Study on Chiller Fault Detection and Diagnosis Method Based on KNN Algorithm and ANOVA

Le Minh Nhut* and Le Ha Dong Quan

Department of Thermal Engineering, Faculty of Vehicle and Energy Engineering, Ho Chi Minh City University of Technology and Education, Thu Duc City, Ho Chi Minh City, Vietnam; Email: lehadongquan@gmail.com (L.H.D.Q.)
*Correspondence: nhutlm@hcmute.edu.vn (L.M.N.)

***Abstract*— As the economy, population, and industry have grown in recent years, more and more water chiller systems have been installed in many buildings throughout the world. However, faults can appear during operation, leading to a reduction in the life of a system and increased energy consumption. As a result, it is necessary to identify and overcome these faults. This paper proposes a chiller fault detection and diagnosis (FDD) method based on the K-nearest neighbors (KNN) algorithm and an analysis of variance (ANOVA) to reduce the number of sensors in a real system and to improve the performance of chiller FDD. A Python program based on the KNN and ANOVA models was developed to simulate and validate the chiller fault detection and diagnosis. The results showed that the correct rates (CRs) of stages 1 and 2 in Case 1 were 99.53% and 99.60%, respectively, whereas the CRs of stages 1 and 2 in Case 2 were 99.08% and 99.48%, respectively. The highest performance of the proposed chiller FDD method was achieved when compared to the CBA method, the EBD-DBN method, and the GDW-SVDD method for Case 2 with slight-severity levels 1 and 2. Furthermore, this method was validated using real data under normal operating conditions and the condenser fouling fault of a centrifugal water-cooled chiller from the Saigon Center building in Vietnam. The results showed that the overall performance of chiller FDD was 97.61%, and the hit rate of the condenser fouling fault was 93.46%. This demonstrated that chiller FDD based on KNN and ANOVA has high reliability and can be used in industry.**

***Keywords*—Heating, Ventilation, and Air-conditioning (HVAC), faults, diagnosis, water chiller, K-nearest Neighbors (KNN) algorithm, Analysis of Variance (ANOVA)**

I. Introduction

Currently, heating, ventilation, and air-conditioning (HVAC) systems are rapidly increasing due to development in the population, tourism, and industries. They consume a large amount of energy, which forms a major part of the total energy used in commercial buildings, hotels, resorts, and industrial factories. HVAC systems consume a large amount of energy: 35–40% in Hong Kong [1]; 50% in the United States [2]; and up to about 60% in countries with high temperature and humidity, such as Singapore [3]. Producing energy impacts the environment negatively. Therefore, saving energy while pushing the economy forward is a long-term aim and has become an interesting topic. A chiller is a complex system, has a high cost, and is the highest energy-consuming component in an HVAC system. Operating a chiller with a fault wastes energy and raises maintenance costs significantly [4]. Successful Fault Detection and Diagnosis (FDD) can save 10–40% of HVAC energy consumption [5]. Moreover, in many industrial fields, chiller faults can cause an entire process to stop [6]. Therefore, establishing chiller FDD on time is necessary, as it maintains the continuous operation of a system and reduces the number of severely damaged parts.

FDD models have been developed in recent decades and can be divided into three main groups [7]: quantitative, qualitative, and data-based methods. Of these, data-driven methods are becoming more and more popular. Han *et al.* [8] used an LS-SVM method to detect and diagnose seven faults in a chiller system. Li *et al.* [9] used a PCA-R-SVDD method with eight temperature measurements. Han et al. [10] proposed a strategy using a deep neural network with simulated annealing. Wang *et al.* [11] proposed an entropy-based discretization-discrete Bayesian network (EBD-DBN) method. Both fault diagnosis and fault action mechanisms based on a classification-based association (CBA) were given by Liu *et al.* [12]. The best performance was 96.53% for a non-condensable fault, and the worst performance was 81.17% for refrigerant overcharge. In addition, this work also provided low results for excess oil at 62.96%, refrigerant leak at 76.74%, and condenser fouling at 76.47% for a slight severity level. Chen *et al.* [13] developed a Global Density-Weighted Support Vector Data Description (GDW-SVDD) method to improve detection accuracy and to reduce the false alarm rate (FAR). They concluded that the FAR reduced from 10.75% to 8.25%, and the fault detection accuracy improved by 3.75%. Huang *et al.* [14] proposed a model to achieve an absolute performance for condenser fouling, but the refrigerant overcharge and refrigerant leak performances were only 67.4% and 80.8%, respectively. Suowei *et al.* [15] developed a Bayesian network classifier with a Probabilistic Boundary (PB-BNC) model and added site information to a Bayesian network classifier with a probabilistic boundary (SI-PB-BNC) model for FDD. The results showed that both models diagnosed a non-

Manuscript received October 21, 2022; revised December 22, 2022; accepted February 13, 2023.

condensable fault with 100% accuracy. However, the PB-BNC model had bad diagnostic results for refrigerant leaks, at 69.4%, and condenser fouling, at 73%. The SI-PB-PBC model was similar, with condenser fouling at 73.6% accuracy. In practice, chiller operation is a complex process and may produce faults. Therefore, a model should have good diagnostic ability, and uniform performances between faults that are reliable are expected. Yan *et al.*'s [16] research provided low performances for condenser-fouling faults: 62.07% accuracy at level 1 and 79.31% at level 2. The model by Xia *et al.* [17] performed at a 69% accuracy for evaporator water flow and at a 68% accuracy for refrigerant leak. The early detection and diagnosis of slightly severe faults is important and necessary. It supports scheduled maintenance services, optimal operation, and energy conservation. For these reasons, a solution is required to overcome these drawbacks. Gao *et al.* [18] analyzed the fault characterization features of a chiller based on a Global Sensitivity Analysis (GSA) and a Cascade Feature-Cleaning and Supplement (CFCS) model to reduce redundancy among sensitive features and to supplement further information for performance promotion. Wang *et al.* [19] developed an FD model by combining a Bayesian Network (BN) and a Principal Component Analysis (PCA). In the developed model, the accuracy was increased the most, by 43%, for condenser fouling at level 1. In addition, Wang *et al.* [20] evaluated the FD accuracies of building energy systems under missing univariate data and missing multivariate data based on an expectation-maximization algorithm and a Bayesian network (EM-BN). The configuration of Factory-Installed (FI) sensors was investigated to improve the fault diagnosis performance of the model [21].

From the aforementioned literature review, although much research has been conducted on chiller Fault Detection and Diagnosis (FDD) models, most of the previous studies have employed too many physical variables, increasing the number of sensors and the amount of information overlap, both of which increase computational cost and reduce the performance of chiller FDD. Furthermore, early detection and diagnosis of slightly severe faults are important because they help in maintenance planning, operational optimization, and energy conservation. The above studies, however, have had low diagnostic performances for slightly severe faults. For the above reason, this paper proposes a chiller fault detection and diagnosis method based on the K-nearest neighbors algorithm and an analysis of variance (FDD-KNN-ANOVA) to reduce the physical variables for improving chiller FDD performance, which can then be applied to the industry.

II. Research Methodology

A. KNN Algorithm for Classification Model

As a classification method, KNN is considered to play a key role in pattern classification [22] and was selected as one of the top 10 data-mining algorithms [23]. The basic idea of KNN is that, if a great majority of neighbors belong to a class, then the data needing classification belong to the same class. The training dataset D contains training samples x_i. Each data sample is described by variables $f_i \in F'$ and labeled with a class label $y_i \in Y$. For each $x_i \in D$, the KNN algorithm calculates the distance between x and x_i. Depending on whether the variable is continuous or discrete, the distance is calculated as follows [24]:

$$d(x_f, x_{if}) = \begin{cases} 0 \text{ if f discrete and } x_f = x_{if} \\ 1 \text{ if f discrete and } x_f \neq x_{if} \\ \left| x_f - x_{if} \right| \text{ if f continuous} \end{cases} \quad (1)$$

In this study, the variables are continuous, so the Euclidean distance is employed. When the K nearest neighbors of the training samples are identified, the label of x is determined by the voting method. There are two voting methods that are used as optimization parameters. These are presented in the following equations.

Majority voting [25]:

$$y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} \delta(v, y_i) \quad (2)$$

Distance-weighted voting [25]:

$$y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} w_i \delta(v, y_i) \quad (3)$$

where y′ is the predicted label for test point data; D_z is the dataset of K-nearest neighbors of the test sample; and x_i and y_i denote the data and the class label in D_z, respectively. In this paper, cross-validation was used to choose the K-value, and w_i is the weight distance of x_i.
In this paper, the weight from the new point x to the point x_i was expressed as follows [26, 27]:

$$w_i = \exp\left(\frac{-\left\| x - x_i \right\|_2^2}{a^2} \right) \quad (4)$$

where a is an optional positive number.

B. Feature Selection Based on ANOVA

Raw data contain a large number of variables that cannot accurately reflect system information when a fault occurs. As a result, they must be removed. Reducing the variables in a chiller FDD model decreases the number of installed sensors, as well as the computational cost. Reaching a low installation cost is a key factor in the deployment of chiller FDD in field applications [28]. Many feature selection methods have been applied, such as cost-sensitive sequential feature selection [28], genetic algorithms [29], and principal component analyses [19], [30]. In this study, the ANOVA conducted is called the F–statistic, and it was used to compare the multiple mean values of the dataset and to visualize whether there existed any significant difference between the mean values of multiple groups, calculated with the following steps.

The mean of squares between groups is expressed as follows [31]:

$$MSB = \left(\sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X} \right)^2 \right) / df \quad (5)$$

The mean of squares within groups is expressed as follows [31]:

$$MSW = \left(\sum_{i=1}^{k} (n_i - 1) \sigma_i^2 \right) / df_w \quad (6)$$

where n_i, $\overline{X}_i$, and σ are the number sample, the mean, and the standard deviation of the ith group, respectively. $\overline{X}$ is the mean of the dataset, df is the degree of freedom, and $df_w = N - k$ (N is the number of samples, and k is the number of groups.

The F-statistic is determined as follows [31]:

$$F = MSB/MSW \quad (7)$$

The greater the F-value, the more reliable the results because it indicates greater differences between the sample averages. Furthermore, in the ANOVA, the F-value is used to calculate the p-value. The p-value is used to decide whether to accept or reject the null hypothesis. The lower the p-value, the more likely it is that the null hypothesis is rejected.

C. Reference Data Preprocessing and Fault Diagnosis Performance Evaluation Indices

Data preprocessing [31] is a critical process in data mining and machine learning [32]. Raw data might contain irrelevant samples that have no contribution to the performance of a diagnostic model. The reasons are damaged equipment, chiller operation containing transient processes, and varied capacity. Thus, it is necessary to perform data preprocessing on the raw data collected from an experimental chiller system.

In this paper, the steady-state detector developed by Han *et al.* [34] to select steady-state data was adopted, in which the outliers were removed by an interquartile range rule algorithm. It identified the outliers by defining a lower threshold ($Q_1 - 1.5 \times (Q_3 - Q_1)$) and an upper threshold ($Q_3 + 1.5 \times (Q_3 - Q_1)$), where Q_1 is the first quartile and Q_3 is the third quartile of each variable. When the data were higher than the upper threshold or lower than the lower threshold, they were eliminated. Furthermore, data standardization can effectively eliminate negative effects and improve the stability of the model [35]. Z-score normalization was used. For an original dataset with N samples and n characteristics, the original matrix Z $\left(Z \in R^{N \times n}\right)$ was normalized to a matrix $\widehat{X}$ by following equation [13]:

$$\begin{cases} m_j = \dfrac{1}{N} \sum_{i=1}^{N} Z_{i,j} \\ \sigma_j = \sqrt{\dfrac{1}{N-1} \sum_{i=1}^{N} \left(Z_{i,j} - M_j \right)^2} \\ x_{i,j} = \dfrac{z_{i,j} - M_j}{\sigma_j} \end{cases} \quad (8)$$

where m_j is the mean, σ_j is the standard deviation of the jth column characteristic, $z_{i,j}$ is a component of the matrix Z, and $\hat{x}_{i,j}$ is a component of the matrix $\widehat{X}$.

D. Evaluation Metrics for Classification Models

A confusion matrix provides a metric to evaluate performance. In particular, it shows which classes are misdiagnosed with other classes so that an author can improve individual efficiency. In addition, a confusion matrix, such as that shown in Table I, can give a better idea of what a diagnosis model evaluates correctly and what types of errors it makes.

TABLE I. CONFUSION MATRIX.

Abbreviation		Predicted label	
		Positive	Negative
Actual label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

TP is the number of positive samples that are correctly predicted, FN is the number of positive samples that are incorrectly predicted, TN is the number of negative samples that are correctly predicted, and FP is the number of negative samples that are incorrectly predicted. Good results correspond to large numbers down the main diagonal and small, off-diagonal elements (ideally zero). For a multi-class classification problem, such as fault diagnosis, positive can be considered as the category of interest and negative represents all the other categories.

In this study, parameters such as accuracy, recall, and F1_Score were used to evaluate the performance of the fault diagnosis model. Accuracy is the ratio of correct predictions to the total predictions made and is calculated using the following equation:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (9)$$

Precision is defined as the number of TP identifications divided by the number of predicted positives and is expressed as follows:

$$Precision = TP/(TP + FP) \quad (10)$$

Recall (sensitivity or true positive rate)explains how many of the actual positive cases are able to be predicted correctly with a model and is defined as the number of TP identifications divided by the total number of actual positives using the following equation:

$$Recall = TP/(TP + FN) \quad (11)$$

F1_Score is the harmonic mean of precision and recall and is given by the following:

$$F1\_Score = \frac{1}{precision} + \frac{1}{recall} = 2 \frac{precision \,.\, recall}{precision + recall} \quad (12)$$

The Correct Rate (CR) and Hit Rate (HR) were used to evaluate the overall and individual performances of the FDD model, respectively. They were calculated from

elements in the confusion matrix. Their definitions and calculations are given in Table II.

TABLE II. DEFINITIONS AND CALCULATIONS OF CR AND HR.

Evaluation indices	Definition	Calculation
CR	The ratio of correct predictions to the total predictions made	CR=(TP+TN)/(TP+FP+TN+FN)
HR	The correctly predicted cases that actually turned out to be positive	HR=TP/(TP+FP)

III. THE DEVELOPED CHILLER FDD METHOD BASED ON KNN MODEL AND ANOVA

A chiller FDD method based on the KNN model and ANOVA was suggested and is illustrated in Fig. 1. This model consisted of three parts, as follows:

- **Data-preprocessing Module**: the chiller's historical data were preprocessed. Physical variables were selected using the authors' extensive knowledge. After each status (normal and fault), the same samples were randomly selected and split into training, validation, and testing datasets.
- **Development Model**: the structure of the model was determined, the training dataset was used to train the model, and the validation dataset was used to find optimized parameters. In addition, an ANOVA was carried out to check the variable sensitivity and to reduce the number of variables.
- **FDD Method**: this method had two stages. In the first stage, a normal status and seven faults were detected and diagnosed. If the model delivered the correct results and its label had faults, it moved forward with stage two. In the second stage, the model diagnosed fault severity levels. The testing dataset was used to confirm and evaluate the FDD model's performance.

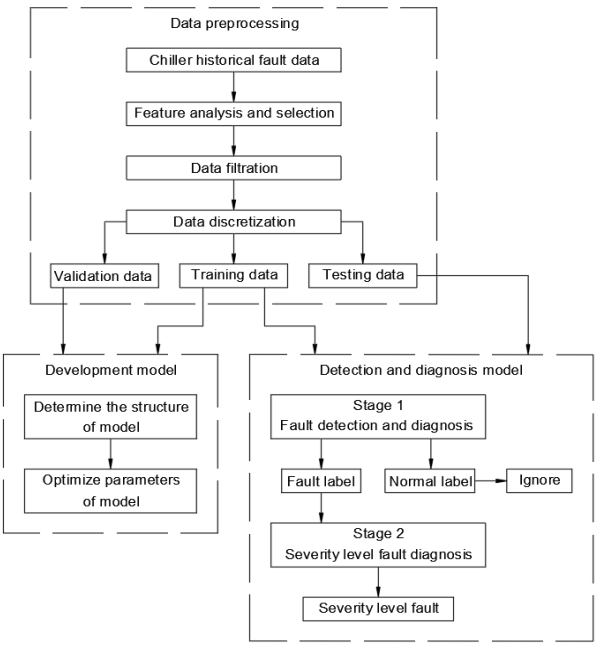

Figure 1. A flowchart of the chiller FDD method based on KNN and ANOVA.

A flowchart of the experimental system is shown in Fig. 2. The experimental data in this study were taken from ASHRAE RP-1043 [36], and a centrifugal water-cooled chiller with a capacity of 90 tons was used. Both the evaporator and the condenser were shell-and-tube designs. Water flowed inside the tubes, while R134a refrigerant flowed outside. The refrigerant was controlled by a thermostatic expansion valve. The experiments were conducted under normal operating conditions and with seven common faults: condenser fouling (ConFoul), excess oil (ExcsOil), reduced condenser water flow (ReduCF), reduced evaporator water flow (ReduEF), non-condensable refrigerant (NonCon), refrigerant leak or undercharge (RefLeak), and refrigerant overcharge (RefOver). Four severity levels of each fault were considered, as given in Table III. The data acquisition interval was 10 seconds, with 64 measurements.

TABLE III. DETAIL OF SEVERITY LEVELS FOR CHILLER FAULTS.

Type	Level 1	Level 2	Level 3	Level 4
ConFoul	12%	20%	30%	45%
ExcsOil	+14%	+32%	+50%	+68%
ReduCF	−10%	−20%	−30%	−40%
ReduEF	−10%	−20%	−30%	−40%
NonCon	+1%	+1.8%	+2.4%	+5.6%
RefLeak	−10%	−20%	−30%	−40%
RefOver	+10%	+20%	+30%	+40%

After using data-preprocessing technologies, 5,500 samples were chosen at random for each status. The results of the data distribution are depicted in Fig. 3. Then, 64%, 16%, and 20% of the data were randomly divided for the training, validation, and testing groups, respectively.

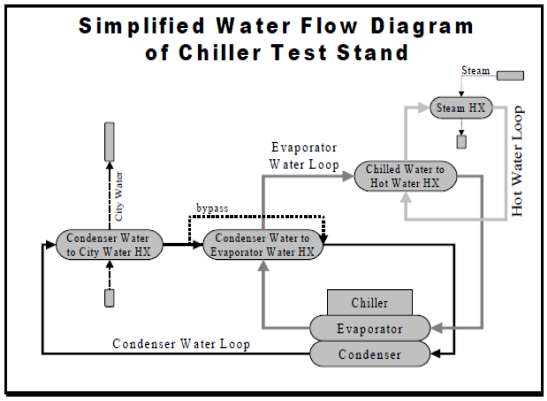

Figure 2. Experimental layout [28].

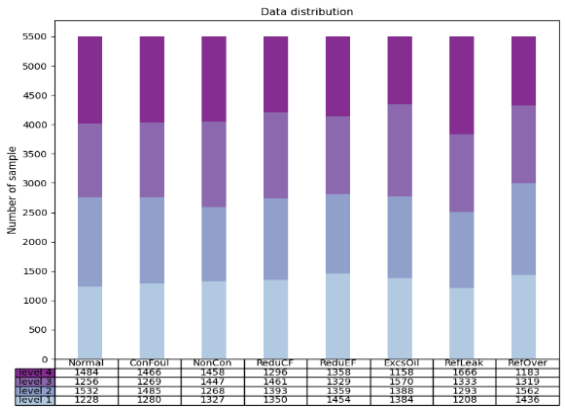

Figure 3. Distribution of the filtered dataset at four severity levels.

After considering sensor cost, popularity, and fault sensitivity, eight primary variables were chosen. Table IV displays these, and the results of the ANOVA are listed in the last two columns. TCI and TCO were expected to be sensitive variables for ReduCF because they reduce the condenser water flow rate while keeping the condenser's entering water temperature and heat rejection rate constant. Similarly, TEI and TEO were expected for to be sensitive for ReduEF. TO feed was responsive to ExcsOil since there are increased mechanical losses in compressors as oil levels rise. Non-condensable gas tends to accumulate in the condenser, which increases the saturated temperature of the refrigerant in the condenser. Consequently, TRC was expected to be sensitive for this fault. Refrigerant faults were diagnosed by the TRE and TR_dis variables. Condenser fouling was more complex. It causes an increase in heat resistance but also a decrease in condenser water flow due to reduced flow area and increased flow resistance [29]. Therefore, a difference between TRC and TCO was expected for it.

TABLE IV. SELECTED VARIABLES AND THEIR DESCRIPTIONS.

Variable	Description	F	P
TO_feed	Temperature of oil feed	3701.90	0
TRC	Saturated refrigerant temperature in condenser	435.74	0
TR_dis	Refrigerant discharge temperature	379.99	0
TCO	Temperature of condenser water outlet	130.43	7.65^{-191}
TRE	Saturated refrigerant temperature in evaporator	97.34	9.40^{-142}
TEO	Temperature of evaporator water outlet	36.42	3.46^{-51}
TCI	Temperature of condenser water inlet	26.78	6.22^{-37}
TEI	Temperature of evaporator water inlet	19.46	3.43^{-26}

A null hypothesis meant that there were no significant differences among the statuses in terms of the average variable ranges. An ANOVA was applied with an interval estimation of 95% (p-value=0.05) to decide whether the null hypothesis was rejected or accepted. The results showed that the p-values for all the variables were less than 0.05, so the null hypothesis was rejected. This meant that variables had significant differences among statuses. Therefore, the selected variables efficiently worked to detect and diagnose statuses in the chiller system. After reviewing the F-values presented in the previous section, the higher the F-value, the greater the significant difference. The TCI and TEI variables contained less information, so they were removed to save costs. Two subsets of eight variables (listed in Table IV) and six variables (TCI and TEI removed) were used in the FDD-KNN model to evaluate model performance after reducing the variables with the ANOVA. In addition, the K-values and two voting methods in the KNN algorithm were determined using the validation dataset. As a result, two case studies were used to evaluate the effectiveness of the proposed strategy. The detailed parameters are shown in Table V.

TABLE V. PARAMETER DESCRIPTIONS FOR FDD MODEL.

Case study	Number variables	K value	Voting
Case 1	8	1	Majority voting
Case 2	6	2	Distance-Weighted Voting

IV. RESULTS AND DISCUSSION

The overall performance of the chiller FDD model is shown in Table VI. As indicated in Table VI, Case 1 performed the best, and the values of the CRs of stage 1 and stage 2 in Case 1 were 99.53% and 99.60%, respectively. While Case 2 performed slightly better, the values of the CRs of stages 1 and 2 in Case 2 were 99.08% and 99.48%, respectively. Although Case 1 was consistently outperformed in both stages, the deviations were minor for both stage 1 and stage 2. This suggests that reducing two variables had a slight effect on the performance of the FDD. Therefore, the ANOVA was successful in removing variables while maintaining performance.

TABLE VI. FDD PERFORMANCES.

Case study	CR (%)		Time cost (s)	
	Stage 1	Stage 2	Training	Testing
Case 1	99.53	99.60	0.1277	0.4249
Case 2	99.08	99.48	0.1047	0.0758

The time consumed by both models in the training phase was the same. Case 1 consumed slightly more training time. In particular, the testing time for Case 1 consumed 3.3 times as much as the training time, but with Case 2, testing consumed 1.4 times less than the training time. That means that reducing the variables helped optimize time consumption. It makes the future promising for online FDD applications because of the timely response.

Fig. 4 shows the confusion matrices for Cases 1 and 2. At a glance, it can be seen that the performance of Case 1 was better than that of Case 2 because the numbers that were placed on the main diagonal were higher. Many faults were often incorrectly diagnosed as being of a normal status because the selected variables did not exhibit clear expressions suggestive of the faults being slightly severe. In particular, the ConFoul fault took a long time to display physical signs. In addition, ConFoul, RefOver, RefLeak, and ExcsOil had similar effects on the system. Therefore, these faults were easily misdiagnosed for each other. RefLeak and RefOver could occur anywhere because refrigerant moves around the system. These faults affected other physical variables, so the fault signs were difficult to judg and easy to diagnose incorrectly.

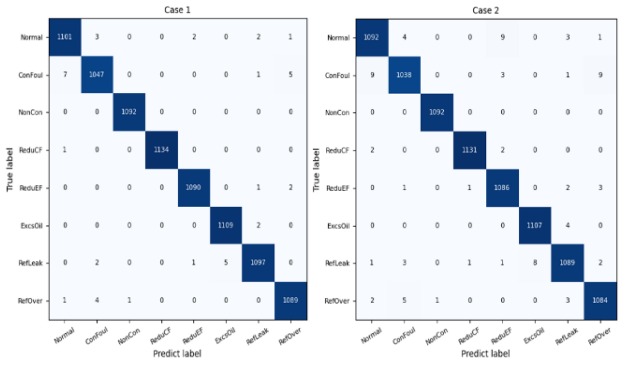

Figure 4. Confusion matrices of Cases 1 and 2 for each status.

The HR values of the two cases are shown in Table VII. The individual performance was the same in both cases for each status. Case 1 performances were almost always higher than those of Case 2 (excepting ExcsOil), but the deviations were small. When reducing the TEI variable that indicated sensitivity to ReduEF, the model without it had a lower result (the reduction was 1.09%). Other faults were affected by reducing the variables, but they were minimal. Especially in ExcsOil, the FDD-KNN models worked well, regardless of the number of variables. Furthermore, the uniformly high HR among the faults was an advantage. This study overcame a disadvantage of many previous studies, which is the difference between individual performance faults.

TABLE VII. INDIVIDUAL FDD HIT RATES OF THE TWO MODELS FOR EACH STATUS.

HR (%)	Normal	ConFoul	ExcsOil	ReduCF
Case 1	99.19	99.15	99.91	100
Case 2	98.73	98.76	99.91	99.82
HR (%)	ReduEF	NonCon	RefLeak	RefOver
Case 1	99.73	99.55	99.46	99.27
Case 2	98.64	99.28	98.82	98.64

Fig. 5 shows the diagnostic results for the severity levels. In general, the higher the severity level, the better the diagnostic performance. Because severe faults affected physical variables, they were easily detected. The ReduCF fault had the best performance, achieving 100% on all levels and in both cases. Reducing the variables did not affect the NonCon, ReduCF, ExcsOil, and RefLeak fault diagnostic results because the individual performances were equal in both cases.

The ConFoul, NonCon, ReduCF, ReduEF, and ExcsOil faults achieved 100% diagnoses in the two cases at severity level 4 (represented by the red line). The hit rates for RefLeak and RefOver in the two cases slightly declined. In particular, RefLeak and RefOver were complex faults. It was difficult to diagnose even severe faults. The cause was the widespread influence caused by the refrigerant. Refrigerant flows through the refrigeration cycle and passes through each component, causing changes and symptoms throughout. Therefore, RefLeak and RefOver required more variables to improve their performances. At severity level 3 (displayed by the black line), faults such as NonCon, ReduCF, ReduEF, and ExcsOil reached 100% in the two cases. The RefLeak and RefOver performances were equal in the two cases, but these had a similar tendency to be lower than other faults. The performances for ConFoul and ReduEF in Case 2 were lower than those in Case 1. Although only ReduCF had an absolute performance in the two cases at severity level 2 (represented by the blue line), other faults had high individual performances (greater than 98%) in the two cases. Many faults were unaffected by reducing the variables, such as NonCon, ExcsOil, and RefLeak. In particular, ReduCF, ExcsOil, and RefLeak all reached 100% for the two cases, regardless of fault diagnosis, at severity level 1 (represented by the green line). Similar to the previous levels, ConFoul and ReduEF decreased in hit rate when reducing the variables. This result demonstrated the effectiveness of the variable-reducing strategy. The main contribution of this study was the high performance diagnostics in the slightly severe faults.

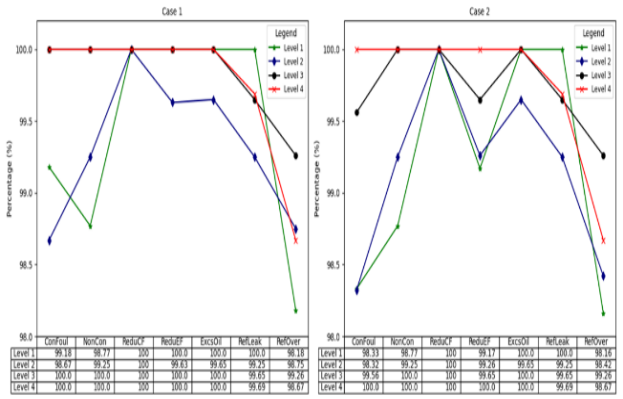

Figure 5. Individual FDD hit rates of the Case 1 and Case 2 models for each status of the severity levels.

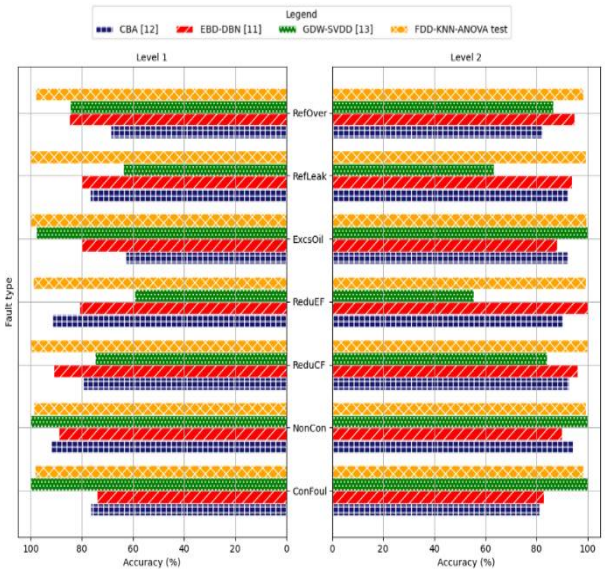

Figure 6. Comparison of FD performances for four different methods at two slight-severity levels.

Fig. 6 compares the chiller FDD performance of the proposed method (FDD-KNN-ANOVA) to those of other methods at two slight-severity levels. Compared to the CBA method, the EBD-DBN method, and the GDW-

SVDD method for Case 2 with slight-severity level 1, the chiller FDD performances of ReduCF, ReduEF, ExcsOil, RefLeak, and RefOver for the FDD-KNN-ANOVA method achieved the highest results: 100%, 99.17%, 100%, 100%, and 98.16%, respectively. For Case 2 with slight-severity level 2, the chiller FDD performances of ReduCF, ReduEF, ExcsOil, RefLeak, and RefOver for the FDD-KNN-ANOVA method achieved the highest results: 100%, 99.26%, 99.65%, 99.25%, and 98.42%, respectively. Although the chiller FDD performances of the GDW-SVDD method were the highest at 100% for the ConFoul, NonCon (levels 1 and 2), and ExcsOil (level 2), the chiller FDD performances for ReduEF and RefLeak were the lowest, corresponding to 59.5% and 63.75% (for level 1) and 55.5% and 63.25% (for level 2), respectively. This was a drawback of this method because the faults of ReduEF and RefLeak often occur in real operation conditions of chiller systems. Therefore, the proposed method (FDD-KNN-ANOVA) overcame the weaknesses of previous studies.

Real data under normal operating conditions and the condenser-fouling fault of a centrifugal water-cooled chiller from the Saigon Center building in Vietnam were used in this study to evaluate the reliability and performance of the proposed FDD-KNN-ANOVA strategy. The data were collected in 2019 from the building management system. After preprocessing the data, 1400 samples were generated (including 70% for the training dataset samples and 30% for the testing dataset samples), which were then used to validate the proposed method. The results are shown in Fig. 7. The blue area represents the data from the model classified as normal, while the red area represents the data from the model classified as ConFoul. The model correctly diagnosed the majority of the data points: the blue and red points were mostly located in the corresponding colored areas. However, there were a few red points of ConFoul that overlapped with the blue area for normal. This can be explained by the fact that these points represented data from when the system was just beginning to form ConFoul faults, the scale layer in the condenser was just forming, and the effect on the system's operation parameters was very small. As a result, these points represented a system with a slight fault, which was difficult to detect.

Fig. 7 also shows the TRC and TCO parameters in the normal and ConFoul states. When the scale layer was thick, the severity level of a fault rose due to increases in the TCO and TRC, causing the data to be distributed toward the right. When the system had normal operating conditions, the TRC and TCO parameters were lower, so the data area showing the normal state moved to the left. The results also indicated that the overall performance of the model was achieved at 97.61%, and the hit rate of condenser-fouling faults was 93.46%. Therefore, the KNN-FDD-ANOVA model was reliable and effective for chiller fault detection and diagnosis.

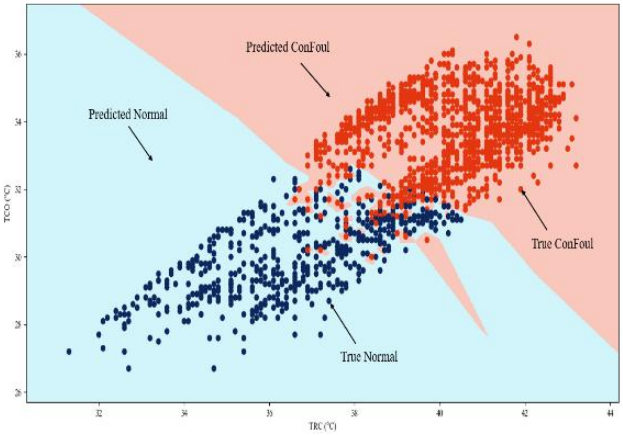

Figure 7. The classifications of normal and ConFoul using real data from the Saigon Center building, Ho Chi Minh City, Vietnam.

V. Conclusions

A chiller Fault Detection and Diagnosis (FDD) method based on the K-Nearest Neighbors (KNN) Algorithm and an Analysis of Variance (ANOVA) to reduce the number of sensors in a real system and to improve the performance of FDD was proposed in this study. Seven faults and four severity levels of each fault were considered. This method was validated using real data under normal operating conditions and the condenser-fouling fault of a centrifugal water-cooled chiller from the Saigon Center building in Vietnam and ASHRAE RP-1043 chiller operation data. The main conclusions of this paper were as follows:

1. The highest chiller FDD performance of the proposed method was achieved when compared to the CBA method, the EBD-DBN method, and the GDW-SVDD method for Case 2 of slight-severity levels 1 and 2.
2. The correct rates (CRs) of stages 1 and 2 in Case 1 were 99.53% and 99.60%, respectively, whereas the CRs of stages 1 and 2 in Case 2 were 99.08% and 99.48%, respectively.
3. The validation based on real data under normal operating conditions and the condenser-fouling fault of a centrifugal water-cooled chiller from the Saigon Center building in Vietnam revealed that the overall performance of the proposed method was 97.61%, and that the hit rate of condenser-fouling faults was 93.46%.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

L.H.D. Quan analyzed the data and L. M. Nhut supervised the research and wrote the paper, all authors had approved the final version.

Acknowledgment

The authors would like to thanks for the support from Ho Chi Minh City University of Technology and Education, Vietnam, as well as the use of their facilities in this study.

REFERENCES

[1] EMSD, "Hong Kong energy end-use data," *The Energy Efficiency Office, Electrical & Mechanical Services Department, Hong Kong,* 2008.

[2] DOE, "Buildings energy data book," *U.S. Department of Energy, Washington DC,* 2011.

[3] R. Lapisa, E. Bozonnet, P. Salagnac, and M. O. Abadie, "Optimized design of low-rise commercial buildings under various climates–Energy performance and passive cooling strategies," *Building and Environment,* vol. 132, pp. 83-95, 2018.

[4] H. Li and J. E. Braun, "Economic evaluation of benefits associated with automated fault detection and diagnosis in rooftop air conditioners," *ASHRAE Trans.,* vol. 113, no. 2, pp. 200–210, 2007.

[5] H. Wang, Y. Chen, C. W. H. Chan, J. Qin, and J. Wang, "Online model-based fault detection and diagnosis strategy for VAV air handling units," *Energy and Buildings,* vol. 55, pp. 252-263, 2012.

[6] D. W. Sun, *Handbook of Frozen Food Processing and Packaging*, *CRC Press,* 2016.

[7] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems, *Science and Technology for the Built Environment,* vol. 24, no. 1, pp. 3-21, 2017.

[8] H. Han, X. Cui, Y. Fan, and H. Qing, "Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features," *Applied Thermal Engineering,* vol. 154, pp. 540-547, 2019.

[9] G. Li, *et al.*, "An improved fault detection method for incipient centrifugal chiller faults using the PCA-R-SVDD algorithm," *Energy and Buildings,* vol. 116, pp. 104-113, 2016.

[10] H. Han, L. Xu, X. Cui, and Y. Fan, "Novel chiller fault diagnosis using deep neural network (DNN) with simulated annealing (SA)," *International Journal of Refrigeration,* vol. 121, pp. 269-278, 2021.

[11] Y. Wang, Z. Wang, S. He, and Z. Wang, "A practical chiller fault diagnosis method based on discrete Bayesian network," *International Journal of Refrigeration,* vol. 102, pp. 159-167, 2019.

[12] J. Liu *et al.*, "Data-driven and association rule mining-based fault diagnosis and action mechanism analysis for building chillers," *Energy and Buildings,* vol. 216, 2020.

[13] K. Chen, Z. Wang, X. Gu, and Z. Wang, "Multicondition operation fault detection for chillers based on global density-weighted support vector data description," *Applied Soft Computing,* vol. 112, 2021.

[14] R. Huang *et al.*, "An effective fault diagnosis method for centrifugal chillers using associative classification," *Applied Thermal Engineering,* vol. 136, pp. 633-642, 2018.

[15] S. He, Z. Wang, Z. Wang, X. Gu, and Z. Yan, "Fault detection and diagnosis of chiller using Bayesian network classifier with probabilistic boundary," *Applied Thermal Engineering,* vol. 107, pp. 37-47, 2016.

[16] K. Yan, Z. Ji, and W. Shen, "Online fault detection methods for chillers combining extended kalman filter and recursive one-class SVM," *Neurocomputing,* vol. 228, pp. 205-212, 2017.

[17] Y. Xia, Q. Ding, N. Jing, Y. Tang, A. Jiang, and S. Jiangzhou, "An enhanced fault detection method for centrifugal chillers using kernel density estimation based kernel entropy component analysis," *International Journal of Refrigeration,* vol. 129, pp. 290-300, 2021.

[18] Y. Gao, H. Han, Z. X. Ren, J. Q. Gao, S. X. Jiang, and Y. T. Yang, "Comprehensive study on sensitive parameters for chiller fault diagnosis," *Energy and Buildings,* vol. 251, 2021.

[19] Z. Wang, L. Wang, K. Liang, and Y. Tan, "Enhanced chiller fault detection using Bayesian network and principal component analysis," *Applied Thermal Engineering,* vol. 141, pp. 898-905, 2018.

[20] Z. Wang, L. Wang, Y. Tan, and J. Yuan, "Fault detection based on Bayesian network and missing data imputation for building energy systems," *Applied Thermal Engineering,* vol. 182, 2021.

[21] Y. Fan, X. Cui, H. Han, and H. Lu, "Feasibility and improvement of fault detection and diagnosis based on factory-installed sensors for chillers," *Applied Thermal Engineering,* vol. 164, 2020.

[22] H. Han, Z. Zhang, X. Cui, and Q. Meng, "Ensemble learning with member optimization for fault diagnosis of a building energy system," *Energy and Buildings,* vol. 226, 2020.

[23] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems,* vol. 14, no. 1, pp. 1-37, 2007.

[24] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers-A Tutorial," *ACM Computing Surveys,* vol. 54, no. 6, pp. 1-25, 2022.

[25] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing,* vol. 391, pp. 234-242, 2020.

[26] W. Zuo, D. Zhang, and K. Wang, "On kernel difference-weighted k-nearest neighbor classification," *Pattern Analysis and Applications,* vol. 11, no. 3-4, pp. 247-257, 2008.

[27] P. Karsmakers, K. Pelckmans, J. Suykens, and H. Van hamme, "Fixed-size kernel logistic regression for phoneme classification," presented at the Interspeech, 2007.

[28] K. Yan, A. Chong, and Y. Mo, "Generative adversarial network for fault detection diagnosis of chillers," *Building and Environment,* vol. 172, 2020.

[29] H. Han, B. Gu, T. Wang, and Z. R. Li, "Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning," *International Journal of Refrigeration,* vol. 34, no. 2, pp. 586-599, 2011.

[30] Y. Zhao, S. Wang, and F. Xiao, "Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD)," *Applied Energy,* vol. 112, pp. 1041-1048, 2013.

[31] S. M. Ross, "Analysis of variance," *Introduction to Probability and Statistics for Engineers and Scientists*, pp. 453-498, 2021.

[32] S. Garc á, J. Luengo, F. Herrera, *Data Preprocessing in Data Mining,* Springer, Berlin, 2015.

[33] I. H. Witten, E. Frank, M. A. Hall, "Data mining: Practical machine learning tools and techniques," *Morgan Kaufmann,* 2016.

[34] H. Han, B. Gu, J. Kang, and Z. R. Li, "Study on a hybrid SVM model for chiller FDD applications," *Applied Thermal Engineering,* vol. 31, no. 4, pp. 582-592, 2011.

[35] Z. Li *et al.*, "Machine learning based diagnosis strategy for refrigerant charge amount malfunction of variable refrigerant flow system," *International Journal of Refrigeration,* vol. 110, pp. 95-105, 2020.

[36] J. E. B. M. C. Comstock, "Development of analysis tools for the evaluation of fault detection and diagnostics in chillers ASHRAE research project RP-1043," *Purdue University, Ray W. Herrick Laboratories, West Lafayette,* 1999.