applied sciences

MDPI

Article

Defect Prediction for Capacitive Equipment in Power System

Qingjun Peng [1,†], Zezhong Zheng [2,*,†] and Hao Hu [2,†]

[1] Electric Power Research Institute of Yunnan Power Grid Corporation, Kunming 650127, China; 13648716143@139.com
[2] School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China; hh222471@163.com
* Correspondence: zezhongzheng@uestc.edu.cn
† These authors contributed equally to this work.

Abstract: As a core component of the smart grid, capacitive equipment plays a critical role in modern power systems. When defects occur, they pose a significant threat to the safety of both other equipment and personnel. Hence, it is of great significance to predict whether defects occur in capacitive equipment in advance. To achieve this goal, we propose a novel method that integrates the weight of evidence (WOE) feature encoding with machine learning (ML). Five models, including support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), multi-layer perceptron (MLP), and linear classification, are employed with WOE features for defect prediction. Furthermore, based on the prediction of equipment with defects, an additional prediction is conducted to determine the potential defect level of the equipment. Experimental results demonstrate that the performance of each algorithm significantly improves with WOE encoding features. Particularly, the RF model with WOE encoding features exhibits optimal performance. In conclusion, the proposed method offers a promising solution for predicting the occurrence of defects and the corresponding defect levels of capacitive equipment. It enables relevant personnel to focus on and inspect equipment predicted to be at risk of defects, thereby preventing major malfunctions.

Keywords: defect prediction; capacitive equipment; WOE encoding; machine learning

Citation: Peng, Q.; Zheng, Z.; Hu, H. Defect Prediction for Capacitive Equipment in Power System. *Appl. Sci.* **2024**, *14*, 1968. https://doi.org/10.3390/app14051968

Academic Editor: Andreas Sumper

Received: 29 December 2023
Revised: 13 February 2024
Accepted: 20 February 2024
Published: 28 February 2024

1. Introduction

Capacitive equipment comprises an essential part of power transmission and transformation infrastructure, encompassing various components such as current transformers, bushings, coupling capacitors, and capacitive voltage transformers. These components represent a significant portion, comprising approximately 40% to 50% of the total equipment within a substation. The reliable operation and electrical safety of capacitive equipment are paramount for the smooth functioning of the substation. Any defects or malfunctions in this equipment can have severe repercussions, impacting the entire substation and potentially endangering surrounding equipment, resulting in substantial losses. The occurrence of such defects can be influenced by various factors, including the equipment's manufacturing date, operating environment, and topography [1].

Currently, research on defects in capacitive equipment primarily emphasizes online monitoring. Through online monitoring methods, personnel can promptly and accurately assess the equipment's condition by collecting monitoring data, enabling the real-time observation of capacitive equipment's actual operation, thus averting accidents caused by equipment defects. Annually, a large number of defective devices occur, and the power department has compiled statistics on such defect data, including information such as equipment type, time of defect occurrence, and type of defect. Therefore, we propose an approach which integrates these defect data and employing machine learning methods to investigate the relationship between device-related information and the occurrence of equipment defects, thereby establishing a foundation for subsequent equipment maintenance.

Undoubtedly, online monitoring methods play a crucial role in promptly detecting and addressing faults in capacitive equipment. Our proposed approach offers an alternative perspective for studying defects in capacitive equipment by leveraging historical defect information to train machine learning models. Subsequently, current device information is utilized as an input to predict the occurrence of device defects in advance, along with the potential defect severity levels. This enables relevant personnel to concentrate on monitoring the operational status of devices predicted to be at risk of experiencing defects. The key contributions of our work are summarized as follows:

- Unlike traditional methods relying on online monitoring and diagnostic techniques, we introduce a proactive approach. By utilizing machine learning algorithms, we predict whether defects will occur in capacitive equipment and their severity level before they manifest. This proactive prediction enables preemptive maintenance and intervention, ultimately enhancing the reliability and safety of the equipment.
- Successful application of the weight of evidence (WOE) feature encoding, based on the scorecard model, for preprocessing capacitive equipment data. This approach enhances the data preparation stage and improves the effectiveness of subsequent analysis.

The remainder of this paper is organized as follows. Section 2 reviews the current research status. Section 3 provides the process of constructing the model and related algorithms. Section 4 introduces the data preprocessing and model construction. Section 5 displays the experimental results and discusses the findings. Finally, Section 6 concludes our work.

2. Related Work

The early online monitoring of capacitive equipment primarily relied on manual inspections and periodic offline testing [2]. This approach incurred significant manpower and time costs, hindering the prompt detection of equipment faults and resulting in equipment damage and downtime. With the increasing prevalence and application scope of capacitive equipment, researchers began investigating parameter-based monitoring methods to evaluate equipment status and performance. The primary indicators monitored include the dielectric loss tangent, leakage current value, and capacitance value [3]. These indicators can be effectively utilized to discern early stage defects in capacitive equipment [4–7]. As sensor and computer technologies advanced, signal-processing-based monitoring methods emerged, analyzing signals from capacitive equipment such as current and voltage to evaluate equipment status and performance [8]. This method involves employing techniques such as spectral analysis and wavelet transform to extract features like amplitude, frequency, and partial discharge quantity for diagnosing faults in capacitive equipment. The application of frequency response analysis (FRA) technology for diagnosing faults in transformer bushings was proposed [9]. By integrating the magnitude and phase information of measured FRA features into a polar plot, more feature parameters were obtained compared to traditional amplitude plots, enabling the finer-level diagnosis of transformer bushing faults and the evaluation of insulation oil degradation status. A capacitor bushing fault diagnosis method based on high-frequency partial discharge measurement technology was proposed, addressing issues encountered in traditional diagnostic methods such as difficulties in detecting defects in the initial stage and determining fault types only when certain or multiple deteriorations have fully occurred [10].

With the continuous development of artificial intelligence (AI) technology, AI-based monitoring methods have garnered attention [11]. These methods utilize techniques such as machine learning and deep learning to analyze the characteristic parameters of capacitive equipment and diagnose them based on historical data and experience. The application of multilayer perceptron to capacitance tomography imaging sensor data exhibited promising performance in fluid classification [12]. Support vector regression (SVR) and artificial neural network (ANN) techniques were employed to identify and compensate for the effect of temperature on the output of capacitive differential pressure sensors [13]. The utilization of artificial neural networks (ANNs) has progressively enhanced the efficiency and effectiveness of power transformer fault diagnosis [14–19].

Online monitoring systems for capacitive equipment have evolved into intelligent systems, integrating various emerging technologies such as data acquisition, processing, communication transmission, and intelligent diagnosis [20]. These systems facilitate comprehensive, accurate, and real-time monitoring and the diagnosis of capacitive equipment, offering relevant warnings and suggestions for fault handling. An online monitoring system for capacitive-type equipment insulation was developed, with hardware based on DSP+FPGA, showing promising results in a simulation environment in a high-voltage laboratory [21]. To enhance the accuracy and reliability of leakage current measurement, an online insulation monitoring system for high-voltage capacitive substation equipment based on the Zigbee wireless sensor network was introduced [22]. The real-time monitoring of various capacitive equipment operating statuses can be achieved through the establishment of an intelligent auxiliary monitoring terminal [23].

While online monitoring enables personnel to make timely and accurate judgments on equipment status based on collected monitoring data, it is impractical to install sensors on all capacitive equipment for real-time monitoring due to limited resources, which would also consume significant manpower and resources. With advancements in technologies such as sensors and artificial intelligence, monitoring techniques for capacitive equipment are undergoing continual evolution. Nevertheless, there remains a scarcity of research concerning the utilization of historical defect data for modeling and analysis, which could enable defect prediction in capacitive equipment [24–28]. Our proposed methodology diverges from the norm by integrating equipment information, geographic data, substation details, etc., and employing machine learning techniques to investigate the correlation between these variables and defects in capacitive equipment. Subsequently, we construct machine learning models to forecast the likelihood of current equipment developing defects.

3. Method

To predict the occurrence and severity of defects in capacitive equipment using machine learning algorithms and identify the optimal defect prediction models, as shown in Figure 1, we initially analyzed and verified the importance of features that affected the occurrence and severity of defects. Subsequently, considering the characteristics of capacitive equipment data, we conducted data cleaning, feature encoding, and data balancing to prepare a dataset for model construction. Following this, we employed the random forest (RF) algorithm to develop a defect prediction model and compare its performance with four other machine learning algorithms. Ultimately, through comparative analysis of model performance, we identified the optimal prediction model.

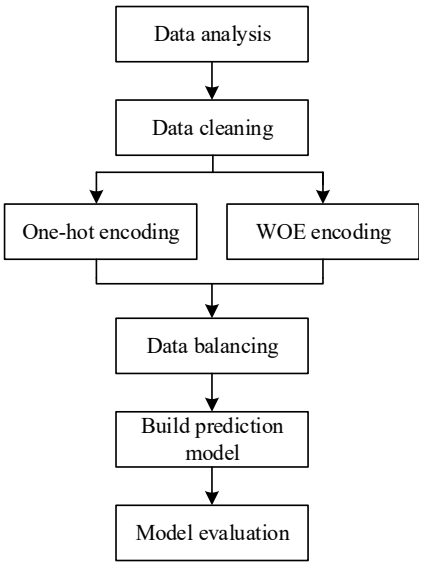

Figure 1. Diagram of the prediction process for defects in capacitive equipment.

3.1. WOE

Weight of evidence (WOE) [29] is a method used to quantify the impact of a specific variable value on the default rate. Widely employed in the financial domain, especially in constructing scorecard models, WOE serves as an encoding technique. For a specific feature that may possess categorical or continuous values, all dataset points are grouped into subgroups based on these feature values. Assuming each data point is associated with a 'Good' or 'Bad' target variable, the WOE value for the i^{th} subgroup is computed as follows:

$$\mathrm{WOE}_i = \ln\left[\frac{\frac{B_i}{B_T}}{\frac{G_i}{G_T}}\right] = \ln\left[\frac{\frac{B_i}{G_i}}{\frac{B_T}{G_T}}\right] \quad (1)$$

where B_i and B_T represent the number of bad data points in the i^{th} subgroup and the entire dataset, respectively, while G_i and G_T denote the number of good data points in the i^{th} subgroup and the entire dataset, respectively. If a value of the feature belongs to the i^{th} subgroup, the original feature value is replaced by WOE_i for subsequent modeling.

3.2. Random Forest

Random forest (RF) is a robust machine learning algorithm that combines multiple decision trees through the idea of ensemble learning [30]. RF is a particular form of the bootstrap aggregating (Bagging) method. Compared with Bagging, it exhibits three key distinctions:

- RF employs classification and regression tree (CART) algorithms as its constituent weak learners;
- RF randomly selects features every time;
- The number of samples selected by RF is identical to that of the training set. Due to its randomness, it can reduce the variance of the model. Therefore, RF exhibits superior generalization and antioverfitting capabilities compared to Bagging [31].

RF boasts rapid training speed, parallelizability, and high efficiency in processing large-scale data.

3.3. Comparison Models

- Liner Classifier: Linear classifiers categorize targets by linearly combining features. The model facilitates decision making by summing the product of each feature and its corresponding weight [32].
- MLP: Multilayer perceptron (MLP) is a forward-structured artificial neural network characterized by its layered structure. It can be conceptualized as a composition of multiple single-layer perceptron. The output layer of one perceptron serves as the input layer for the subsequent perceptron, with the final output layer representing the overall output of the MLP [33].
- SVM: Support vector machines (SVMs) are algorithms rooted in statistical theory, proficient in solving classification and regression problems with small-scale data. SVM addresses inner product operations in high-dimensional spaces by employing a kernel function, facilitating the effective implementation of nonlinear classification [34].
- XGBoost: Extreme gradient boosting (XGBoost) algorithm employs ensemble thinking to combine multiple weak learners into a strong learner through specific methodologies. XGBoost comprises multiple classification and regression trees (CARTs) and can handle diverse problems, including classification and regression [35].

4. Experiments

4.1. Data Collection and Preprocessing

During the operation of power systems, a significant amount of operational data and maintenance records are generated, forming a repository of historical data. These data comprise diverse information related to capacitive equipment, characterized by their

varied attributes. This study utilized data obtained from the power grid department, encompassing a broad spectrum of information, including equipment name, power supply bureau, latitude and longitude of the equipment, voltage level, production date, running state, etc. After sorting through the data, the following variables were selected for modeling:

- Equipment Name: Isolating switch, C phase current transformer, circuit breaker, etc.;
- Power Supply Bureau: The power supply bureau to which the equipment belongs, such as Kunming Power Supply Bureau (501) and Qujing Power Supply Bureau (502);
- Equipment Type: Optical current transformer, oil-filled transformer, DC current transformer, etc.;
- Full Name: The comprehensive name of the equipment along with its corresponding category, for instance, 'substation equipment/primary equipment/combined electrical appliance/COMPASS/current transformer';
- Equipment Type Remarks: The designation of equipment types, such as main transformer bushing (B A GT10 GT11 KH00) and current transformer (B A GG00 GG20 GT70);
- Equipment Model: The specific model information of the equipment, such as LZZBJ-35W, SZ11-4000/35, etc.;
- Manufacturer: The name of the equipment manufacturer;
- Topography: The geographic environment of the equipment's location, classified into six types: high mountain, hill, plain, river network, paddy field, and mountain, represented by numbers 1–6, respectively;
- Equipment longitude, latitude, and altitude;
- Pollution Level: The pollution level in the area where the equipment is situated, categorized into five levels;
- Substation: The substation to which the equipment belongs, e.g., 110 kV Lunan substation;
- Running State: Refers to the operational status of the substation, represented by the numbers 1–9, indicating operation, under construction, standby, etc.;
- Voltage Level: Indicates the rated voltage of the equipment, represented by the numbers 1–18, corresponding to voltage levels of 10,000 V, 110,000 V, 220,000 V, etc.;
- Voltage Type: Specifies whether the voltage is DC or AC, '1' indicates DC, '2' denotes AC, and '3' signifies that the voltage type is not distinguished. For example, 500,000 V voltage encompasses both AC and DC;
- Production Date and Commissioning Date: The date of the equipment leaving the factory and the date of the equipment being put into operation, respectively;
- Defect Occurrence Time: The timestamp when equipment defects occur;
- Years of Operation: For normal equipment, it represents the duration between the commissioning year and the current year. For faulty equipment, it signifies the duration between the commissioning year and the year the fault occurred;
- Defect Occurrence: A binary classification variable serving as the output for the defect occurrence prediction model, with two possible values: 'defect' and 'normal';
- Defect Level: A four-class variable used as the output for the defect level prediction model, comprising the following categories: 'urgent', 'critical', 'general', and 'others'.

The collected sample variables underwent statistical analysis, revealing 11,715 samples with defects and 648,288 normal samples. The defective samples were categorized into four levels, as depicted in Table 1. The defect level was determined through manual inspections of faulty equipment, categorizing them based on the severity of defects. 'Urgent' denotes equipment with critical defects posing immediate threats to both equipment and personnel safety, requiring urgent attention, whereas 'others' signifies equipment with minor defects having minimal impact and causing no disruption to the normal operations. Table 2 presents some typical defect data. The equipments are from Electric Power Research Institute of Yunnan Power Grid Corporation (Kunming, China).

Table 1. Sample size of different defect levels.

Level	Urgent	Critical	General	Others
Size	3275	1496	5894	1050

Table 2. Examples of original capacitive equipment defect data.

Power Supply Bureau	Voltage Level	Defect Level	Defect Type	Equipment Type	Manufacturer
502	10,000	General	Bird nest	Overhead conductor	Jinbei Electric Co., Ltd.
502	400	Critical	Insufficient safe distance	Low-voltage overhead conductor	Kunming Cable Group Co., Ltd.
502	35,000	Urgent	Low insulation	Isolating switch	Yunnan Yunkai Electric Co., Ltd.
502	110,000	General	Visible gas in the Buchholz relay	Oil-filled transformer	Jiangsu Huapeng Transformer Co., Ltd.

4.1.1. Data Cleaning

The raw data may contain a substantial number of invalid samples, and direct utilization in analysis and modeling may substantially impact the performance of the prediction model. Thus, data cleaning is essential. To understand the situation of missing features, we utilized Missingno [36], a visualization tool from the Pyecharts package, to graphically display the extent of missing features. The significance and extent of missing values were manually evaluated. Following this assessment, missing values in the samples were addressed. For instance, the latitude and longitude features of the equipment are crucial and should be retained whenever possible. In cases where these features have a relatively high missing rate, we match a missing value sample with high similarity to a complete dataset sample. Subsequently, we utilized the corresponding value from the complete dataset to fill in the missing value in the sample with missing data. The data format was then standardized for subsequent processing. For Chinese encoding, data storage documents were encoded in Chinese internal code specification (GBK) and saved as comma-separated value (CSV) files for ease of programming and readability. Subsequently, incorrectly recorded data were rectified, abnormal data were removed, and duplicate values were manually evaluated for retention or deletion.

After data cleaning, a total of 24 variables were selected as input variables, including equipment name, substation, equipment type, full name, equipment type remarks, equipment model, manufacturer, equipment longitude, equipment latitude, equipment altitude, running state, voltage level, voltage type, year, month, and day of production, week of the production date, year, month, and day of commissioning date, week of the commissioning date, year of operation, power supply bureau, and topography. The first seven variables are of string type, while the remaining variables are of floating-point or integer type.

4.1.2. Feature Encoding

Since the machine learning algorithm we employ exclusively deals with numerical data, it becomes crucial to perform feature encoding on each variable. There are two common encoding methods: one-hot encoding [37] and label encoding [38]. Label encoding preserves the original feature dimensions, conserves space, and minimizes information loss. It is noteworthy that label encoding may result in significant information loss if the sample order is altered. In contrast, one-hot encoding effectively handles categorical data and expands the feature space. For our study, we implement one-hot encoding and compare it with WOE encoding.

The procedure of WOE encoding is depicted in Figure 2. Initially, the cleaned dataset is read and divided into three equal parts, labeled as dataset D1, D2, and D3. The defect prediction model follows a binary classification approach, employing the output variable

'whether there is a defect' as the target variable to compute WOE using Equation (1). The defect level prediction model adopts a four-class classification model. However, the direct computation of WOE values for the four defect severity levels did not yield satisfactory results. Therefore, we transformed the four-class classification into binary classification for each defect level by establishing four '1 vs. rest' cases for WOE computation as follows: Four target variables were established, namely 'level_1', 'level_2', 'level_3', and 'level_4'. For 'level_1', a value of '1' was assigned to the defect level 'urgent', while the other three levels were assigned a value of '0' to calculate a WOE value for 'level_1'. This procedure was repeated for the other defect levels. Consequently, the original feature yielded four WOE values for the defect level prediction model.

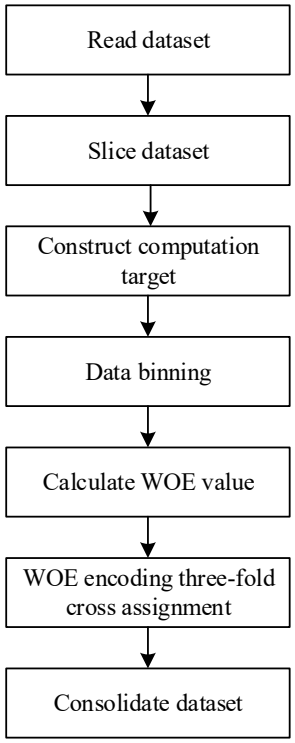

Figure 2. Flowchart of WOE feature encoding.

WOE computation was executed using a three-fold cross-assignment approach to mitigate overfitting, as shown in Figure 3. Specifically, the WOE values for the data points in dataset D3 were computed based on datasets D1 and D2. Likewise, the WOE values for D1 were determined based on D2 and D3, and so forth. Examples of WOE values for defect detection and defect level prediction are presented in Tables 3 and 4, respectively. Furthermore, Table 3 also provides some examples of one-hot encoding.

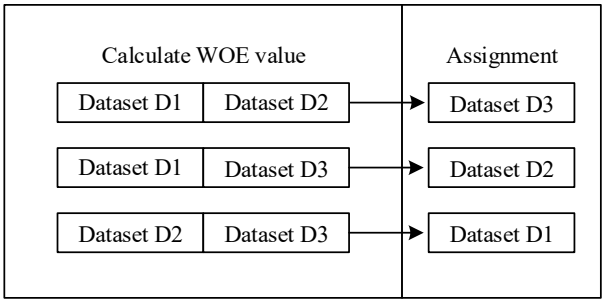

Figure 3. Three-fold cross-assignment.

Table 3. Two encoding results for power supply bureau feature in the defect occurrence prediction model.

Bureau	One-Hot Encoding	WOE Encoding
501	(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)	−0.012523
502	(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0)	−0.313688
503	(0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0)	0.571820
504	(0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0)	0.091393
505	(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0)	−0.229169
506	(0, 0,0,0,0,0,0,0,0,0,1,0,0,0,0,0)	−0.112641
507	(0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0)	0.034807
508	(0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0)	0.851002
509	(0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0)	0.136587
510	(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)	−0.052263
511	(0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0)	0.450399
512	(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0)	−0.483160
513	(0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0)	−0.954793
514	(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0)	0.526281
515	(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	1.029099
516	(1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	−0.939009
522	(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	−0.850388
581	(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)	−0.533197

Table 4. WOE encoding results for power supply bureau feature in the defect level prediction model.

Bureau	PSB_woe1	PSB_woe2	PSB_woe3	PSB_woe4
501	−0.320753	−0.434713	0.276863	0.359008
502	−0.320477	−0.174722	0.381339	−0.212646
503	0.605673	−1.446667	−0.698807	1.056935
504	−0.631471	0.821530	−0.337750	0.645300
505	0.204939	−0.212179	0.104558	−0.808023
506	−0.210769	−0.422391	0.293096	0.106945
507	0.228539	−0.118984	0.070539	−0.974570
508	0.620354	−0.900355	−0.329595	0.182347
509	0.041635	0.283675	0.085593	−1.448681
510	0.281236	0.786925	−0.430517	−1.793746
511	−1.677792	0.140298	0.551238	0.59351
512	0.064487	−0.088544	0.456459	−0.631664
513	0.119139	0.105754	0.193433	−3.619674
514	1.516963	0.225044	−1.452342	−2.123319
515	−2.454321	0.236879	0.14353	0.911634
516	0.030586	−0.136746	0.166161	−0.50136
522	−0.107284	−1.470198	0.723976	−1.016985
581	−0.367959	0.903592	−0.463598	0.428509

4.1.3. Data Balancing

Our datasets are imbalanced, with varying numbers of data points across different target variables. In the presence of such data imbalance, many machine learning classification algorithms may yield suboptimal results. To address this issue, several techniques, including sampling, data synthesis, weighting, and others, are commonly employed to balance the data [39]. However, these techniques have their inherent limitations. Sampling, for instance, may potentially compromise the model's generalization ability or lead to data loss. Similarly, determining appropriate weights for the weighting method can pose to be challenging. On the other hand, the data synthesis method aims to generate new data from existing data. In this study, we utilize the synthetic minority oversampling technique (SMOTE) [40], an improved algorithm based on the random oversampling algorithm. The basic idea of the SMOTE algorithm involves analyzing minority samples and synthesizing new samples based on them.

4.2. Defect Prediction Model Based on RF

In this experiment, the hardware configuration includes Intel(R) Xeon(R) Gold 5115 CPU @ 2.40Ghz (INTEL Corporation, Santa Clara, CA, USA), 32GB RAM, and NVIDIA RTX 1060 GPU (NVIDIA Corporation, Santa Clara, CA, USA), with Python as the programming language. The objective of this study is to forecast the occurrence of defects in capacitive equipment within a specified time period. The dataset utilized comprises both defect and nondefect data. Supervised learning is conducted using the RF algorithm, leveraging historical data of capacitive equipment to construct a defect occurrence prediction model. Parameter tuning is executed utilizing grid search and random search techniques. Additionally, a defect level model is established using the RF algorithm on the defect dataset, predicting four levels of defects: urgent, critical, general, and others. The same parameter tuning methodology employed for the defect occurrence prediction model is applied. To address the task of defect level prediction, the problem is transformed into four individual binary classification problems, as depicted in Figure 4.

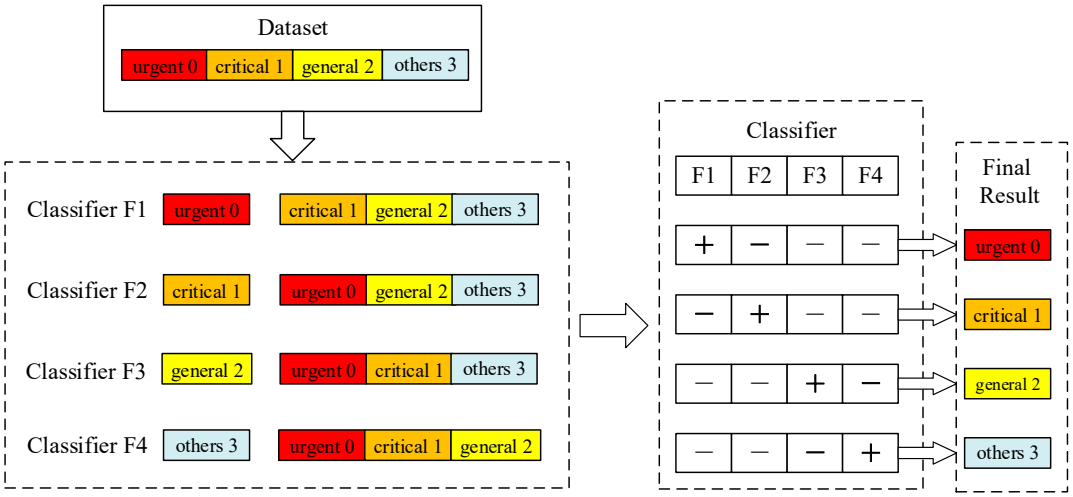

Figure 4. The strategy of transforming a four-classification problem into a binary classification problem.

During the training phase, four separate classifiers are trained using a one-vs-all approach. In each iteration, one class is considered to be the positive class, while all examples of the other classes serve as negative instances. For example, classifier F1 is trained to classify 'urgent' samples as a distinct class while grouping the remaining samples of 'critical', 'general', and 'others' into another class. During the testing phase, if only one classifier predicts a positive class, the corresponding class label is selected as the final classification result.

The defect occurrence prediction model based on RF can be trained using the RandomForestClassifier module from Scikit-learn library. Among all the parameters, 'Max_features' and 'N_estimators' exert the most significant impact on the prediction model. 'Max_features' determines the maximum number of features available for each individual decision tree, and increasing its value generally improves the model's performance. With 24 features in each sample, the decision tree can be trained using any combination of these features. 'N_estimators' refers to the number of subtrees to be created, and increasing its value can enhance the performance of the model. However, there is a saturation point beyond which the prediction accuracy will not further improve.

To obtain the best-performing model, it is necessary to determine the optimal values for 'Max_features' and 'N_estimators'. In this study, 70% of the samples are utilized as the training set, while the remaining 30% are used as the testing set. To enhance computational efficiency, 20,000 samples were randomly selected for parameter tuning. 'Max_features' is set to 'none', 'sqrt', and '15', while 'N_estimators' varies from 10 to 300. 'none' indicates no restriction on the maximum number of features and is set to 24. 'sqrt' represents the square

root of the maximum number of selected features and is set to four. The value '15' is self-set based on the characteristics of the capacitive data. The out-of-bag (OOB) error rate is used to measure the accuracy of the model, with lower error rates indicating higher accuracy. The OOB error curves for different numbers of subtrees are shown in Figures 5 and 6.

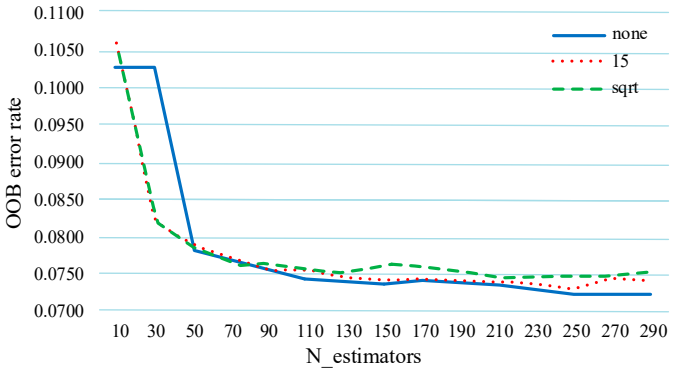

Figure 5. OOB error curve of RF algorithm in defect occurrence prediction under different parameters.

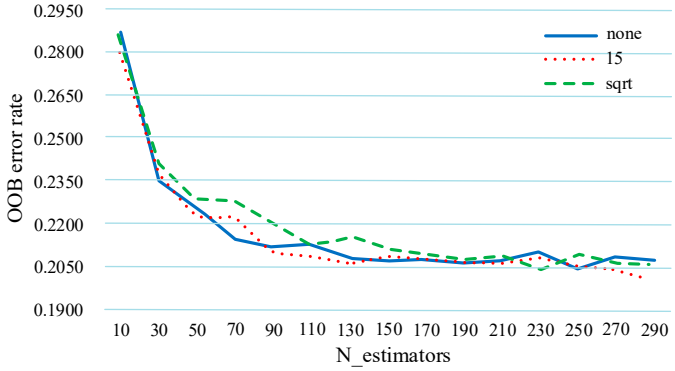

Figure 6. OOB error curve of RF algorithm in defect level prediction under different parameters.

Based on the findings presented in Figure 5, it can be observed that the defect occurrence prediction model tends to reach a state of convergence when the number of subtrees exceeds 110. Similarly, as depicted in Figure 6, the defect level prediction model exhibits a convergence trend when the number of subtrees surpasses approximately 130. Subsequently, further fine-tuning of the parameters is performed using grid search, resulting in the determination of the optimal parameter combinations. Finally, the defect occurrence prediction model demonstrates optimal performance with the following parameter settings: 'N_estimators' = 110 and 'Max_features' = 15. On the other hand, for the defect level prediction model based on RF, the optimal parameter configuration is found to be 'N_estimators' = 150 and 'Max_features' = 17.

4.3. Performance Metrics

In this study, four evaluation metrics are used to measure the prediction performance of the models, including accuracy, precision, recall, and F1-Score, calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

where *TP* refers to the number of positive examples that are correctly classified as positive examples, *FP* represents the number of negative examples that are incorrectly classified as positive examples, *TN* is the number of negative examples that are correctly classified as negative examples, and *FN* denotes the number of positive examples that are misclassified as negative examples.

5. Experiment Results and Discussion

5.1. Results of Defect Occurrence Prediction

The defect occurrence prediction model is employed to predict whether a device will exhibit defects within a certain time threshold, which is a binary classification model with predicted outcomes of defect and normal. To assess the prediction performance of the model, four evaluation metrics are used. In our work, the dataset was partitioned into training and testing sets with a ratio of 7:3. Based on WOE encoding and one-hot encoding, we utilized the RF algorithm to construct a prediction model. Additionally, MLP, SVM, XGBoost, and linear classification algorithms were used for experimental comparison. The defect occurrence prediction results of the five models on the testing set, based on one-hot encoding, are shown in Table 5. Meanwhile, the prediction results of the models utilizing WOE encoding are presented in Table 6.

Table 5. Evaluation of the defect occurrence prediction models based on one-hot encoding.

Models	Accuracy	Precision	Recall	F1-Score
Linear	0.72	0.95	0.64	0.82
XGBoost	0.85	0.98	0.84	0.90
SVM	0.74	0.98	0.70	0.83
MLP	0.86	0.98	0.83	0.91
RF	0.92	0.98	0.92	0.94

Table 6. Evaluation of the defect occurrence prediction models based on WOE encoding.

Models	Accuracy	Precision	Recall	F1-Score
Linear	0.79	0.98	0.69	0.86
XGBoost	0.93	0.98	0.89	0.93
SVM	0.83	0.98	0.73	0.89
MLP	0.88	0.98	0.87	0.92
RF	0.96	0.98	0.97	0.97

The experimental results reveal that the performance of each algorithm is improved with WOE encoding. The accuracy of SVM, XGBoost, and linear classification improved by over 0.07, while MLP and RF improved by 0.02 and 0.04, respectively. These results underscore the superior performance of RF based on WOE, achieving an accuracy of 0.96. Consequently, the WOE_RF algorithm emerges as the optimal model for predicting the occurrence of defects in capacitive equipment.

5.2. Results of Defect Level Prediction

The defect level prediction model serves to further predict the potential levels of defects. This model categorizes the defects into levels of urgency, critical, general, and others utilizing the same evaluation indicators as the defect occurrence prediction model. The prediction results of the defect level prediction models on the testing set, employing one-hot encoding, are presented in Table 7. The prediction results of the five models utilizing WOE encoding are displayed in Table 8.

Table 7. Evaluation of the defect level prediction models based on one-hot encoding.

Models	Accuracy	Precision	Recall	F1-Score
Linear	0.44	0.43	0.44	0.42
XGBoost	0.61	0.60	0.61	0.60
SVM	0.55	0.54	0.55	0.54
MLP	0.61	0.61	0.60	0.61
RF	0.71	0.71	0.70	0.71

Table 8. Evaluation of the defect level prediction models based on WOE encoding.

Models	Accuracy	Precision	Recall	F1-Score
Linear	0.46	0.46	0.45	0.45
XGBoost	0.62	0.62	0.62	0.62
SVM	0.73	0.72	0.73	0.72
MLP	0.66	0.66	0.66	0.66
RF	0.78	0.79	0.78	0.78

After contrasting the data in both tables, we observed a notable enhancement in classification accuracy across all five algorithms for defect level prediction with the application of WOE encoding. Moreover, when compared to other algorithms, RF exhibited the most robust overall classification performance in predicting defect levels. The significance of the defect level prediction model lies in its ability to assist personnel in performing the precise maintenance of capacitive devices.

6. Conclusions

Research on capacitive equipment defects primarily focuses on real-time online detection using various technologies. However, due to limited resources, it is impractical to install sensors on all capacitive equipment for real-time monitoring, which would require substantial manpower and resources. Our proposed method takes a different approach by integrating device information, geographical data, and substation information, utilizing machine learning techniques to explore the relationship between these factors and capacitive equipment defects. Subsequently, we developed machine learning models to predict whether current equipment is likely to develop defects and the severity of these defects. We applied WOE encoding from the finance domain to the analysis of capacitive equipment data, resulting in improved model performance. The experimental results demonstrate a significant enhancement in the performance of five algorithms with the use of WOE encoding. Upon comparison, the RF model employing WOE encoding emerged as the optimal defect occurrence prediction model and level prediction model. These findings are reliable and offer valuable insights for power grid companies in their production processes. In future endeavors, exploring the combination of various algorithms could augment the classification ability of the prediction model. Additionally, further experiments employing alternative machine learning algorithms or deep learning methods could yield an even more effective defect prediction model for capacitive equipment.

Author Contributions: Conceptualization, Q.P. and Z.Z.; resources, Q.P.; writing—original draft preparation, H.H.; writing—review and editing, H.H. and Z.Z.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Science and Technology Project of Yunnan Province (grant number 202202AD080004).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: Author Qingjun Peng was employed by the company Electric Power Research Institute of Yunnan Power Grid Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relation-ships that could be construed as a potential conflict of interest.

References

1. Xie, C.; Peng, Q.; Zheng, Z.; Li, Z.; Wang, Z.; Li, M.; Jiang, L.; Liu, Q.; Li, X. Relationship between defects of capacitive equipment and geomorphology. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4508–4511.
2. Xu, Y.; Wang, J.; Liu, W.; Jiang, Y.; Xu, T.; Zhang, J. Online monitoring device for partial discharge in high-voltage switchgear based on capacitance coupling method. In Proceedings of the 2019 IEEE 3rd International Conference on Green Energy and Applications (ICGEA), Taiyuan, China, 16–18 March 2019; pp. 56–60.
3. Yang, Y. Study on measurement error test method of on-line insulation monitoring device for capacitive equipment. *Appl. Mech. Mater.* **2013**, *373–375*, 844–847. [CrossRef]
4. Lachman, M.F.; Walter, W.; von Guggenberg, P.A. Online diagnostics of high-voltage bushings and current transformers using the sum current method. *IEEE Trans. Power Deliv.* **2000**, *15*, 155–162. [CrossRef]
5. Zhang, H.; Tan, K.; Dong, F.; Wang, J. The analysis of on-line monitored results for capacitive type of equipment. In Proceedings of the 2001 International Symposium on Electrical Insulating Materials (ISEIM 2001). 2001 Asian Conference on Electrical Insulating Diagnosis (ACEID 2001). 33rd Symposium on Electrical and Ele, Himeji, Japan, 22–22 November 2001; pp. 805–808.
6. Chen, T.; Zhang, B.; Wang, J.; Liu, J. New on-line high precision tanδ monitoring system for capacitive equipment. *Automat. Electr. Power Syst.* **2004**, *15*, 67–70.
7. Gao, Q.; Ding, P.; Han, Y.; Geng, B. Development of distributed on-line monitoring system for dielectric loss tangent of high voltage capacitive apparatus. In Proceedings of the 2008 International Conference on Condition Monitoring and Diagnosis, Beijing, China, 21–24 April 2008; pp. 1179–1182.
8. Damião, L.; Guimarães, J.; Ferraz, G.; Bortoni, E.; Rossi, R.; Capelini, R.; Salustiano, R.; Tavares, E. Online monitoring of partial discharges in power transformers using capacitive coupling in the tap of condenser bushings. *Energies* **2020**, *13*, 4351. [CrossRef]
9. Aljohani, O.; Abu-Siada, A. Application of digital image processing to detect transformer bushing faults and oil degradation using FRA polar plot signature. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 428–436. [CrossRef]
10. Lee, J.G.; Koo, J.-H.; Han, K.-S.; Choi, W. Development of transformer bushing diagnosis system based on high frequency PD measurement. In Proceedings of the 2022 9th International Conference on Condition Monitoring and Diagnosis (CMD), Kitakyushu, Japan, 13–18 November 2022; pp. 372–375.
11. Yang, Z.; Zhang, Z.; Xue, W.; Chen, Y.; Mou, X.; Yang, Q. Study on power equipment condition based maintenance (CBM) technology in smart grid. In Proceedings of the 2021 3rd International Conference on Smart Power & Internet Energy Systems (SPIES), Shanghai, China, 25–28 September 2021; pp. 291–295.
12. Mokhtar, K.Z.; Saleh, J.M.; Talib, H.; Ali, N.O. Flow regime classification using artificial neural network trained on electrical capacitance tomography sensor data. *Comput. Inf. Sci.* **2008**, *1*, 25–32.
13. Hashemi, M.; Ghaisari, J.; Salighehdar, A. Identification and compensation of a capacitive differential pressure sensor based on support vector regression using particle swarm optimization. *Intell. Autom. Soft Comput.* **2012**, *18*, 263–277. [CrossRef]
14. Cheng, J.; Li, A.; Duan, Z. Transformer fault diagnosis based on improved evidence theory and neural network integrated method. *Power Syst. Prot. Control* **2013**, *41*, 92–96.
15. Ghoneim, S.S.M.; Taha, I.B. Artificial neural networks for power transformers fault diagnosis based on IEC code using dissolved gas analysis. *Int. J. Control Automat. Syst.* **2015**, *4*, 18–21.
16. Yang, X.; Chen, W.; Li, A.; Yang, C.; Xie, Z.; Dong, H. BA-PNN-based methods for power transformer fault diagnosis. *Adv. Eng. Inform.* **2019**, *39*, 178–185. [CrossRef]
17. Jin, Y.; Wu, H.; Zheng, J.; Zhang, J.; Liu, Z. Power transformer fault diagnosis based on improved BP neural network. *Electronics* **2023**, *12*, 3526. [CrossRef]
18. Yang, P.; Wang, T.; Yang, H.; Meng, C.; Zhang, H.; Cheng, L. The performance of electronic current transformer fault diagnosis model: Using an improved whale optimization algorithm and RBF neural network. *Electronics* **2023**, *12*, 1066. [CrossRef]
19. Fang, H.; Deng, J.; Chen, D.; Jiang, W.; Shao, S.; Tang, M.; Liu, J. You can get smaller: A lightweight self-activation convolution unit modified by transformer for fault diagnosis. *Adv. Eng. Inform.* **2023**, *55*, 101890. [CrossRef]
20. Shi, H.; Bai, C.; Xie, Y.; Li, W.; Zhang, H.; Liu, Y.; Zheng, Y.; Zhang, S.; Zhang, Y.; Lu, H.; et al. Capacitive–inductive magnetic plug sensor with high adaptability for online debris monitoring. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–8. [CrossRef]
21. Hao, X.; Zhang, G.; Zhang, W.; Dong, M.; Liu, G. Online monitoring technology for the insulation condition of capacitive-type substation equipment. In Proceedings of the 2008 International Conference on Condition Monitoring and Diagnosis, Beijing, China, 21–24 April 2008; pp. 1220–1223.
22. Li, J.; Jiao, S.; Wen, Y.; Wang, H. Online insulation monitoring system of high-voltage capacitive substation equipment based on WSN. In Proceedings of the CICED 2010 Proceedings, Nanjing, China, 13–16 September 2010; pp. 1–6.
23. Xu, M.; Guo, Y.; Zhang, D.; Cao, R.; Shen, Y.; Han, S. Research on high voltage online monitoring system for dielectric loss of capacitive equipment in substation. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *769*, 042025. [CrossRef]

24. Sun, C.; Bi, W.; Zhou, Q.; Liao, R.; Chen, W. New gray prediction parameter model and its application in electrical insulation fault prediction. *Control Theory Appl.* **2003**, *20*, 798–801.
25. Marino, P.; Sigiienza, C.; Poza, F.; Vazquez, F.; Machado, F. Supporting information system for power transformer fault forecasting applications. In Proceedings of the IECON'03. 29th Annual Conference of the IEEE Industrial Electronics Society (IEEE Cat. No. 03CH37468), Roanoke, VA, USA, 2–6 November 2003; Volume 2, pp. 1899–1904.
26. Xu, Z.; Peng, D.; Xu, P. Fault prediction of power electronics module based on RELM-AdaBoost. In Proceedings of the 2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE), Shenzhen, China, 27–29 May 2022; pp. 11–14.
27. Di, Y.; Jin, C.; Bagheri, B.; Shi, Z.; Ardakani, H.D.; Tang, Z.; Lee, J. Fault prediction of power electronics modules and systems under complex working conditions. *Comput. Ind.* **2018**, *97*, 1–9. [CrossRef]
28. Peng, J.; Zhou, F.; Xiang, H.; Ma, Y.; Zheng, Z.; Jiang, S.; Li, J. Prediction of defects occurrence time for capacitive device. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 14–15 December 2019; pp. 226–229.
29. Wu, Y.; Pan, Y. Application analysis of credit scoring of financial institutions based on machine learning model. *Complexity* **2021**, *2021*, 1–12. [CrossRef]
30. Surhone, L.; Tennoe, M.; Henssonow, S. Random forest. *Mach. Learn.* **2010**, *45*, 5–32.
31. Yu, X.; Hyyppa, J.; Vastaranta, M.; Holopainen, M.; Viitala, R. Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 28–37. [CrossRef]
32. Yuan, G.X.; Ho, C.H.; Lin, C.J. Recent advances of large-scale linear classification. *Proc. IEEE* **2012**, *100*, 2584–2603. [CrossRef]
33. Pal, S.; Mitra, S. Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **1992**, *3*, 683–697. [CrossRef] [PubMed]
34. Zhang, S.; Wang, Y.; Liu, M.; Bao, Z. Data-based line trip fault prediction in power systems using LSTM networks and SVM. *IEEE Access* **2017**, *6*, 7675–7686. [CrossRef]
35. Chen, M.; Liu, Q.; Chen, S.; Liu, Y.; Zhang, C.; Liu, R. XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. *IEEE Access* **2019**, *7*, 13149–13158. [CrossRef]
36. Bilogur, A. Missingno: A missing data visualization suite. *J. Open Source Softw.* **2018**, *3*, 547. [CrossRef]
37. Rodríguez, P.; Bautista, M.A.; Gonzàlez, J.; Escalera, S. Beyond one-hot encoding: Lower dimensional target embedding. *Image Vis. Comput.* **2018**, *75*, 21–31. [CrossRef]
38. Zhang, Y.; Schneider, J. Multi-Label output codes using canonical correlation analysis. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; Volume 15, pp. 873–882.
39. He, H.; Garcia, E. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
40. Blagus, R.; Lusa, L. Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. In Proceedings of the 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012; Volume 2, pp. 89–94.