

Article

Fault Diagnosis of Centrifugal Chiller Based on Extreme Gradient Boosting

Yaxiang Liu ¹, Tao Liang ¹, Mengxin Zhang ², Nijie Jing ², Yudong Xia ² and Qiang Ding ^{2,*}¹ Shandong Electric Power Engineering Consulting Institute Co., Ltd., Jinan 250013, China² Institute of Energy Utilization and Automation, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: dingqiang@hdu.edu.cn; Tel.: +86-571-86919133

Abstract: Centrifugal chillers have been widely used in medium- and large-scale air conditioning projects. However, equipment running with faults will result in additional energy consumption. Meanwhile, it is difficult to diagnose the minor faults of the equipment. Therefore, the Extreme Gradient Boost (XGBoost) algorithm was used to solve the above problem in this article. The ASHRAE RP-1043 dataset was employed for research, utilizing the feature splitting principle of XGBoost to reduce the data dimension to 23 dimensions. Subsequently, the five important parameters of the XGBoost algorithm were optimized using Multi-swarm Cooperative Particle Swarm Optimization (MSPSO). The minor fault diagnosis model, MSPSO-XGBoost, was established. The results show that the ability of the proposed MSPSO-XGBoost model to diagnose eight different states is uniform, and the diagnostic accuracy of the model reaches 99.67%. The accuracy rate is significantly improved compared to that of the support vector machine (SVM) and back propagation neural network (BPNN) diagnostic models.

Keywords: centrifugal chillers; fault diagnosis; extreme gradient boosting; MSPSO-XGBoost



Citation: Liu, Y.; Liang, T.; Zhang, M.; Jing, N.; Xia, Y.; Ding, Q. Fault Diagnosis of Centrifugal Chiller Based on Extreme Gradient Boosting. *Buildings* **2024**, *14*, 1835. <https://doi.org/10.3390/buildings14061835>

Academic Editors: Sungkyun Jung, Piljae Im and Yeoboom Yoon

Received: 9 April 2024

Revised: 8 June 2024

Accepted: 11 June 2024

Published: 17 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Centrifugal chillers have the characteristics of a high energy efficiency ratio and large single-machine capacity, which makes them the most used model in medium and large air-conditioning systems. Due to the continuous improvements in modern industrial levels, the structure and system of centrifugal chillers have become more complex [1]. This will increase energy consumption by 20% to 50% for the centrifugal chillers without timely troubleshooting [2]. Therefore, it is critical to use the micro-fault diagnosis technology in the early fault diagnosis of chillers. It can determine the type of fault occurrence quickly and accurately and shorten the maintenance time effectively. It plays a role in guaranteeing refrigeration efficiency, reducing equipment loss, and saving energy [3].

The minor fault diagnosis techniques may be classified into three types, namely engineering experience, mechanisms, and data-driven models. Profiting from the large data technology industry, fault diagnosis with data-driven models has become the mainstream method. Data-based fault diagnosis is essentially a search for the mapping relationship between the monitoring data and the unit state without relying on a priori knowledge of the system. However, for complex systems such as chillers, there is no simple correspondence between monitoring data and unit status [4]. In addition, the operating conditions of chiller units are varied, and certain operating parameters are highly similar in the micro-fault state and the no-fault state in incipient faults [5]. Due to the complexity and uncertainty of the system, machine learning algorithms like Artificial Neural Networks (ANNs) [6] and support vector machines (SVMs) [7] were applied to the diagnosis of minor faults, and some results have been achieved. In the study of chiller fault diagnosis, study [8] used an ANN to diagnose various faults in chillers and found that the diagnosis of system faults is more difficult than local faults. An SVM is used in chiller fault diagnosis with information

from factory-installed sensors. And, the experimental results show that the information from eight sensors can satisfy the needs of the diagnosis task [9].

SVMs [10] are based on statistical learning theory [11] and skilled in different classes of the small sample dataset by finding the optimal hyperplane. However, SVMs are not appropriate for training large sample datasets because they cannot manage the complicated nonlinear correlations that large-scale datasets—especially those with high dimensionality—present in the high-dimensional space [12]. ANNs [8,13] have strong robustness and flexibility, learning complex nonlinear relationships between data by adjusting the connection weights between neurons. However, ANNs usually obtain sub-optimal solutions and often obtain non-global minimum values, which means that during the training process, they may fall into local optima. Therefore, in order to obtain a suitable model, ANNs need a lot of data for training. ANNs are generally used for classification problems in the form of data such as images, text, etc., and are not optimal for dealing with the problem of classifying discrete table data.

To address the above problems, integrated learning algorithms can be used in the fields of fault diagnosis and signal classification [14]. In integrated learning algorithms, the type of data can be detected and classified by a weak classifier with the advantages of a fast training speed and the ease of adjusting parameters; however, the accuracy is not high. A strong classifier is constructed by combining multiple weak classifiers, thus improving the overall generalization performance. Typical integrated learning algorithms include Random Forest (RF) and Extreme Gradient Boosting (XGBoost). XGBoost is proposed by Chen [15] on the basis of a Boosting algorithm based on GBDT. XGBoost has better modelling capabilities, higher computational accuracy, and faster training speed [16], and is suitable for the classification of table data, compared to other data types such as images and signals. XGBoost has been popular in fields such as statistics, data mining, and machine learning since its release. As a typical representative of Boosting technology in ensemble learning, XGBoost can effectively handle large-scale machine learning tasks. Since the introduction of XGBoost, it has been widely used in various research fields, such as cancer diagnosis [17], credit risk assessment [18], and macro genomics [19], due to its performance advantages and affordable time and memory complexity. However, the predictive performance of XGBoost without parameter optimization is often unsatisfactory due to its low fit with the dataset, resulting in poor generalization and adaptability. Its predictive performance highly depends on the tuning of hyper-parameters. Therefore, it is necessary to propose an efficient method to optimize these hyper-parameters [20].

However, there is little research on using the XGBoost method in the fault diagnosis of chillers [21]. This research aims to study the XGBoost algorithm and its parameter optimization. For seven typical minor faults in centrifugal chillers, this paper builds an XGBoost diagnostic framework and analyzes the impact of hyper-parameters on model performance. It implements the optimization of the hyper-parameters using the MSPSO algorithm, thus establishing the MSPSO-XGBoost diagnostic model. Consequently, the MSPSO-XGBoost model is further compared and analyzed with the SVM and BPNN models to verify the effectiveness of the proposed method in the micro-fault diagnosis of chillers.

The paper is organized as follows. Firstly, a description of the principle content for the research process is introduced in Section 2. Then, the description of the implementation and validation process of fault diagnosis is introduced in Section 3. Finally, the main contributions of the current study are summarized in Section 4.

2. Basic Principles

2.1. XGBoost

The core idea of the Boosting framework is to form a stronger classifier by combining multiple weak classifiers. The weak classifier chosen by XGBoost is Classification and Regression Trees (CARTs).

As shown in Figure 1, XGBoost belongs to a type of boosting tree model. Initially, a tree independently predicts a value based on the actual value and then obtains the deviation

between the actual and forecast values. After that, the deviation is used as the actual value of the second tree, and new deviations are continuously obtained. When t trees are added,

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i), 0 < \eta < 1 \quad (1)$$

where $f_t(x_i)$ is the discriminant function of the t -th tree for the i -th datapoint, $\hat{y}_i^{(t)}$ is the discriminant result of the strong model integrated by t decision tree models, and η is the learning rate.

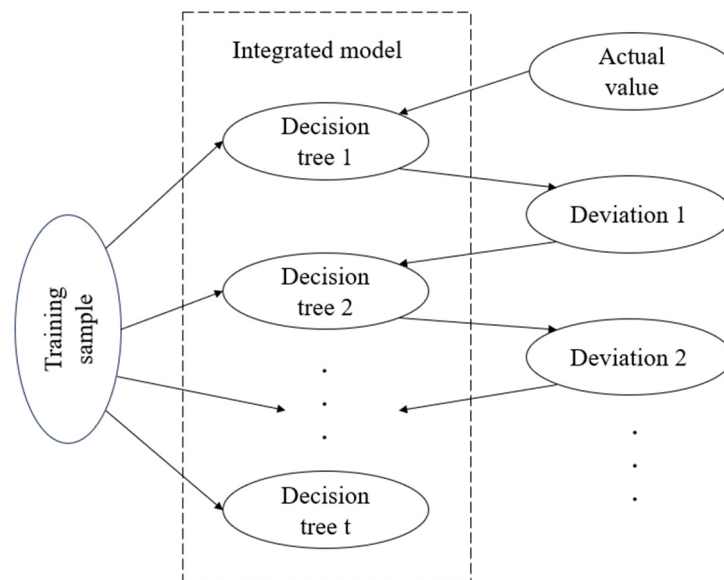


Figure 1. XGBoost principle.

The essence of CART is to construct a binary tree to divide the feature space; training CART is to find an optimal binary tree structure to make the minimum objective function. Similarly, the XGBoost objective function is defined as follows:

$$Obj(t) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{i=1}^t \Omega(f_i) \quad (2)$$

The first term on the right side of the formula is the deviation between the actual value and the predicted value, where n is the number of training samples. The last term on the right side of the formula is the regularization term, which mainly serves to suppress model complexity.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

where T is the depth of the current subtree and ω is the node value of the leaf node. γ represents the node segmentation threshold and λ is the L2 regularization coefficient.

According to the objective function, the optimal output obtained is as follows:

$$Obj^* = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (4)$$

$$G_j = \sum_{i \in I_j} \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (5)$$

$$H_j = \sum_{i \in I_j} \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (6)$$

Equation (4) is the rating function of the tree, with smaller values indicating better structure. XGBoost relies on the greedy strategy of recursive node splitting to generate sub-model trees. Equation (7) is the expression of the splitting profit Gain. When selecting the optimal splitting Gain, the feature with the largest difference in Obj^* values before and after splitting is considered, and the feature with the largest Gain value is used as the splitting point to construct a CART.

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (7)$$

The index L represents the optimal value of the objective function when the value is divided into the left subtree, and the index R represents the optimal value of the objective function when the value is divided into the right subtree.

When generating XGBoost models, it is possible to count the instances in which features engage in splitting, thereby establishing a ranking of feature contribution, because the process of developing CART adheres to inference rules and has practical physical significance and interpretability. Then, the top-ranked features are selected to construct low-dimensional training samples, which can reduce feature redundancy and improve model training speed.

2.2. Principles of MSPSO-XGBoost

The XGBoost algorithm is an efficient classification algorithm. The performance of the XGBoost model is highly dependent on the optimization of the hyper-parameters, and it has a large number of parameters. Hence, it is difficult to adjust using the empirical method and cannot achieve the global optimum [22]. Based on the basic principles of XGBoost and existing studies [22,23] on optimizing XGBoost parameters, five parameters were chosen that affect the classification accuracy significantly, as listed in Table 1.

Table 1. XGBoost algorithm partial parameter information table.

Parameter	Range	Describe
eta	[0, 1]	Learning rate: η , reducing the weight of each step. If a learning rate is too high or too low, it is impossible to find the position of the minimum loss function.
gamma	[0, ∞]	Gain threshold: γ , gamma specifies the loss reduction which is necessary to split tree nodes rightly in a loss function. In other words, it is a parameter that contributes to making an algorithm conservative.
max_depth	[0, ∞]	The maximum depth of the tree; the larger the max_depth value is, the more a model learns a very characteristic relation for a particular sample. The parameter is used to adjust over-fitting.
min_child_weight	[0, ∞]	The minimum weight of leaf nodes; when the value is large, the model can avoid learning the local optimal solution.
n_estimators	[100, 500]	The number of sub-model trees; if the value is too small, the problem of model underfitting will occur, and if the value is too large, the calculation amount will be greatly increased.

PSO [24] is a commonly used heuristic optimization algorithm that has the characteristics of a simple implementation mechanism, strong interpretability, and fast convergence speed. Its speed and position update formulas are as follows:

$$V_i^{t+1} = \omega V_i^t + c_1 r_1 (P_g^t - X_i^t) + c_2 r_2 (P_i^t - X_i^t) \quad (8)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (9)$$

where c_1 and c_2 are the learning factors, r_1 and r_2 are the random factors, ω is the inertia weight coefficient, V_i^t and X_i^t are the velocity and position of the particle at the current time,

V_i^{t+1} and X_i^{t+1} are the updated velocity and position, P_i^t is the historical optimal position of the i -th particle, and P_g^t is the optimal particle position of all the particles.

When facing complex problems, reducing ω according to inertia can easily trap the algorithm into local optima. Based on the literature [23,25,26], and inspired by the idea of group decision-making, MSPSO is employed to address this issue. The particle swarm is divided into multiple subgroups, each of which independently seeks optimization. After each iteration, the optimal particle information for each subgroup is shared, and the most adaptable particle is selected as the global guiding particle. When updating the speed, the global guide particles guide all particles, and the formula for updating the speed is as follows:

$$V_i^{t+1} = \omega V_i^t + c_1 r_1 (P_i^t - X_i^t) + c_2 r_2 (P_i^t - X_i^t) + c_3 r_3 (P_G^t - X_i^t) \quad (10)$$

where c_3 is the learning factor for the i -th particle to learn the global guide particle, r_3 is a random factor with values in the $[0, 1]$ interval, and P_G^t is the optimal position for the global guide particle.

The steps to optimize XGBoost using MSPSO are the following:

Step 1: First, initialize, set the number of particles to N , divide them equally into S subgroups, set the particle dimension to D , and iterate T times;

Step 2: Calculate the fitness value of particles at time $t = 0$, that is, the accuracy of each particle's corresponding model;

Step 3: S subgroup updates speed and position in parallel according to Equations (9) and (10);

Step 4: Process boundary constraints and calculate the fitness value of particles at time $t + 1$;

Step 5: Determines whether the current iteration count has reached its maximum setting. If so, end the optimization and output the model; otherwise, jump to step 3.

2.3. Evaluation Indicators and Model Establishment Process

The multi-classification confusion matrix is shown in Table 2. A to I show classification number (for example, B represents the number of misclassified datapoints in C1 to C2). Based on the confusion matrix, the accuracy rate (AR), precision rate (PR), and recall rate (RR) are defined.

Table 2. Confusion matrix.

		Diagnosed Faults		
		C1	C2	C3
True Faults	C1	A	B	C
	C2	D	E	F
	C3	G	H	I

Define AR to measure the overall classification performance of diagnostic models:

$$AR = (A + E + I) / (A + B + C + D + E + F + G + H + I) \quad (11)$$

Taking C1 as an example, define PR and RR to measure the diagnostic model's ability to separate and identify different fault states:

$$PR = A / (A + D + G) \quad (12)$$

$$RR = A / (A + B + C) \quad (13)$$

Figure 2 shows the flowchart for training a fault diagnosis model, which includes two parts: data preprocessing and the optimization of the model. Data preprocessing includes reading data from a database, adding labels, feature selection, and dataset splitting.

In the model optimization, the training set and particle positions are used to establish the model, while the test set is input into the model to obtain the confusion matrix, which can evaluate the model. Based on the overall diagnostic accuracy of the model, the performance is evaluated. The evaluation indicators of the model are taken as the fitness function of the particles and the particle positions are adjusted according to Equations (9) and (10). The particle positions are repeatedly updated to establish the diagnostic model until the maximum number of iterations stops training; then, the diagnostic model is output.

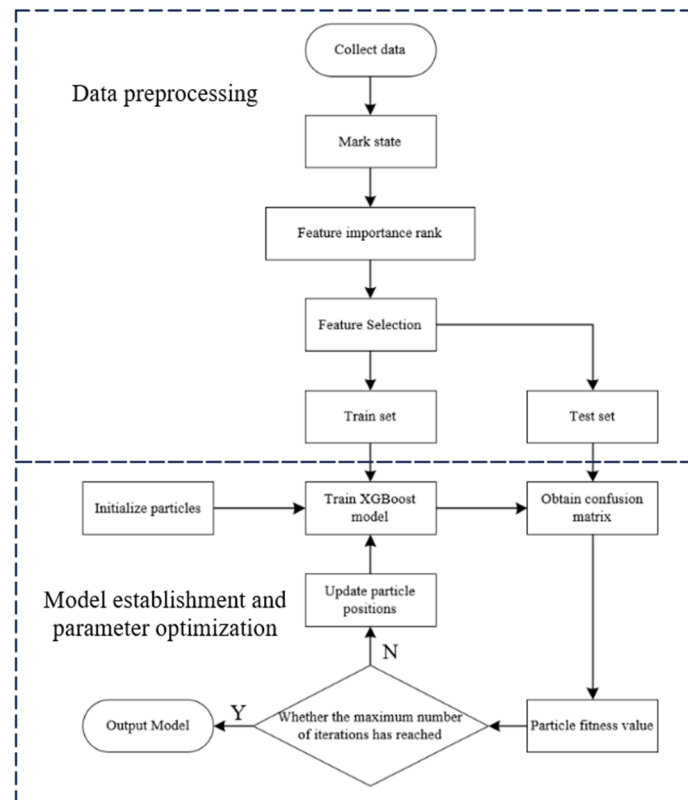


Figure 2. Model training flowchart.

3. Examples of Fault Diagnosis for Chiller Units

The minimum classification loss function is chosen as the penalty function for the diagnostic model. This model is based on the Python 3.6.15 platform and references the Numpy 1.19.5, Scikitlearn 0.17.1, Pandas 1.2.4, Matplotlib 3.4.3, and Seaborn 0.12.2 [27] scientific computing packages.

3.1. Fault Sample Information

The dataset used in this article comes from the ASHRAE RP-1043 chiller fault simulation experiment [28], and the system sketch is shown in Figure 3.

In Figure 3, a centrifugal chiller with a capacity of 90 refrigeration tons can be used to simulate the cooling effect of the cooling tower by exchanging heat between tap water and cooling water. Steam and hot water are used to simulate user load, and 64 parameters are recorded, including 48 measured parameters and 16 calculated parameters such as compressor efficiency and heat exchanger efficiency. The parameters are numbered 0–7 in the order shown in Table 3 according to reference [28]. The experiment simulated four types of local faults and three types of system faults and simulated four degradation levels from small to large for these seven faults to obtain operational data. In order to explore the diagnostic ability of minor faults in chillers, seven types of minimum simulated monitoring data with different fault levels were selected as the research samples for this article. The

fault description is detailed in Table 3, and the percentage after the unit status description text in the table represents the degree of change compared to the normal state.

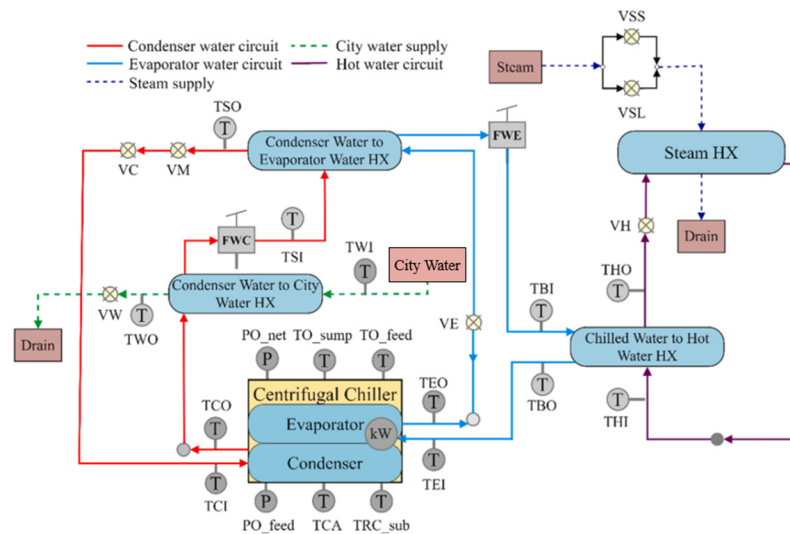


Figure 3. RP-1043chiller system fault simulation experimental device schematic diagram.

Table 3. RP-1043 fault information.

Type of Fault	Unit Status	Abbreviation	Label
—	Normal state	Normal	0
partial fault	Reduced condenser water flow (−10%)	FWC	1
	Reduced evaporator water flow (−10%)	FWE	2
	Condenser fouling (12%)	CF	3
	Non-condensable gas in refrigerant (1%)	NC	4
system fault	Refrigerant leak (−10%)	RL	5
	Refrigerant overcharge (+10%)	RO	6
	Excess oil (+14%)	EO	7

The experimental period for each state is up to 14 h, and the data collection interval is 10 s. Due to the drastic fluctuations in some parameters during the brief moments of starting and stopping the chiller unit, the data validity is not high. Therefore, this part of the data is removed when selecting the dataset. In the remaining data, 4000 samples are uniformly extracted for each of the eight states mentioned in Table 3. This forms a training sample set of $32,000 \times 64$ dimensions, ensuring a balanced number of samples among different classes.

3.2. Data Preprocessing

Among the 64 parameters, the calculated parameters are all obtained from the measured parameters, and the parameters numbered 0–8 are repeated measurements of temperature parameters related to the condenser and evaporator using a Resistance Temperature Detector (RTD) and thermistors. In addition, the complex nonlinear relationship between the parameters of the refrigeration system results in a highly coupled situation among the 64 parameters. The contribution degree is defined as the ratio of the splitting times of a feature participating in the decision tree to the total splitting times of the model. All features are ranked in order to lower the input feature dimension of the training set. Table 4 shows that the cumulative contribution is the sum of the current feature contribution and all feature contributions before the current feature ranking. The accuracy is the overall accuracy of the XGBoost model classification under the current cumulative contribution, and then the optimal feature dimension is determined together with the training time.

Table 4. Original data feature parameter contribution analysis.

Number	Contribution Degree	Ranking	Accumulated Contribution	Accuracy	Training Time/s
56	18.42%	1	18.42%	26.87%	6.28
25	14.11%	2	32.53%	39.76%	7.56
24	10.64%	3	43.17%	48.04%	9.14
47	8.66%	4	51.83%	73.17%	12.03
4	5.31%	5	57.14%	79.28%	13.12
45	4.08%	6	61.22%	83.67%	14.99
58	3.74%	7	64.96%	85.71%	16.17
27	3.14%	8	68.10%	87.19%	16.81
48	2.66%	9	70.76%	88.46%	17.93
28	2.42%	10	73.18%	89.35%	20.20
39	2.31%	11	75.49%	90.02%	24.43
33	2.11%	12	77.60%	90.12%	26.51
3	1.86%	13	76.69%	90.39%	26.99
20	1.73%	14	79.46%	90.81%	28.80
30	1.62%	15	79.97%	91.18%	30.42
49	1.51%	16	81.08%	91.61%	32.51
9	1.39%	17	82.47%	91.82%	33.69
7	1.31%	18	83.78%	92.01%	33.99
46	1.26%	19	85.04%	92.43%	34.48
36	1.23%	20	86.27%	92.54%	35.34
11	1.18%	21	87.45%	92.89%	38.86
16	1.09%	22	88.54%	93.15%	39.21
18	1.00%	23	89.54%	93.35%	42.75
52	0.96%	24	90.50%	93.23%	43.93
32	0.87%	25	91.37%	92.11%	45.41
...					
63	0%	64	100%	93.51%	79.16

Table 4 shows that as the number of features increases, the model training time continues to increase, with accuracy first increasing and then decreasing. The cumulative contribution reaches 89.54% with the feature number at 23, and the XGBoost model has the best accuracy rate of 93.35%. Considering that the original feature parameters, such as pressure and temperature, belong to slowly changing process parameters, and the training time is 41.75 s, which meets the practical requirements of engineering applications, the top 23 features ranked in contribution are selected to form a new low-dimensional learning sample.

The set of training samples after dimensionality reduction is a matrix of $32,000 \times 23$. Based on the size of the sample and the perspective of statistical learning, cross-validation is adopted and the samples are randomly and evenly divided into a training set of $24,000 \times 23$ and a test set of 8000×23 in a 3:1 ratio. First, the training set data were input into the given algorithm to train the fault diagnosis model, then the test set data were input into the trained diagnostic model and the confusion matrix and other evaluation indicators were used to evaluate the performance of fault diagnosis models.

3.3. XGBoost Parameter Optimization

Table 5 shows the optimization results of XGBoost parameters using MSPSO and PSO, respectively. Figure 4 shows the accuracy curves of XGBoost optimized via both methods. As shown in Figure 4, the PSO algorithm has a faster convergence speed in the early stages but it falls into a local optimum at 20 iterations and is unable to escape. The classification accuracy of the XGBoost algorithm optimized using PSO increased from 93.35% to 97.71% after 50 iterations. The MSPSO algorithm has a slow convergence speed in the early stages but particles search for the optimal solution through multi-group cooperation, which can effectively prevent the local optima [29]. The accuracy of the XGBoost optimized using MSPSO can be considerably enhanced when compared to the PSO-optimized version. The

classification accuracy of the MSPSO-optimized XGBoost algorithm is now 99.67%, up from 93.35% previously.

Table 5. Model parameter optimization results.

Parameter	Default	PSO	MSPSO
Learning rate (η)	0.3	0.12	0.033
Gain threshold (γ)	0	0.11	0.02
The maximum depth of the tree	6	4	3
Minimum weight of leaf nodes	1	2	3
Number of sub-model trees	100	345	203

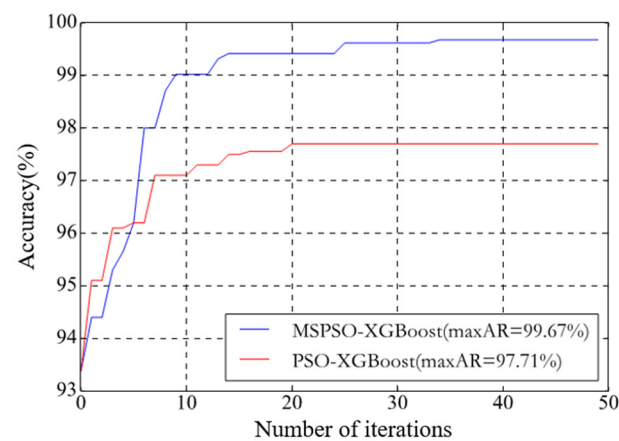


Figure 4. MSPSO and PSO optimize the XGBoost accuracy curve.

3.4. Diagnostic Results and Analysis

Figure 5 illustrates the diagnostic accuracy of fault models that were trained using the same dataset with the SVM, back propagation neural network (BPNN), XGBoost, PSO-XGBoost, and MSPSO-XGBoost algorithms for eight distinct states. The SVM and the BPNN parameter sets are shown, and the parameters are listed in Tables 6 and 7. The AR in Figure 5 represents the overall diagnostic accuracy of the five models.

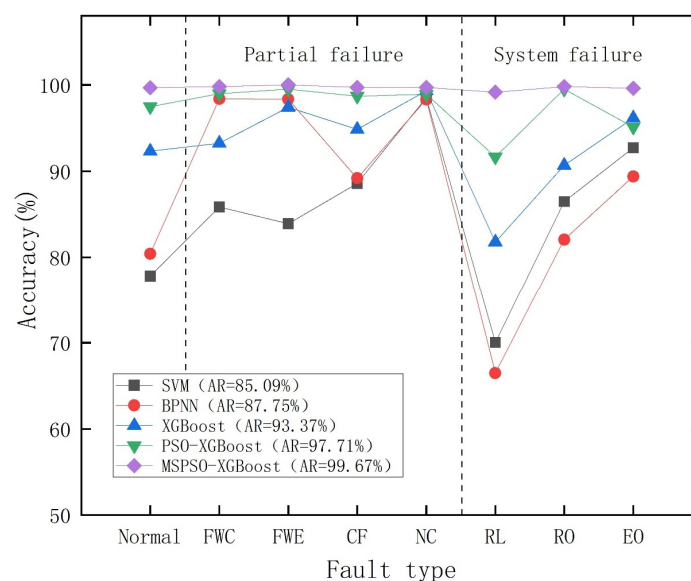


Figure 5. The precision rate of the diagnosis of 8 states using 5 algorithms.

Table 6. SVM parameter values.

Kernel Function Type	Kernel Function Parameters (γ)	Penalty Factor (C)
Gaussian kernel	0.6	32

Table 7. BPNN parameter values.

Hyper-Parameters	Value
Number of input layer neurons	23
Number of hidden layers	3
Number of neurons per hidden layer	12
The activation function of the hidden layer	sigmoid
Number of output layer neurons	8

From Figure 5, some results can be concluded.

1. The performance of the local fault diagnosis is superior to that of system fault diagnosis, indicating that the difficulty of diagnosing local faults is lower than that of system faults in the minor faults of chillers. The reason for this phenomenon is that local faults can usually be judged with a small number or even a single parameter change, while system faults often require more parameter changes to be judged comprehensively. From the perspective of engineering practice, for example, FWC failure is usually based on the chilled water flow and cooling water flow changes can be relatively accurate judgements, while the RL failure needs to be combined with the degree of subcooling, chilled water flow, cooling water flow, compressor operating power, the condenser temperature, the evaporator temperature, condensing pressure, oil supply temperature, oil pressure temperature, and oil supply pressure parameter changes to make a comprehensive diagnosis decision;
2. According to the literature [30,31], it can be concluded that BPNNs are superior to SVMs in the diagnosis of local faults, but inferior to SVMs in the diagnosis of system faults. Except for slightly lower diagnostic accuracy in FWC and FWE faults compared to BPNNs, XGBoost outperforms BPNNs and SVMs in diagnostic accuracy in other faults. In terms of overall accuracy ratio (AR), XGBoost is better than BPNNs and SVMs for chiller micro-fault diagnosis;
3. Compared with XGBoost, the PSO-XGBoost has improved diagnosis accuracy of all faults except EO; the accuracy of MSPSO-XGBoost diagnosis in eight states is better than XGBoost. The diagnosis performance of the MSPSO-XGBoost is prior to PSO-XGBoost in three types of system faults and normal states, demonstrating that MSPSO-XGBoost is more capable of diagnosing micro-faults in chillers than PSO-XGBoost.
4. Compared to SVMs and BPNNs, the AR of MSPSO-XGBoost is improved by 14.58% and 11.92%, respectively. The gap in diagnostic performance for various types of faults is narrowed by XGBoost after parameter optimization.

Tables 8–12 show the confusion matrices of SVMs (Gaussian kernel), BPNNs, XGBoost, PSO-XGBoost, and MSPSO-XGBoost. The confusion matrix is partitioned into nine parts with four black dashed lines according to the three levels of normal, local, and systematic faults, which is known as the nine-grid. In this case, the middle part of the nine-grid (4×4 matrix) represents the classification results of local faults, while the bottom-right corner (3×3 matrix) represents the classification results of systematic faults.

Table 8. SVM (Gaussian kernel) confusion matrix.

	Number	True Faults								Precision Rate	Recall Rate
		0	1	2	3	4	5	6	7		
Diagnosed Faults	0 Normal	752	42	57	0	0	64	6	45	77.84%	78.08%
	1 FWC	39	850	12	2	1	46	23	17	85.85%	85.00%
	2 FWE	89	11	808	0	1	21	11	22	83.90%	87.06%
	3 CF	7	4	5	875	0	54	22	21	88.56%	92.59%
	4 NC	0	3	3	3	1002	0	6	0	98.52%	99.30%
	5 RL	46	71	28	39	0	746	88	47	70.04%	71.93%
	6 RO	6	13	9	25	5	71	883	9	86.48%	84.98%
	7 EO	24	6	6	1	0	35	0	918	92.72%	85.07%

Table 9. BPNN confusion matrix.

	Number	True Faults								Precision Rate	Recall Rate
		0	1	2	3	4	5	6	7		
Diagnosed Faults	0 Normal	777	2	37	0	0	105	5	40	80.43%	79.12%
	1 FWC	2	974	1	2	5	3	3	0	98.38%	98.58%
	2 FWE	10	0	947	0	1	2	0	3	98.33%	94.22%
	3 CF	3	0	0	881	5	57	29	13	89.17%	86.54%
	4 NC	0	2	0	5	1000	0	9	1	98.32%	98.23%
	5 RL	151	4	10	66	0	708	89	37	66.47%	69.54%
	6 RO	13	6	2	54	7	88	838	13	82.07%	85.67%
	7 EO	26	0	8	10	0	55	6	885	89.39%	89.21%

Table 10. XGBoost confusion matrix.

	Number	True Faults								Precision Rate	Recall Rate
		0	1	2	3	4	5	6	7		
Diagnosed Faults	0 Normal	892	22	0	0	2	28	17	5	92.33%	88.84%
	1 FWC	47	923	0	0	2	7	9	2	93.23%	96.44%
	2 FWE	13	5	938	0	0	2	1	4	97.40%	99.57%
	3 CF	0	0	2	937	1	30	15	3	94.83%	99.57%
	4 NC	0	0	2	2	1010	0	1	2	99.31%	99.51%
	5 RL	29	4	0	2	0	871	126	33	81.78%	86.66%
	6 RO	8	2	0	0	0	58	927	26	90.69%	83.49%
	7 EO	15	1	0	0	0	9	13	952	96.16%	92.69%

Table 11. PSO-XGBoost confusion matrix.

	Number	True Faults								Precision Rate	Recall Rate
		0	1	2	3	4	5	6	7		
Diagnosed Faults	0 Normal	937	2	0	4	0	21	1	1	96.99%	99.26%
	1 FWC	0	979	0	0	9	0	1	1	98.89%	99.69%
	2 FWE	0	0	958	2	0	0	0	3	99.48%	100.00%
	3 CF	0	1	0	975	12	0	0	0	98.68%	99.19%
	4 NC	1	0	0	2	1004	0	10	0	98.72%	97.95%
	5 RL	4	0	0	0	0	976	80	5	91.64%	97.80%
	6 RO	1	0	0	0	0	0	1015	5	99.41%	88.26%
	7 EO	1	0	0	0	0	1	43	945	95.45%	97.72%

Table 12. MSPSO-XGBoost confusion matrix.

	Number	True Faults								Precision Rate	Recall Rate
		0	1	2	3	4	5	6	7		
Diagnosed Faults	0 Normal	963	0	0	1	0	1	0	1	99.68%	99.27%
	1 FWC	0	988	0	0	0	0	1	1	99.79%	99.89%
	2 FWE	0	0	963	0	0	0	0	0	100.00%	100.00%
	3 CF	0	1	0	985	2	0	0	0	99.69%	99.69%
	4 NC	1	0	0	2	1013	0	1	0	99.70%	99.80%
	5 RL	4	0	0	0	0	1056	1	4	99.15%	99.81%
	6 RO	1	0	0	0	0	0	1019	1	99.80%	99.51%
	7 EO	1	0	0	0	0	1	2	986	99.59%	99.29%

From Table 8, it can be found that the diagnostic accuracy of NC is significantly better than the other faults in the table, while the diagnostic accuracies of NC faults in Tables 8–12 are quite close to each other. From the perspective of refrigeration principles, the variables that will be affected first are analyzed when an NC fault occurs. It is found that they coincide with the four variables listed in the top five contributions in Table 4. The raw data feature parameter numbers, namely 56, 25, 24, and 4, are the VE, FWE, FWC, and TWEO variables in the RP-1043 dataset, respectively. This phenomenon indicates that the key features have a greater impact on the fault diagnosis performance compared to the diagnostic model. It is also found that for NC faults XGBoost, PSO-XGBoost, and MSPSO-XGBoost diagnostic performance gradually and slightly decreases. It can be concluded that the overall diagnostic performance index of the model will be improved by optimizing the model parameters, but it may not necessarily be improved for certain types of faults.

As can be seen from Table 10, the XGBoost diagnostic model is prone to confuse normal with the three system-level faults of RL, RO, and EO due to the fact that in the early stage of the occurrence of the tiny faults, the monitoring data of the system faults are coupled with the monitoring data of the normal state, which results in the difficulty of the diagnosis and the existence of a high rate of misclassification. From the comparison between Tables 8 and 9, it is obvious that in the BPNN and SVM diagnostic models, there is a coupling between the normal state and system faults, as well as a certain coupling with the local faults of FWC and FWE. Table 10 shows that the XGBoost diagnostic easily confuses normal with the three system-level faults of RL, RO, and EO. The reason is that in the early stages of micro-faults, the monitoring data of system faults are coupled with the monitoring data of normal states, resulting in high diagnostic difficulty and a high misclassification rate. It can be concluded that the normal state is coupled with other faults, possibly due to the fact that experiments on fault-free data are conducted intermittently between various fault experiments. Various faults can cause changes in the physical properties of the chiller unit, resulting in a constantly changing physical state of the fault-free state, which is coupled with other fault states. Therefore, the diagnosis of minor faults at the system level of chillers is difficult.

In analyzing Tables 8–12, it was found that compared to SVMs and BPNNs, XGBoost has the highest overall correctness for local fault diagnosis, with improved recall and precision for each system fault diagnosis. PSO-XGBoost has shown considerable improvement in the diagnosis of most fault categories, but it still needs to be improved in some system faults, like RL and EO. The recall and accuracy of FWE for local faults under the MSPSO-XGBoost composite model diagnosis are 100%, and the classification performance of other local faults is likewise optimal. For system failures, the recall and accuracy rates of the three types of failures have been improved, with the accuracy rate of RL increasing to 99.81% and the recall rate of RO increasing to 99.51%. Therefore, MSPSO-XGBoost can significantly enhance the accuracy of minor fault diagnosis in chillers and facilitate timely detection.

Table 13 shows the comparison of the research of this work and the references; it can be concluded that the MSPSO-XGBoost method has a significant improvement in diagnosis, compared to the SVM and BPNN models.

Table 13. Comparative analysis of present work with other references.

Reference	Classifier	Sample Size	Ratio of Training Set to Test Set	0 Normal	1 FWC	2 FWE	3 CF	4 NC	5 RL	6 RO	7 EO	Accuracy (Minor Grade)
[32]	SVM	41,528	4:1	0.924	0.998	0.998	0.994	0.947	0.949	0.922	0.929	95.8%
[33]	BPNN	12,000	2:1	0.9243	0.9634	0.9981	0.9899	0.9922	0.9153	0.9482	0.9025	95.50%
Present Work	XGBoost	32,000	3:1	0.9233	0.9233	0.9740	0.9483	0.9931	0.8178	0.9069	0.9616	93.37%
Present Work	MPSO-XGBoost	32,000	3:1	0.9968	0.9979	1.000	0.9969	0.9970	0.9915	0.9980	0.9959	99.67%

4. Conclusions

A minor fault diagnosis model for centrifugal chillers based on the XGBoost algorithm was proposed in this paper. Five important parameters of the XGBoost algorithm were optimized using MSPSO. The minor fault diagnosis model (MSPSO-XGBoost) was established. By comparing and analyzing the diagnostic results of the XGBoost, PSO-XGBoost, MSPSO-XGBoost, BPNN, and SVM models, the following conclusions are drawn:

1. The cumulative contribution reaches 89.54% with the feature number at 23. The XGBoost model has the best accuracy rate of 93.35%;
2. The MSPSO algorithm is a good choice for optimizing XGBoost parameters. Compared to PSO, it may effectively prevent the local optimal solution, while MSPSO has a slower initial convergence speed;
3. For chillers, the high similarity of the data between minor faults and the normal state leads to distinguishing difficulty. Therefore, system-level minor faults are more difficult to distinguish compared to local minor faults;
4. The diagnostic ability of the proposed MSPSO-XGBoost model on the eight different states is uniform. The diagnostic accuracy of the model reaches 99.67%. The classification performance of MSPSO-XGBoost is superior to the SVM and BPNN diagnostic models;
5. Critical features have a greater impact on fault diagnosis performance compared to the diagnostic model. Optimizing the model parameters will improve the overall diagnostic performance metrics of the model, but it may not necessarily improve for certain types of faults.

Author Contributions: Conceptualization, Q.D.; methodology, Y.L.; software, Y.X.; validation, Y.X.; formal analysis, N.J. and Q.D.; investigation, T.L., M.Z. and N.J.; resources, M.Z.; data curation, T.L.; writing—original draft preparation, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Yaxiang Liu and Tao Liang were employed by the company Shandong Electric Power Engineering Consulting Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Dou, H.; Zmeureanu, R. Detection and diagnosis of multiple-dependent faults (MDFDD) of water-cooled centrifugal chillers using grey-box model-based method. *Energies* **2022**, *16*, 210. [\[CrossRef\]](#)
- Singh, V.; Mathur, J.; Bhatia, A. A comprehensive review: Fault detection, diagnostics, prognostics, and fault modeling in HVAC systems. *Int. J. Refrig.* **2022**, *144*, 283–295. [\[CrossRef\]](#)
- Zhao, Y.; Zhang, C.; Zhang, Y.; Wang, Z.; Li, J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy Built Environ.* **2020**, *1*, 149–164. [\[CrossRef\]](#)
- Wang, Z.; Dong, Y.; Liu, W.; Ma, Z. A novel fault diagnosis approach for chillers based on 1-D convolutional neural network and gated recurrent unit. *Sensors* **2020**, *20*, 2458. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, G.; Yao, Q.; Fan, C.; Zhou, C.; Wu, G.; Zhou, Z.; Fang, X. An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems. *Build. Environ.* **2021**, *203*, 108057. [\[CrossRef\]](#)
- Thango, B.A. On the Application of Artificial Neural Network for Classification of Incipient Faults in Dissolved Gas Analysis of Power Transformers. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 839–851. [\[CrossRef\]](#)
- Gao, T.; Yang, J.; Jiang, S. A novel incipient fault diagnosis method for analog circuits based on GMKL-SVM and wavelet fusion features. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 3502315. [\[CrossRef\]](#)
- Wang, Z.; Guo, J.; Zhou, S.; Xia, P. Performance Evaluation of Chiller Fault Detection and Diagnosis Using Only Field-Installed Sensors. *Processes* **2023**, *11*, 3299. [\[CrossRef\]](#)
- Fan, Y.; Cui, X.; Han, H.; Lu, H. Feasibility and improvement of fault detection and diagnosis based on factory-installed sensors for chillers. *Appl. Therm. Eng.* **2020**, *164*, 114506. [\[CrossRef\]](#)
- Han, H.; Cui, X.; Fan, Y.; Qing, H. Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features. *Appl. Therm. Eng.* **2019**, *154*, 540–547. [\[CrossRef\]](#)
- Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gong, W.; Chen, H.; Zhang, Z.; Zhang, M.; Wang, R.; Guan, C.; Wang, Q. A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion. *Sensors* **2019**, *19*, 1693. [\[CrossRef\]](#)
- Kukreja, H.; Bharath, N.; Siddesh, C.S.; Kuldeep, S. An introduction to artificial neural network. *Int. J. Adv. Res. Innov. Ideas Educ.* **2016**, *1*, 27–30. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mienye, I.D.; Sun, Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access* **2022**, *10*, 99129–99149. [\[CrossRef\]](#)
- Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
- Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520. [\[CrossRef\]](#)
- Wang, C.W.; Lee, Y.C.; Calista, E.; Zhou, F.; Zhu, H.; Suzuki, R.; Cheng, S.P. A benchmark for comparing precision medicine methods in thyroid cancer diagnosis using tissue microarrays. *Bioinformatics* **2018**, *34*, 1767–1773. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, W.; Lesner, C.; Ran, A.; Rukonic, M.; Xue, J.; Shiu, E. Using small business banking data for explainable credit risk scoring. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13396–13401. [\[CrossRef\]](#)
- Wassan, J.T.; Wang, H.; Browne, F.; Zheng, H. A comprehensive study on predicting functional role of metagenomes using machine learning methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *16*, 751–763. [\[CrossRef\]](#) [\[PubMed\]](#)
- Qiu, C.; Zhang, L.; Li, M.; Zhang, P.; Zheng, X. Elevator Fault Diagnosis Method Based on IAO-XGBoost under Unbalanced Samples. *Appl. Sci.* **2023**, *13*, 10968. [\[CrossRef\]](#)
- Zhang, S.; Zhu, X.; Anduv, B.; Jin, X.; Du, Z. Fault detection and diagnosis for the screw chillers using multi-region XGBoost model. *Sci. Technol. Built Environ.* **2021**, *27*, 608–623. [\[CrossRef\]](#)
- Wang, Y.; Ni, X.S. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv* **2019**, arXiv:1901.08433. [\[CrossRef\]](#)
- Kaligambe, A.; Fujita, G.; Tagami, K. Indoor Room Temperature and Relative Humidity Estimation in a Commercial Building Using the XGBoost Machine Learning Algorithm. In Proceedings of the 2022 IEEE PES/IAS PowerAfrica, Kigali, Rwanda, 22–26 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5. [\[CrossRef\]](#)
- Jain, M.; Saihpal, V.; Singh, N.; Singh, S.B. An overview of variants and advancements of PSO algorithm. *Appl. Sci.* **2022**, *12*, 8392. [\[CrossRef\]](#)
- Gad, A.G. Particle swarm optimization algorithm and its applications: A systematic review. *Arch. Comput. Methods Eng.* **2022**, *29*, 2531–2561. [\[CrossRef\]](#)
- Yang, X.; Li, H.; Yu, X. A dynamic multi-swarm cooperation particle swarm optimization with dimension mutation for complex optimization problem. *International J. Mach. Learn. Cybern.* **2022**, *13*, 2581–2608. [\[CrossRef\]](#)
- Hao, J.; Ho, T.K. Machine learning made easy: A review of scikit-learn package in python programming language. *J. Educ. Behav. Stat.* **2019**, *44*, 348–361. [\[CrossRef\]](#)
- Comstock, M.C.; Braun, J.E. Development of analysis tools for the evaluation of fault detection and diagnostics for chillers. In ASHRAE Research Project 1043-RP, HL 99-20, Report #4036-3; Purdue University: West Lafayette, IN, USA, 1999.

29. Gu, K.; Wang, J.; Qian, H.; Su, X. Study on intelligent diagnosis of rotor fault causes with the PSO-XGBoost algorithm. *Math. Probl. Eng.* **2021**, *2021*, 9963146. [[CrossRef](#)]
30. Ren, Z.; Han, H.; Cui, X.; Qing, H.; Ye, H. Application of PSO-LSSVM and hybrid programming to fault diagnosis of refrigeration systems. *Sci. Technol. Built Environ.* **2021**, *27*, 592–607. [[CrossRef](#)]
31. Sun, S.; Li, G.; Chen, H.; Huang, Q.; Shi, S.; Hu, W. A hybrid ICA-BPNN-based FDD strategy for refrigerant charge faults in variable refrigerant flow system. *Appl. Therm. Eng.* **2017**, *127*, 718–728. [[CrossRef](#)]
32. Sun, X.; Yan, K.; Zhou, X. Fault detection and diagnosis of chillers with s&d convolutional neural network. In Proceedings of the 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Rhodes, Greece, 2–6 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 829–836. [[CrossRef](#)]
33. Liang, Q.; Han, H.; Cui, X.; Qing, H.; Fan, Y. Comparative study of probabilistic neural network and back propagation network for fault diagnosis of refrigeration systems. *Sci. Technol. Built Environ.* **2018**, *24*, 448–457. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.