EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Application of various machine learning models for fault detection in the refrigeration system of a brewing company

Anusak Luekiangkhamla[1, 2], Natee Panagant[2, 3], Sujin Bureerat[2, 3] and Nantiwat Pholdee*[2, 3]

[1]Research in Machine House Department of Power Plant, Singha Beverage Company, Nakhon Pathom 73130, Thailand
[2]Department of Mechanical Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand
[3]Sustainable Infrastructure Research and Development Center, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand

Received 16 December 2022
Revised 3 March 2023
Accepted 16 March 2023

Abstract

Industrial refrigeration systems typically exhibit a higher energy consumption rate compared to other systems. It is a very important system, and for the brewing industry, it is at the heart of brewing as, in each process, the temperature of the beer must be controlled. Therefore, a health monitoring or faults detection system is a necessary tool for predictive and preventive maintenance to avoid faults in the system. This work presents machine learning models for faults prediction of the refrigeration system in a brewing firm. The raw data of processing parameters are collected, while the system health for each data set is identified and used for machine learning model training. Three operation cases based on the actual operation of the brewery's refrigeration system are analyzed in this work. Several machine learning models including Naïve Bayes (NB), Generalized Linear Models (GLMs), Logistic Regression (LR), Fast Large Margin (FLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT), and Support Vector Machine (SVM) are applied for refrigeration system fault detection in the three operation cases while their performances are investigated. The numerical simulation is performed based on RapidMiner Studio. According to the experiment of the three operation cases, the deep learning model was found to be the most accurate and required the least amount of time for the analysis. The accuracy percentages are 86.7%, 90.2%, and 82.5%, while the running times are 15 seconds, 27 seconds, and 15 seconds, respectively. This work can be considered as the baseline for future studies on applied machine learning models for fault detection in refrigeration systems.

Keywords: Predictive maintenance, Preventive management, Refrigeration system, Machine learning model, Data analysis, Fault detection

1. Introduction

Machining learning is a technique widely used in various applications to make classifications [1], predictions [2], or decisions [3] based on big data training to improve system performance. Several machine learning techniques have been successfully proposed in the global research community, while many industries are starting to apply machine learning in various subsystems in order to improve system performance [4]. For a refrigeration system, machine learning is successfully used for both system performance prediction and fault detection [5-8]. Studies have utilized various models such as the Multi-Layer Perceptron Artificial Neural Network (ANN-MLP), Support Vector Regression (SVR), Decision Tree (DT), and Random Forest (RF) to predict the performance of the evaporative condenser [6, 7]. Additionally, fault detection in refrigeration systems has been investigated using several models including convolutional neural networks (CNNs), SVM, principal components analysis-SVM, linear discriminant analysis-SVM, and linear discriminant analysis classifiers [8], where the models were able to identify up to twenty different types of faults in the system.

Although some machine learning has been applied to refrigeration systems for several purposes, most of the research work focused on modeling a single machine learning technique to attain the prediction or detection accuracy required, while the comparative performance of various existing machine learning techniques is rarely studied. Since numerous machining learning techniques have been proposed worldwide, to apply machine learning for specific applications, the performance comparison of several machine learning techniques still remains to be studied.

In this work, the machine learning model for faults prediction in the refrigeration system of a brewing company is presented, while the performances of several machine learning techniques are investigated. Firstly, training data is prepared. The raw data of the processing parameters are collected and the system health for each data set is identified. Then, several machine learning models are constructed based on the training. The machine learning models used in this study are Naïve Bayes (NB), Generalized Linear Models (GLMs), Logistic Regression (LR), Fast Large Margin (FLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT), and Support Vector Machine (SVM). Three operation cases based on the actual operation of the brewery's refrigeration system are used as a case study for investigating the performance of several machine learning models for fault detection in the refrigeration system.

*Corresponding author.
Email address: nantiwat@kku.ac.th
doi: 10.14456/easr.2023.16

2. Materials and methods

2.1 Refrigeration system

In this study, the refrigeration system is a vapor compression refrigeration system, illustrated in the flowchart in Figure 1. For each refrigeration cycle, the compressor receives low pressure and temperature refrigerant gas, and releases high pressure and temperature gas to the evaporating condenser. The evaporating condenser is responsible for decreasing the temperature of the refrigerant in the environment and releasing high-pressure, low-temperature liquid refrigerant, which flows in the pipe to the receiver tank. Then, the liquid refrigerant flows through the expansion valve to the economizer to separate the gas and liquid refrigerant. The refrigerant gas at high pressure and high temperature is sent to the compressor while the liquid refrigerant flows through the expansion valve to the separator tank. The liquid refrigerants and gas refrigerants are separated again at this stage. The liquid is taken to the plate heat exchanger and returned to the separator tank in a mixed state while the gas refrigerant is vaporized and returns to the heat exchanger again in the liquid state.

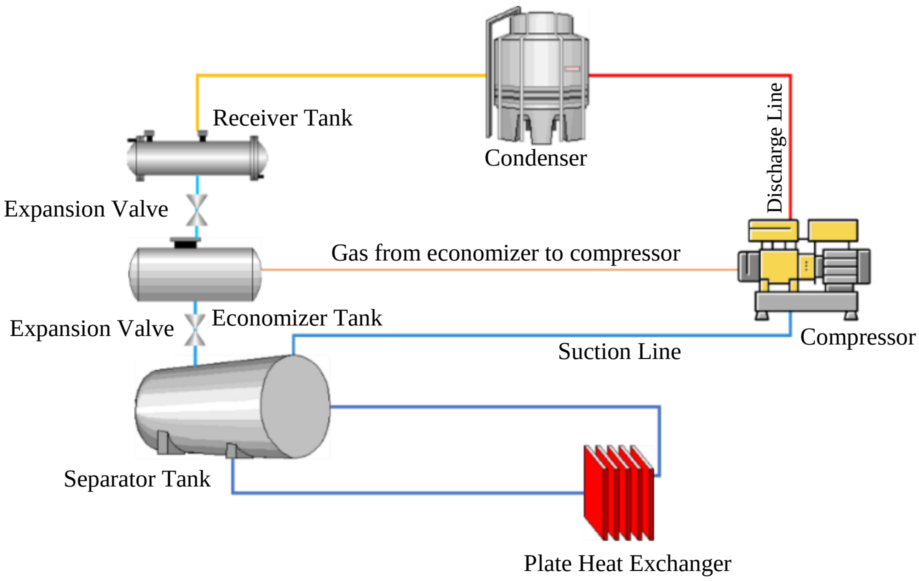

Figure 1 The cycle of the refrigeration system.

2.2 System fault detection

The refrigeration system studied is a large refrigeration system which has a large amount of measuring instruments, and several processing parameters. In this work, 27 processing parameters are selected as key features of the refrigeration system. These can be divided according to the fault of each device: a compressor fault with 14 parameters, a condenser fault with 3 parameters, a receiver fault with 2 parameters, an economizer fault with 3 parameters, and a separator fault with 4 parameters as detailed in Table 1. The raw data of the 27 parameters are collected based on the real operation of the refrigeration system, and the coefficient of performance (COP) of the system is estimated based on Equation (1). Faults of the system are identified based on the COP. For each set of processing parameters that obtain a COP close to the lower limit set point, a fault will be identified and action will need to be taken to prevent an actual fault in the system.

$$COP = \frac{Cooling\ Effect}{Power\ Input} \tag{1}$$

Table 1 Important parameters for feature modeling.

Type of fault	Symbol	Parameters	Symbol	Parameters
Compressor fault	R1	Suction pressure	R8	Evaporating temperature
	R2	Discharge pressure	R9	Condensing temperature
	R3	Oil pressure	R10	Oil filter pressure
	R4	Suction temperature	R11	Motor current
	R5	Discharge temperature	R12	Compressor power
	R6	Oil temperature	R13	Flow rate
	R7	Eco temperature	R14	Mass flow
Condenser fault	R15	Condenser pressure	R17	Ambient temperature
	R16	Humidity		
Receiver fault	R18	Receiver pressure	R19	Receiver level
Economizer fault	R20	Economizer pressure	R22	Economizer level
	R21	Economizer temperature		
Separator fault	R23	Separator pressure	R25	Separator level
	R24	Separator temperature	R26	Separator pressure pump
Status fault	R27	Status		

2.3 Machine learning model

In this work, RapidMiner Studio was used for investigating the performance of the various machine learning models. This program is a capable tool in the field of data management and data analysis. It is open-source software that is free to use, while a commercial license for professional paid support is also available. Another advantage is that machine learning can be modeled easily based on the graphic user interface (GUI) and coding required. Therefore, the software is suitable for the performance investigation of several machine learning models for the present application. Figure 2 shows the flow chart of the data analysis and machine learning modeling using RapidMiner Studio. The processes start with importing data and data cleaning (adding missing data, correcting, repairing, or removing incorrect or irrelevant data from a data set). Then, the feature engineering process is activated and the machine learning models are constructed.

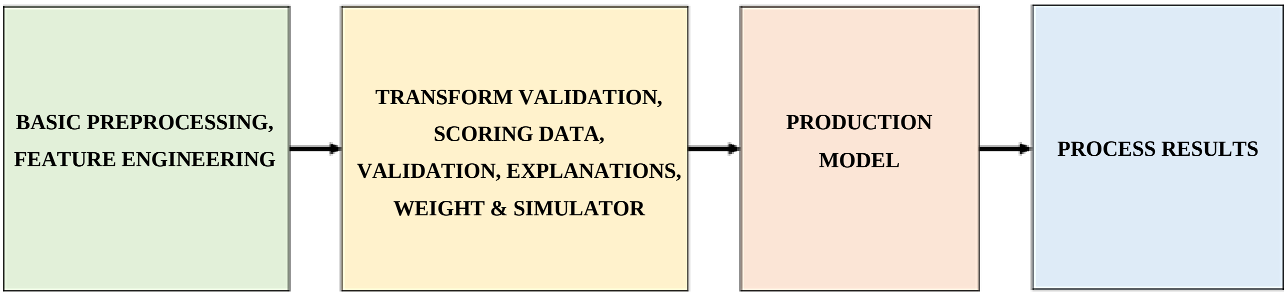

Figure 2 Block diagram of modeling processes.

In this work, 9 machine learning models are applied for fault detection in the refrigeration system and their performances are investigated. The machine learning models used are as follows:

Naïve Bayes (NB)

Naïve Bayes is a machine learning algorithm in the probabilistic classifiers category which classifies data based on the Bayes theorem [9]. The probabilistic model based on the Bayes theorem can be expresssed as follows:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{2}$$

where c and x are the class and attribute. $P(c|x)$ and $P(x|c)$ are respectively the posterior and likelihood probabilities, while $P(c)$ and $P(x)$ are the prior probability and predicted prior probability.

Generalized Linear Models (GLMs)

A generalized linear model is a combination of general classes of models which include logistic regression, Poisson regression, and multiple linear regression [10], and was first proposed by Nelder and Wedderburn. The model consists of three components: a random component, a systematic component, and a link function. The model prediction can be expressed as follows:

$$E(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{3}$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are regression coefficients while $\mathbf{x}=[x_1,\ldots, x_n]^T$ is the vector of independent variables.

Logistic Regression (LR)

Logistic regression is a type of supervised learning algorithm in machine learning that is used for solving binary classification problems [11]. The algorithm models the relationship between the independent variables and the binary outcome by utilizing a logistic function. For multiple independent variables, the logistic regression model can be expressed as follows:

$$P(\mathbf{x}) = \frac{1}{(1 + e^{-(\beta_0 + \beta_1 x_1 +,\ldots,\beta_n x_n)})} \tag{4}$$

where $P(x)$ = the probability estimation function, either 0 or 1, and $\mathbf{x}=[x_1,\ldots, x_n]^T$ is the vector of independent variables. The variables $\beta_0, \beta_1, \ldots, \beta_n$ are regression coefficients.

Deep Learning (DL)

Deep learning is a type of machine learning that utilizes multiple layers to successively uncover higher-level features from raw input data. With their roots in neural network theory, deep learning models make use of many hidden neurons and layers (usually more than two) compared to traditional neural networks. Additionally, they apply advanced learning algorithms such as autoencoders and restricted Boltzmann machines, which are not found in classical neural networks [12]. In supervised learning tasks, deep learning eliminates the need for manual feature engineering by automatically transforming data into compact intermediate representations similar to principal components, and creating layered structures that reduce redundant information in the data representation.

Decision Tree (DT)

A DT algorithm is widely used classification method in machine learning [13]. It works by constructing a tree-like model that depicts decisions and their outcomes based on input features. In the tree, internal nodes are tested on features, where branches represent the results of the tests, and leaf nodes represent class labels or predictions. The algorithm repeatedly divides the data into subsets based on the feature that provides the most information gain, until a stopping criterion such as a minimum number of samples per leaf or maximum tree depth is reached. The final tree structure can be used for making predictions on new data.

Random Forest (RF)

Random Forest (RF) is a powerful ensemble machine learning technique for both classification and regression problems [12, 14]. It functions by combining multiple decision trees into a single model and producing a prediction that is based on the most frequent outcome from individual trees. To build the individual trees, the model selects a random subset of the training data and a random subset of the features at each split, which infuses randomness into the model and reduces the risk of overfitting. The performance of the RT model is highly dependent on the selection of two key parameters: the number of decision trees generated and the number of features selected for each tree.

Gradient Boosted Trees (GBT)

A gradient boosted tree (GBT) is an advanced decision tree method that employs an optimization algorithm in the modeling process [12]. The GBT model is constructed by combining multiple simple trees, where each subsequent tree aims to correct the mistakes made by the previous tree. The trees are grown using gradient descent optimization, where the loss function is minimized iteratively, and the model tries to improve the prediction with each iteration. The prediction of the final model is the weighted sum of predictions of individual trees. GBTs are known to be powerful models that can handle both linear and non-linear relationships and handle complex interactions between features well.

Support Vector Machine (SVM)

The SVM is a powerful supervised machine learning algorithm [15]. It aims to find the optimal boundary, also known as a decision surface, that separates the data into different classes. The decision surface is defined by maximizing the margin between the closest data points of each class, referred to as support vectors. These support vectors are crucial in defining the decision surface and the algorithm balances the separation of the classes to avoid overfitting. SVM has the capability to handle both linear and non-linear boundaries, and is effective in high-dimensional data. Additionally, SVM is equipped to deal with imbalanced datasets and has a strong theoretical foundation.

Fast Large Margin (FLM)

FLM is a machine learning algorithm that can be classified as linear SVM [12]. The FLM model operates by constructing a decision boundary with a large margin, which refers to the distance between the decision boundary and the closest data points of each class. The large margin helps to increase the generalization ability of the model, reducing the risk of overfitting.

2.4 Numerical experiment

In this work, the fault detection problem of the refrigeration system is divided into three cases based on the actual three operation cases in the brewing company. Case I was the operating data of the parameters contained in compressor 1, when compressor 1 is working alone. Case II is the operating data of the parameters contained in compressor 2, when compressor 2 is operating alone. Case III is the operation data when compressor 1 and compressor 2 are operated simultaneously. These sample data based on the three cases are collected from real-time operation and recorded every 5 minutes for 3 days. The total sampling data for each case are 864 samples, 60% of which are used for training the machine learning models, and 40% for evaluation of the machine learning models' performance. Note that, for each case, the parameter sets that are not correlated with the model must be automatically eliminated during the model training process. The performances of the machine learning models are investigated based on the classification accuracy, which can be expressed as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5}$$

which can be calculated based on the confusion matrix, the tool for evaluating the results of the predictions, as shown in Figure 3, i.e.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

where

TP (True Positives)	= the prediction results match the actual values in cases where the prediction is true while the actual value is true.
TN (True Negatives)	= the prediction results match the actual values in case where the prediction is false while the actual value is false.
FP (False Positives)	= the prediction results do not match the actual values in cases where the prediction is true while the actual value is false.
FN (False Negatives)	= the prediction results do not match the actual values in cases where the prediction is false while the actual value is true.

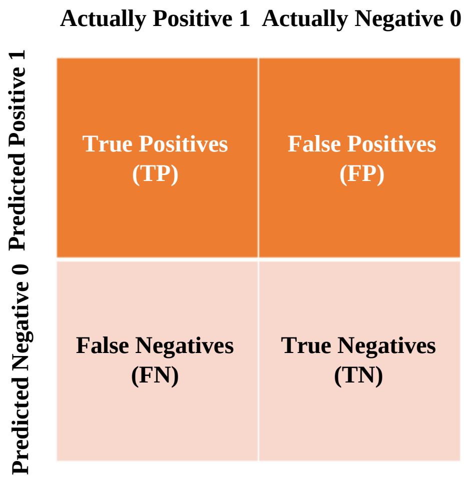

Figure 3 Confusion matrix of calculate accuracy.

3. Results and discussion

After performing the numerical experiment with the nine machine learning models on the three case studies of fault detection in the refrigeration system, the performance of the machine learning models based on the classification accuracy percentage and computational time are reported in Tables 2–4 for Case I, Case II and Case III, respectively. For Case I, based on Table 2, the most accurate machine learning model is GBT, while the second and third best are DL and NB, respectively. The computation times of GBT, DL and NB are 68, 15 and 17 seconds, respectively. For Case II, the results based on Table 3 show that the most accurate model is GBT, while the second and third best are DL and RF, respectively. The computation times of GBT, DL and RF are 82, 27 and 52 seconds, respectively. For Case III, the results based on Table 4 show that the best and second-best are GBT and DL, respectively, while the third best is NB, which is similar to Case I. The computation times of GBT, DL and NB for this case are 63, 15 and 11 seconds, respectively. In summary, the experimental results of the three cases show that GBT was the most accurate model, with accuracy of 88.8%, 91.6%, and 87.8%, respectively, as shown in Table 5. For each case, the parameter that was found to have no effect on the machine learning model is the motor current (R11).

For each case among the nine machine learning models, the SVM is found to be the worst model and is thus not suitable for this application in terms of accuracy and computational time. Although the SVM has been found to be accurate in some research work that is related to its application for a refrigeration system, such as that reported by Zahra et al. [8] and Zhengfei et al. [16], it is not guaranteed that the algorithm will always be accurate in all situations, as stated in the so-called "no free lunch" theory. Differences in the amount of training data, feature parameters and also the operation conditions led to differences in the models' performance.

This study found that GBT and DL are the best machine learning models for fault detection in the refrigeration system. GBT obtained the most accurate results for each case study, but with a high computational time, while DL obtained slightly lower accuracy for all the case studies with a very low computation time. Therefore, DL is considered to be the most suitable for this application in cases where the computation time is limited.

Table 2 Comparative accuracy and total time of dataset 1

Model	Accuracy	Total time	Ranking
NB	80.4%	17 s	3^{rd}
GLMs	77.0%	11 s	7^{th}
LR	71.7%	10 s	8^{th}
FLM	79.0%	13 s	4^{th}
DL	86.7%	15 s	2^{nd}
DT	78.8%	10 s	5^{th}
RF	78.3%	28 s	6^{th}
GBT	**88.8%**	**68 s**	**1^{st}(winner)**
SVM	66.8%	114 s	9^{th}

Table 3 Comparative accuracy and total time of dataset 2

Model	Accuracy	Total time	Ranking
NB	81.3%	17 s	5^{th}
GLMs	74.0%	17 s	7^{th}
LR	70.8%	17 s	8^{th}
FLM	82.5%	22 s	4^{th}
DL	90.2%	27 s	2^{nd}
DT	81.3%	18 s	6^{th}
RF	88.6%	52 s	3^{rd}
GBT	**91.6%**	**82 s**	**1^{st}(winner)**
SVM	63.8%	403 s	9^{th}

Table 4 Comparative accuracy and total time of dataset 3

Model	Accuracy	Total time	Ranking
NB	80.8%	11 s	3rd
GLMs	75.9%	10 s	6th
LR	55.1%	13 s	8th
FLM	72.0%	12 s	7th
DL	82.5%	15 s	2nd
DT	76.6%	10 s	5th
RF	80.4%	93 s	4th
GBT	**87.8%**	**63 s**	**1st(winner)**
SVM	53.5%	140 s	9th

Table 5 The best ranking results of all three cases.

Case	Model	Accuracy	Total time
Case I	GBT	88.8%	68 s
Case II	GBT	91.6%	82 s
Case III	GBT	87.8%	63 s

4. Conclusions

In this work, machine learning is successfully applied for fault detection in the refrigeration system of a brewing company. Several machine learning models are used to classify faults in the refrigeration system and their performances are investigated. Three case studies of the real operation condition are assigned, while the training and testing data are collected from real process operations. The numerical experiment is performed based on the open-source software, RapidMiner Studio. The results obtained reveal that the GBT model is the most accurate for all cases, but with a high computational time, while DL is slightly inferior to GBT but benefits from a very low computation time. Overall, DL can be considered as the most practical machine learning model for this application in cases where the computing time is limited. Of the nine machine learning models used in this study, SVM is found to be the worst and is not suitable for this application in terms of either accuracy or computational time. This work may serve as the baseline for future studies of the application of machine learning for fault detection in refrigeration systems. The development of an effective and efficient machine learning model for the application is interesting future work.

5. Acknowledgements

This research was supported by the RDI project (Research Development Innovation Management) and cooperation with Singha Beverage Company and the Engineering faculty of Khon Kaen University. The authors are also grateful for support from the National Research Council of Thailand (NRCT), Grant No. N42A650549.

6. References

[1] Jabbar MA. Breast cancer data classification using ensemble machine learning. Eng Appl Sci Res. 2021;48(1):65-72.
[2] Vorapatratorn S, Suchato A, Punyabukkana P. Fast obstacle detection system for the blind using depth image and machine learning. Eng Appl Sci Res. 2021;48(5):593-603.
[3] Ighravwe DE, Mashao D. Using a neural network model to determine electricity sales under renewable energy systems penetration consideration. Eng Appl Sci Res. 2021;48(1):73-82.
[4] Noyunsan C, Katanyukul T, Saikaew K. Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes. Eng Appl Sci Res. 2018;45(3):221-9.
[5] Prabha C, Arunkumar SP, Prabu K. Prediction of effect of refrigeration system parameters on the performance of a vapour compression refrigeration system using design of experiments. Int J Mech Prod Eng Res Dev. 2018;8(4):567-74.
[6] Behnam P, Faegh M, Shafii MB, Khiadani M. A comparative study of various machine learning methods for performance prediction of an evaporative condenser. Int J Refrig. 2021;126:280-90.
[7] Gupta AK, Kashyap S, Sarkar J. Machine learning model of regenerative evaporative cooler for performance prediction based on experimental investigation. Int J Refrig. 2022;137:178-87.
[8] Soltani Z, Sørensen KK, Leth J, Bendtsen JD. Fault detection and diagnosis in refrigeration systems using machine learning algorithms. Int J Refrig. 2022;144:34-45.
[9] Barathula S, Chaitanya S, Srinivasan K. Evaluation of machine learning models in the classification of pool boiling regimes up to critical heat flux based on boiling acoustics. Int J Heat Mass Transf. 2023;201:123623.
[10] Qaqish BF, Preisser JS. Generalized linear models for independent and dependent responses. In: Sen PK, Rao CR, editors. Handbook of Statistics Volume 18. Amsterdam: Elsevier; 2000. p. 193-222.
[11] Ambrish G, Ganesh B, Ganesh A, Srinivas C, Dhanraj, Mensinkal K. Logistic regression technique for prediction of cardiovascular disease. Glob Transit Proc. 2022;3(1):127-30.
[12] Nayak SK, Pradhan BK, Banerjee I, Pal K. Analysis of heart rate variability to understand the effect of cannabis consumption on Indian male paddy-field workers. Biomed Signal Process Control. 2020;62:102072.
[13] Bansal M, Goyal A, Choudhary A. A comparative analysis of K-Nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decis Anal J. 2022;3:100071.
[14] Robles-Velasco A, Cortés P, Muñuzuri J, De Baets B. Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. Expert Syst Appl. 2023;213:119050.
[15] Hu A, Liao H, Guan W, Dong J, Qian X. Support vector machine model based on OTSU segmentation algorithm in diagnosing bronchiectasis with chronic airway infections. J Radiat Res Appl Sci. 2023;16:100500.
[16] Li Z, Shi S, Chen H, Wei W, Wang Y, Liu Q, et al. Machine learning based diagnosis strategy for refrigerant charge amount malfunction of variable refrigerant flow system. Int J Refrig. 2020;110:95-105.