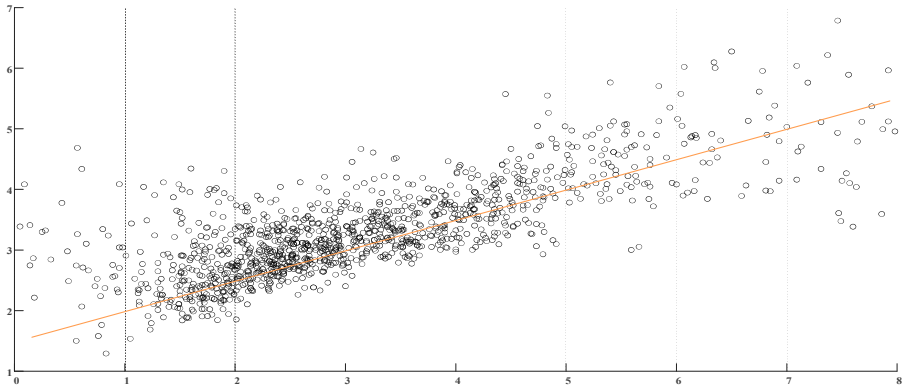

Figure 9. Calculation results of random forest (cleaning rate 95.97%).

4.3. Analysis Based on Traditional Random Forest in Lower Cleaning Rate

Therefore, samples need to be retained as much as possible considering the calculation results of various characteristic factors. As in Figure 10, by analyzing the distribution of various characteristic factors, it is found that the line length of 17% low-voltage grids is zero, which obviously is not true. Meanwhile, a power factor of 33% low-voltage grids is zero while load shape factor of 14.75% low-voltage grids is also zero, which are against practical operation rules of low-voltage grids. A three-phase unbalance degree of 4.5% low-voltage grids is, besides, actually higher than 200, which does not comply with the calculation principle of the three-phase unbalance degree in low-voltage grids.

Although data missing and abnormality exist in the mass of sample sets, much useful information disappears if a forced cleaning strategy is implemented, and model accuracy is influenced. By removing abnormal samples not complying with characteristic calculation principles and preserving those samples with partial properties, 89,067 of data samples are retained, which constitute 53.56% of the total samples. The cleaned sample data are fed into the modified random forest algorithm model. As shown in Figure 11, the leaf node number is set to 5, while the number of decision trees is 110 and optimum RMSE is 1.7319. Model fitting error is much higher than when the forced cleaning strategy is used.

As shown in Figure 12, from the distribution diagram of the model calculated value and the observed value of test samples, a certain linear correlation between the model calculated value and the observed value is shown. The correlation coefficient is only 0.4166, representing the relatively low accuracy of the model prediction. Good calculation accuracy is achieved in the high distribution density range of [1, 4] and a lack of the minimum

accuracy in the range of [0, 1] and [4, 8]. The experimental results above demonstrate that the fitting accuracy in the low distribution density range is not improved when the cleaning rate of abnormal samples drops. On the contrary, the performance of the model is influenced due to the longer training time caused by the larger training sample size.

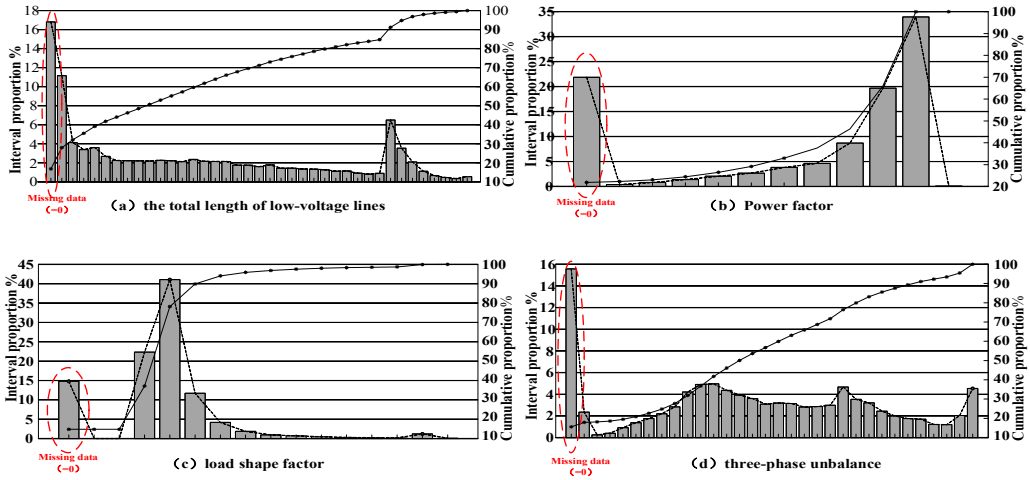

Figure 10. Distribution of characteristic factors.

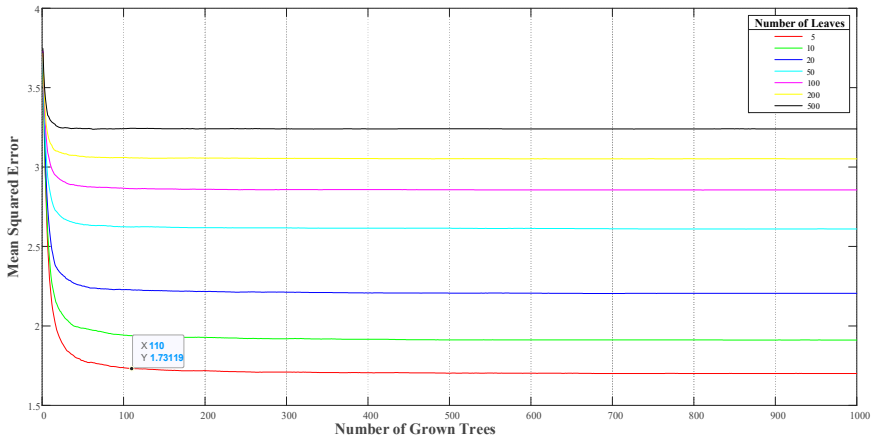

Figure 11. Parameter selection of random forest (cleaning rate 46.44%).

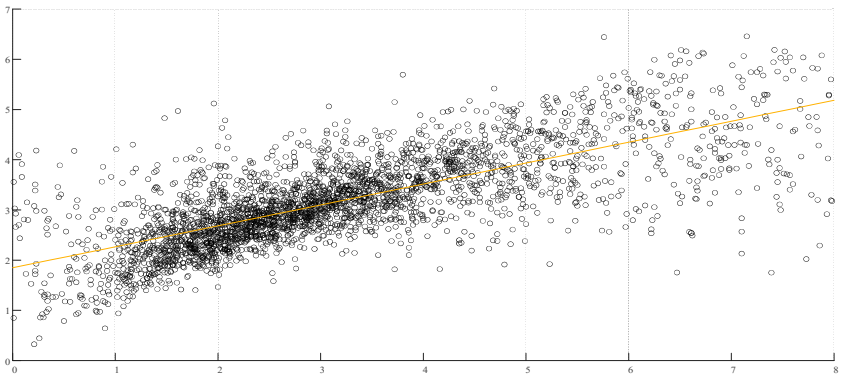

Figure 12. Calculation results based on random forest (cleaning rate 46.44%).

4.4. Analysis Based on Improved Random Forest in Lower Cleaning Rate

When the cleaned sample data as the original data is fed into the modified random forest algorithm model, as shown in Figure 13, the leaf node number is set to 5, while the number of decision trees is 75, and optimum RMSE is 1.2639.

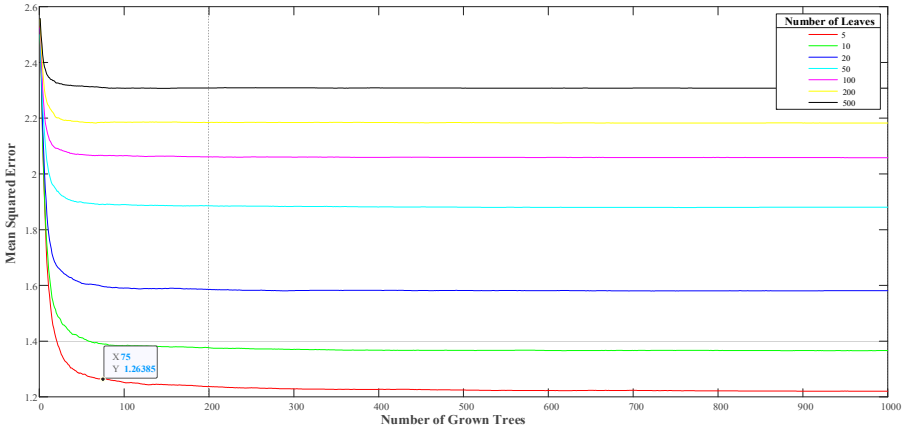

Figure 13. Parameter selection of improved random forest (cleaning rate 46.44%).

The calculated value and observed value of test samples are distributed in Figure 14. Obviously, there is a linear correlation between the model calculated value and the observed value. The correlation coefficient is 0.6733. Excellent linear correlation features appear both in the high distribution density range of [1, 5] and the low distribution density range of [5, 8]. It is verified that the model prediction performance can meet the accuracy requirement of the theoretical line loss model for low-voltage grids.

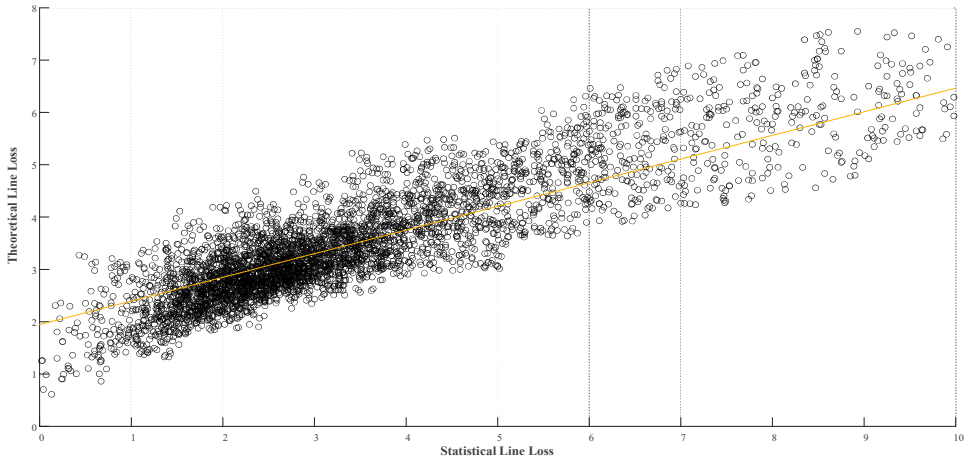

Figure 14. Calculation results based on modified random forest (cleaning rate 46.44%).

4.5. Discussion

As demonstrated by the results, the improved random forest method, by optimizing the decision tree, can help improve the accuracy of the theoretical line loss calculation when samples are missing in large amounts. The results of the abovementioned methods are compared in Table 4.

Table 4. Results of different methods.

	Cleaning Rate	RMSE	Correlation Coefficient
BPNN	95.97%	1.9875	0.4021
SVM	95.97%	1.7621	0.4978
KNN	95.97%	2.0255	0.3687
RF	95.97%	1.3239	0.4522
RF	46.44%	1.7319	0.4366
Proposed Method	46.44%	1.2639	0.6733

The results can be summarized as follows:

(1) According to definitions and calculation principles of electrical characteristics for low-voltage grids, a reasonable range of various characteristic are formed. The cleaning of abnormal sample data out of range is then performed. With the change of the sample data cleaning rule, model training effects using the random forest algorithm under the cleaning rates of 95.57% and 46.44% are compared. The accuracy errors of the model are 1.3239 and 1.7319, respectively.

(2) The issue of the characteristic factor missing using modified random forest algorithm is solved. Furthermore, the model is trained by the modified random algorithm, and model accuracy error is only 1.2161 compared to other approaches when the sample data cleaning rate is 46.44%.

(3) Correlation between the model calculated value and the observed value reached 0.6711 when the improved random forest algorithm was used in the situation of a lower sample cleaning rate, which was much higher than the other two situations (0.4522 and 0.4366). Meanwhile, it can also show the good calculation accuracy of improved random forest algorithm in different line loss intervals.

(4) More characteristic samples can be preserved when using the modified random forest algorithm to deal with samples featuring characteristics missing than by using forced cleaning. Therefore, better accuracy can be obtained during model training and calculation, which demonstrates that it is more effective to calculate and analyze low-voltage grids' theoretical line loss using the method proposed in this paper.

5. Conclusions

In this paper, an improved random forest method was proposed for the calculation of theoretical line loss in low-voltage grids. The main work of this paper included the following:

(1) The reasonable electric characteristic factors of the low-voltage grids were constructed according to the physical and operational influencing mechanism of theoretical line loss. The concept of power supply torque was proposed for the first time to analyze the influence mechanism more accurately by coupling the physical factors and the operational factors.

(2) The random forest algorithm was improved by modifying the property classifying process of the decision tree and optimizing the weight factor allocation method when sample data is missing. The problems of a high characteristic data integrity requirement was solved and the accuracy of the model was improved when a large amount of samples are missing. When the sample data cleaning rate changes from 95.57% to 46.44%, the accuracy of the traditional random forest increases from 1.3239 to 1.7319. However, the accuracy error of the improved random forest is only 1.2161 when the classifying process of the decision tree is modified.

We conclude that the proposed method can more accurately calculate the theoretical line loss of low-voltage grids when samples are missing in a large amount due to the different management of basic data in different areas. Further work may aim at optimizing the parameters of the algorithm model according to other data problems constantly found in practical application so as to improve the accuracy of the theoretical line loss and guide the actual loss reduction work more accurately.

Author Contributions: Conceptualization, L.H. and G.Z.; methodology, L.H.; validation, J.Z., Y.Z. and L.L.; formal analysis, L.H.; investigation, L.H.; resources, J.Z.; writing—review and editing, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key-Area Research and Development Program of Guangdong Province under No. 2020B0101130023.

Data Availability Statement: Available when request.

Acknowledgments: The authors would like to express their gratitude for the valuable recommendations made by the reviewers to improve the quality of this paper.

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

References

1. Ding, X.H.; Luo, Y.F.; Liu, W.; Shi, L.Z. Proposals on improving the current methods for calculating line losses of distribution network. *Autom. Electr. Power Syst.* **2001**, *25*, 57–60.
2. Fu, X.Q.; Chen, H.Y. Energy losses estimation using equivalent time of average current loss method. *Trans. China Electrotech. Soc.* **2015**, *30*, 377–382.
3. Zhang, K.K.; Yang, X.Y.; Bu, C.R.; Ru, W.; Liu, C.J.; Yang, Y.; Chen, Y. Theoretical analysis on distribution network loss based on load measurement and counter measures to reduce the loss. *Proc. CSEE* **2013**, *33*, 59–63.
4. Liu, T.L.; Wang, S.; Zhang, Z.; Zhu, J.F. Newton-Raphson method for theoretical line loss calculation of low-voltage distribution transformer district by using the load electrical energy. *Power Syst. Prot. Control* **2015**, *43*, 143–148.
5. Zhang, Y.; Wu, Y.F.; Zhang, F.; Yao, X.D.; Liu, A.; Tang, L.; Mo, J.G. A real-time three-phase line loss calculation method for distribution network based on feeder terminal unit. *Energy Rep.* **2022**, *8*, 146–152.
6. Zhang, Y.; Zhu, Y.; Bai, X.Q.; Hua, W. CIM-based Data-sharing Scheme for Online Calculation of Theoretical Line Loss. *Energy Procedia* **2012**, *16*, 1619–1626. [CrossRef]
7. Marcio, A.R.; André, L.V.G.; Miguel, E.M.U.; Eduardo, C.G.; Leonardo, M.O.Q. Technical loss estimation approach in power distribution systems using load model in frequency domain. *Electr. Power Syst. Res.* **2022**, *209*, 107982.
8. Bao, H.; Ma, Q. Physical distribution mechanism of network loss for power system. *Proc. CSEE* **2005**, *25*, 82–86.
9. Pablo, A.; Matias, A.K.; Selina, K. Flexibility management in the low-voltage distribution grid as a tool in the process of decarbonization through electrification. *Energy Rep.* **2022**, *8*, 248–256.
10. Kim, Y.J. Development and analysis of a sensitivity matrix of a three-phase voltage unbalance factor. *IEEE Trans. Power Syst.* **2018**, *33*, 3192–3195. [CrossRef]
11. Dai, Z.; Lin, W. Adaptive estimation of three-phase grid voltage parameters under unbalanced faults and harmonic disturbances. *IEEE Trans. Power Electron.* **2017**, *32*, 5613–5627. [CrossRef]
12. Karami, E.; Gharehpetian, G.B.; Madrigal, M.; Chavez, J.D.J. Dynamic phasor-based analysis of unbalanced three-phase systems in presence of harmonic distortion. *IEEE Trans. Power Syst.* **2018**, *33*, 6642–6654. [CrossRef]
13. Tan, Y.; Wang, Z. Incorporating unbalanced operation constraints of three-phase distributed generation. *IEEE Trans. Power Syst.* **2019**, *34*, 2449–2452. [CrossRef]
14. Xu, C.; Song, X.; Tao, Y.; Yang, Q.Q. Research on influencing factors of line loss rate of regional distribution network based on apriori-interpretative structural model. *Energy Rep.* **2022**, *8*, 53–64.
15. Xi, C.; Song, C.H.; Wang, T.R. Spatiotemporal analysis of line loss rate: A case study in China. *Energy Rep.* **2021**, *7*, 7048–7059.
16. Wen, F.S.; Han, Z.X. The calculation of energy losses in distribution systems based upon a clustering algorithm and an artificial neutral network model. *Proc. CSEE* **1993**, *13*, 41–50.
17. Jiang, H.L.; An, M.; Liu, X.J.; Zhao, X.; Zhang, J.H. The calculation of energy losses in distribution systems based on RBF network with dynamic clustering algorithm. *Proc. CSEE* **2005**, *25*, 35–39.
18. Li, Y.; Liu, L.P.; Li, B.Q.; Yi, J.; Wang, Z.Z.; Tian, S.M. Calculation of Line Loss Rate in Transformer District Based on Improved K-Means Clustering Algorithm and BP Neural Network. *Proc. Chin. Soc. Electr. Eng.* **2016**, *36*, 4543–4552.
19. Ma, L.Y.; Liu, J.H.; Lu, Z.G.; Wang, H.Y.; Yuan, Q.F.; Yang, L.P. Theoretical line loss calculation method of low voltage transform district based on deep belief network. *Electr. Power Autom. Equip.* **2020**, *40*, 7.
20. Wang, S.X.; Zhou, K.; Su, Y. Line loss rate estimation method of transformer district based on random forest algorithm. *Electr. Power Autom. Equip.* **2017**, *37*, 39–45.
21. Zhao, Q.M. The Calculation of Line Loss Rate in Transformer District Based on Affinity Propagation Algorithm and Random Forest Regression. *Proc. CSU-EPSA* **2020**, *32*, 94–98.
22. Hu, W.; Guo, Q.; Wang, W.; Wang, W.; Song, S. Loss reduction strategy and evaluation system based on reasonable line loss interval of transformer area. *Appl. Energy* **2022**, *306*, 118123. [CrossRef]
23. Bernard, S.; Adam, S.; Heutte, L. Dynamic random forests. *Pattern Recognit. Lett.* **2012**, *33*, 1580–1586. [CrossRef]
24. Bonissone, P.; Cadenas, J.M.; Garrido, M.C.; DíAzvalladares, R.A. A fuzzy random forest. *Int. J. Approx. Reason.* **2010**, *51*, 729–747. [CrossRef]

25. Ibrahim, I.A.; Khatib, T. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. *Energy Convers. Manag.* **2017**, *138*, 413–425. [CrossRef]
26. Ristin, M.; Guillaumin, M.; Gall, J.; Van, G.L. Incremental learning of random forests for large-scale image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 490–503. [CrossRef]
27. Anaissi, A.; Kennedy, P.J.; Goyal, M.; Catchpoole, D.R. A balanced iterative random forest for gene selection from microarray data. *BMC Bioinform.* **2013**, *14*, 261. [CrossRef]