13th International Conference on Information Science and Technology (ICIST 2023)
Cairo, Egypt, December 8-14, 2023

HVAC System Fault Diagnosis via Feature Selection and Classification

Xuntan Ye*
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
xuntanye2-c@my.cityu.edu.hk

Hongzong Li*
Department of Computer Science
City University of Hong Kong, Kowloon, Hong Kong
Shenzhen Research Institute
City University of Hong Kong, Shenzhen, China
hongzli2-c@my.cityu.edu.hk

Jun Wang
Department of Computer Science
& School of Data Science
City University of Hong Kong
Kowloon, Hong Kong
Shenzhen Research Institute
City University of Hong Kong, Shenzhen, China
jwang.cs@cityu.edu.hk

2023 13th International Conference on Information Science and Technology (ICIST) | 979-8-3503-1392-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIST59754.2023.10367087

***Abstract*—Fault detection and diagnosis play a crucial role in energy savings in heating, ventilation, and air conditioning systems. It enables timely and appropriate repairs to prevent system malfunctioning and reduce energy waste. In this paper, a proposed feature selection algorithm or a neurodynamic optimization algorithm based on information gain is used for selecting a certain number of significant features, and then a stacking classifier is utilized to classify data into several different fault types by using the selected features. Experimental results are elaborated to demonstrate the superior performance of the proposed method against baselines in terms of accuracy on most of the datasets.**

***Index Terms*—Fault detection and diagnosis, heating ventilation and air conditioning (HVAC) systems, feature selection, classification**

I. Introduction

A heating, ventilation, and air conditioning (HVAC) system is a mechanical and thermal system that maintains suitable indoor air quality by controlling the levels of humidity, temperature, and air circulation [1]. HVAC systems are subject to numerous faults, such as cooling coil valve stuck, outdoor air damper stuck, and air handling unit duct leaking [2]. Faulty HVAC systems not only create inconvenience for users but also waste energy. Timely fault diagnosis (FD) can facilitate faster repair and reduce energy consumption [3].

Several FD methods are used in HVAC systems, including rule-based approaches [4]–[6], data-driven approaches [7]–[11], and model-based approaches [6], [12]–[15]. Rule-based approaches rely on predefined rules or thresholds to detect faults, such as hierarchical rule-based methods [4]. Data-driven approaches analyze historical data to identify patterns or anomalies that may indicate faults, such as the interval-valued features based machine learning technique [16]. Model-based approaches utilize mathematical models of HVAC systems to simulate and compare actual system behavior with expected behavior to detect faults. The approaches have their advantages and disadvantages. Rule-based approaches are simple to implement and are conceptually simple [17]. They are based on expert knowledge and can be effective in detecting known faults. However, they may not effectively detect complex or unknown faults and may generate false alarms or miss subtle ones [17]. Data-driven approaches rely on historical data for pattern recognition. They can effectively detect unknown faults and adapt to changing system conditions [18]. However, they may require extensive data collection and may not be suitable for systems with limited data availability [18]. Model-based approaches are based on physics-based or data-driven models of HVAC systems and can provide accurate and detailed fault detection and diagnosis. However, they may require complex modeling and may not be applicable to all HVAC system types [19]. The development and tuning of models are time-consuming and costly [17].

The data from HVAC systems usually comprise highly related features. It is necessary to carry out feature selection in HVAC FD. In this paper, a two-stage approach is proposed for HVAC FD, including a feature selection stage and a fault classification stage. A feature selection method is proposed by combining the affinity propagation algorithm and the ANOVA F-test. The proposed feature selection method or the neurodynamic optimization algorithm based on information gain (NIG) [20] is used for selecting significant features and reducing

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region of China under Grant 11203721.
*These authors contributed to the work equally and should be regarded as co-first authors

the dimensionality of the dataset. The resulting dimensionality-reduced dataset is classified into different fault types or normal data by using the stacking classifier. The experimental results demonstrate that both the proposed feature selection method and N-IG with stacking classifier (N-IG-SC) outperform the baselines in terms of accuracy.

The rest of the paper is organized as follows. Section II describes N-IG, affinity propagation algorithm, and ANOVA F-test. Section III introduces HVAC FD methods. Section IV reports experimental results on three datasets, followed by conclusions and future works in Section V.

II. Background information

Let $F = (F_1, F_2, F_3, ...F_m) \in R^{n\times m}$ denote the features of samples, $Y = (y_1, y_2, y_3, ..., y_n)^T \in R^n$ denote the target label, n is the number of samples, and m is the number of features.

A. Neurodynamic Optimization Algorithm Based on Information Gain

N-IG is a supervised feature selection method [20]. It minimizes redundancies between the features and maximizes relevancy between the features and the target label at the same time. The method ensures that the selected features are most relevant to the target label and least redundant by minimizing a fractional function with feature redundancy measures as the numerator and feature relevancy measures as the denominator. The feature relevancy is measured by the information gain, and the feature redundancy is measured using a similarity coefficient matrix Q:

$$Q = \delta I_p + S,$$

where I_p is an identity matrix, S is a similarity coefficient matrix. To ensure that the Q matrix is positive semidefinite, δI_p is added, where δ is defined as $\delta \geq -\min\{0, \lambda_{\min}(S)\}$ where $\lambda_{\min}(S)$ denotes the minimum eigenvalue of S. S is defined element-wisely as follows:

$$S_{ij} = \max\left\{0, \frac{I(F_i; F_j; Y)}{H(F_i) + H(F_j)}\right\},$$

where $H(F_i)$ is the entropy of F_i:

$$H(F_i) = -\sum_{f\in F_i} p(f)\log p(f),$$

where $F_i = (f_1, f_2, ..., f_n)$, $p(f)$ is the probability density function of f. $I(F_i; F_j; Y)$ is the multi-information between the i-th feature, the j-th feature, and target label Y:

$$\begin{aligned} I(F_i; F_j; Y) &= I(F_i; Y) + I(F_j; Y) - I(F_i, F_j; Y), \\ &= I(F_i; Y) - I(F_i, F_j \mid Y), \end{aligned}$$

where the joint mutual information $I(F_i, F_j; Y)$ is stated as follows:

$$I(F_i, F_j; Y) = \sum_{f\in F_i}\sum_{f'\in F_j}\sum_{y\in Y} p(f, f', y)\log\frac{p(f, f', y)}{p(f, f')p(y)},$$

and the conditional mutual information $I(F_i; F_j|Y)$ is stated as follows:

$$I(F_i; F_j|Y) = \sum_{f\in F_i}\sum_{f'\in F_j}\sum_{y\in Y} p(f, f', y)\log\frac{p(f, f'|y)}{p(f|y)p(f'|y)},$$

where $p(f, f')$ is the joint probability of f and f', $p(f|y)$ is the probability of f given y, $I(F_i; Y)$ is the multi-information that shows the amount of information shared by F_i and Y :

$$I(F_i; Y) = \sum_{f\in F_i}\sum_{y\in Y} p(f, y)\log\frac{p(f, y)}{p(f)p(y)},$$

If F_i, F_j are fully correlated with respect to the target y, then $S_{ij} = 1$. When F_i, F_j are not correlated at all, $S_{ij} = 0$.

By using relevancy maximization and redundancy minimization, a fractional program is stated as follows:

$$\begin{aligned} \min\ & \frac{w^T Q w}{\rho w}, \\ \text{s.t. } & e^T w = 1, \\ & w \geq 0, \end{aligned}$$

where Q is the similarity coefficient matrix, e is a vector of 1s, w is a vector of the scores of features to be determined, and ρ is the vector of feature relevancy defined using the information gain score:

$$\rho(F_i) = I(F_i; Y).$$

The optimization problem is solved by a two-layer recurrent neural network integrated with a projection neural network, generating a vector w^* containing feature importance scores. Then, a k-Winners-Take-All (kWTA) network selects the k-most typical features according to the features' weight values (the value of w) in w^* [20]. Fig. 1 shows the flowchart of N-IG.

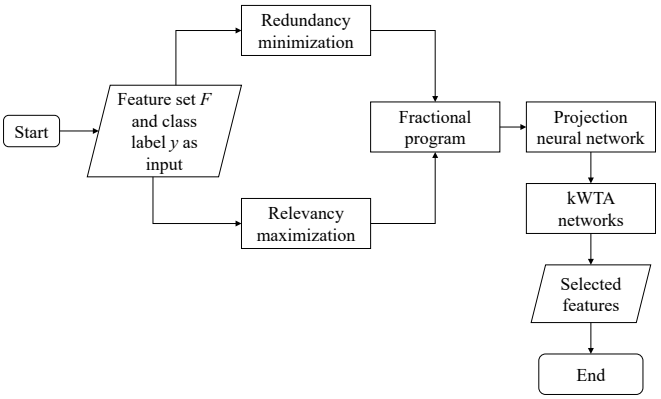

Fig. 1: A flowchart of N-IG [20].

B. Affinity Propagation Algorithm

Clustering is to group similar data into homogeneous clusters [21]–[23]. The affinity propagation algorithm is a clustering algorithm that utilizes the mechanism of message-passing between data points [24]. It does not require the number of

clusters to be determined in advance. It takes a matrix of similarity values as its input, where $s(i,k)$ is the similarity used as initial input to the algorithm, indicating how well the data point with index k is suited to be the exemplar for data point i. To minimize the squared error, the similarity between F_i and F_k is defined as: [24]

$$s(i,k) = -||F_i - F_k||^2. \tag{1}$$

$r(i,k)$ is defined as the responsibility of the data point with index k to the data point with index i. The responsibility means how well the data point with index k (candidate exemplars) can serve as the exemplar for the data point with index i, with consideration to the other possible exemplar points [24]. The value of $r(i,k)$ is updated as follows:

$$\tilde{r}(i,k) \leftarrow s(i,k) - \max_{k',\, s.t.\ k' \neq k} \{a(i,k')_\tau + s(i,k')\}, \tag{2}$$

where τ is defined as the number of iterations, $a(i,k)$ is the availability of the i-th data point to select the k-th data point as its exemplar [24]. $a(i,k)$ is updated as follows:

$$\tilde{a}(i,k) \leftarrow \begin{cases} \min\Big\{0, r(k,k)_\tau + \sum\limits_{i',\, s.t.\ i' \notin \{i,k\}} \max\{0, r(i',k)_\tau\}\Big\}, i \neq k, \\ \sum\limits_{i',\, s.t.\ i' \neq k} \max\{0, r(i',k)_\tau\}, i = k. \end{cases} \tag{3}$$

To avoid oscillations in the updating process, a damping factor λ is used for updating the values of $r(i,k)$ and $a(i,k)$ as follows:

$$r(i,k)_{\tau+1} = \lambda r(i,k)_\tau + (1-\lambda)\tilde{r}(i,k), \tag{4}$$

$$a(i,k)_{\tau+1} = \lambda a(i,k)_\tau + (1-\lambda)\tilde{a}(i,k). \tag{5}$$

For point i, the value k that maximizes $r(i,k) + a(i,k)$ is the exemplar for i if $k \neq i$. If $k = i$, then i-th point is an exemplar.

C. Analysis of Variance (ANOVA) F-test

The ANOVA F-test evaluates the differences of means between groups within a dataset. The different groups are determined by the categorical target label which separates the data samples into different groups by their fault types. The F-value in one-way ANOVA under feature k is stated as follows:

$$\mathcal{F}_k = \frac{S_{bk}^2}{S_{wk}^2}, \tag{6}$$

where k is the index of feature, S_{bk}^2 is the variance between different groups of samples, and S_{wk}^2 is the variance within groups. S_{bk}^2 and S_{wk}^2 are stated as follows:

$$S_{bk}^2 = \frac{\sum_{i=1}^{p} n_i(\overline{x_{ik}} - \overline{X_k})^2}{p-1},$$

where p is the number of groups, $\overline{X_k}$ is the mean of all samples for feature k, $\overline{x_{ik}}$ is the mean of group i under feature k, n_i is the size of samples for group i, $p-1$ is the degree of freedom [25].

$$S_{wk}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ijk} - \overline{x_{ik}})^2}{\sum_{i=1}^{p}(n_i) - p},$$

where p is the number of groups, n_i is the number of samples for group i, $\overline{x_{ik}}$ is the mean of samples within group i under feature k, x_{ijk} is the value of sample j in group i under feature k, $\sum_{i=1}^{p}(n_i) - p$ is the degree of freedom [25].

D. A Stacking Classifier

Classification is the process of categorizing items or data into distinct categories based on shared characteristics or attributes [26]. A stacking classifier is an ensemble involving two kinds of classifiers: base and meta classifiers [27]. The number of base classifiers is usually between one and ten, and the number of meta classifiers is one. Several base classifiers are used to generate prediction results independently called meta-features. The meta-classifier then uses the meta-features to generate the final prediction result. In the literature, a stacking classifier will perform better than a standalone classifier, such as XGBoost [28] and LightGBM [29]. The stacking classifier has been used in solving problems such as diabetes prediction [30], membrane protein type prediction [31], image classification [32], depression prediction [33], music genre classification [34], and achieves satisfactory results.

III. Methodology

In this paper, the Affinity propagation – Analysis of Variance (AP-ANOVA) feature selection method is proposed to address the need to determine the number of features required by hyperparameter tuning. Inspired by the simultaneous relevancy maximization and redundancy minimization principle [20], the proposed AP-ANOVA algorithm clusters all the features into several feature subsets by using the affinity propagation algorithm. Features in each feature subset are considered to be related to each other, and only one feature needs to be selected in each subset to prevent feature redundancies. Then, the ANOVA F-value is used to evaluate the relationship between the features in each subset and the target label. Each feature with the highest ANOVA F-value is selected in each feature subset. A higher ANOVA F-value means the feature and the target label are less similar, and a feature different from the target label means the feature is more useful for label prediction. The feature with the highest ANOVA F-value in a cluster is then selected. After step 3, the features selected are then used for fault classification.

A. System Flow

Fig. 2 shows the system flow for the HVAC system FD in this paper. A two-stage approach is used for HVAC system FD. The first stage is the feature selection stage. Two different feature selection algorithms are used, AP-ANOVA and N-IG. The dataset after the feature selection is then split into the training set and the testing set. The second stage of HVAC FD involves classifying the data samples into different fault types and the normal type. The training set is used to train

Algorithm 1: AP-ANOVA algorithm

Input: All features f in the dimension $R^{n\times m}$, target Y in the dimension R^n.

Output: Indices of selected features $[f_0, f_1, f_2, \ldots, f_s]$ where s is the number of selected features.

Calculate the similarity matrix using Eqn. 1;

$\tau \leftarrow 0$;

repeat

Update $r(i,k)_\tau$ using Eqn. 2 and 4;

Update $a(i,k)_\tau$ using Eqn. 3 and 5;

$\tau \leftarrow \tau + 1$;

until convergence

foreach *data point* i **do**

Select exemplar k with the largest value of $r(i,k) + a(i,k)$;

end

for $i = 1$ *to* n **do**

Calculate ANOVA F-value of feature F_i using Eqn. 6;

end

In each cluster, select one feature with the highest ANOVA-F value;

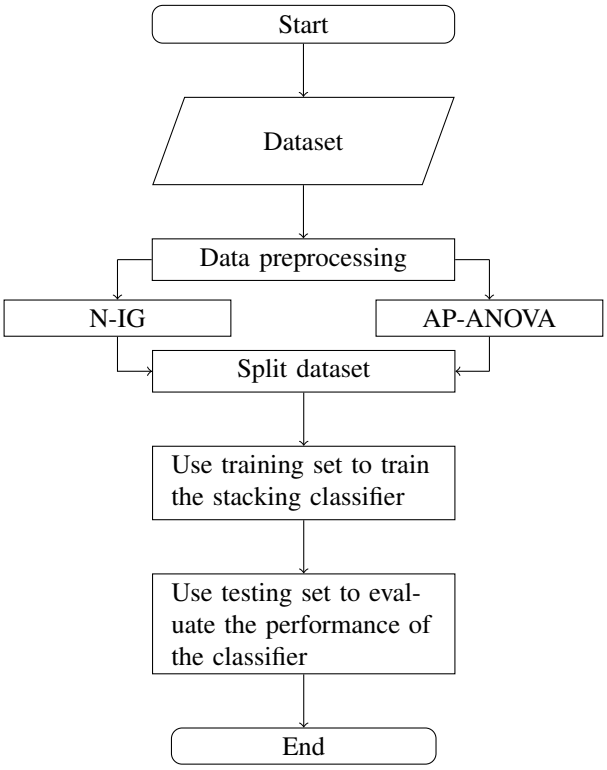

Fig. 2: A flowchart of AP-ANOVA

the stacking classifier. Repeated stratified cross-validation is used with five splits and two repeats to prevent overfitting. Afterward, the trained stacking classifier is used to predict the labels for the data samples in the testing set.

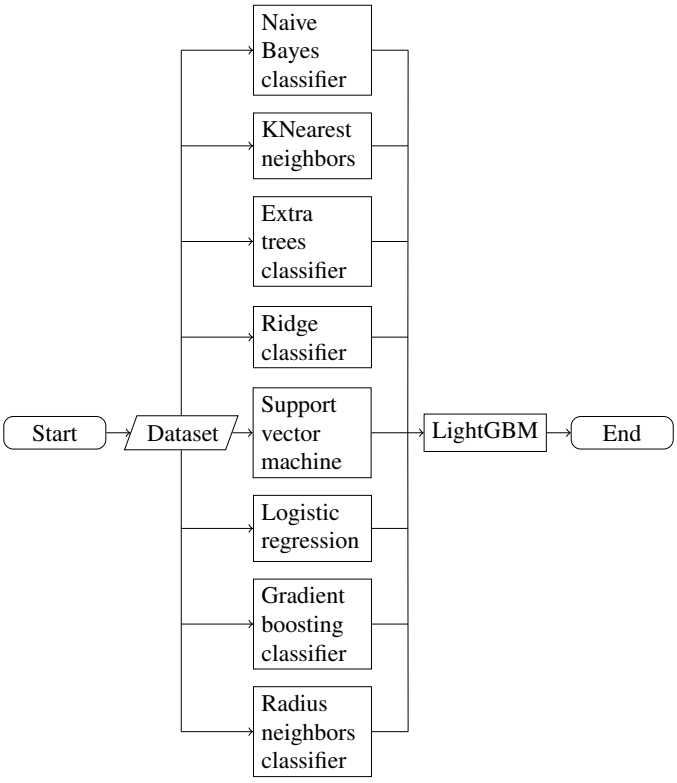

Fig. 3: A stacking classifier

The stacking classifier used in this experiment is shown in Fig. 3. A wide range of base classifiers is used, including tree-base algorithms (Extra Trees and Gradient Boosting), support vector machine, probability-based model (Naïve Bayes), nearest neighbors algorithms (K-Neighbors and Radius Neighbors Classifier), and linear models (Ridge Classifier and Logistic regression (LR)). Two methods used in the literature [31], [35], [36], XGBoost and LightGBM, are initially considered for the meta-classifiers. LightGBM has demonstrated faster computational speed, better memory utilization, and slightly better classification performance than XGBoost in [29]. Therefore, LightGBM is used as the meta-classifier in the stacking classifier.

IV. Experimental results

A. Setups

In experiments, three datasets from ASHRAE Research Project 1312 (RP1312)[1] are used. RP1312 includes two identical air handling units (AHUs), namely AHU-A and AHU-B. AHU-A operates with different kinds of faults, and AHU-B operates normally without any faults. Each type of fault in AHU-A is observed for a single day [37]. The data collection for RP1312 spans three distinct seasons: the summer of 2007, the spring of 2008, and the winter of 2008. Each season presents different typical faults. In summer, the typical

[1]https://www.techstreet.com/standards/rp-1312-tools-for-evaluating-fault-detection-and-diagnostic-methods-for-air-handling-units?product_id=1833299

faults include a stuck cooling coil valve, unstable cooling coil valve control, and complete failure of return fans. In spring, the typical faults include unstable heating and cooling sequences, etc. In winter, the typical faults are related to heating coils, including reduced heating coil capacity and heating coil fouling [2]. Table I lists the details of the three datasets in RP1312, including the number of normal samples, the number of fault samples, and the number of fault types. As shown in Table I, the number of normal samples and the number of fault samples are the same in the three datasets. The number of typical faults varies across seasons, with the spring of 2008 dataset containing the highest number of typical faults, and the winter of 2008 dataset containing the lowest number of typical faults. The datasets contain infinity, negative infinity, and NaN (Not a Number) values. To facilitate feature extraction, the infinity values are replaced with 1, the negative infinity values are replaced with -1, and the NaN values are removed. Each dataset comprises data samples from both AHU-A and AHU-B, with the normal samples labeled as 100 and the faulty samples labeled within the range of $[0, l-1]$ according to their fault types, where l is the number of faults in each dataset. Consequently, the FD problem becomes a classification problem with $l+1$ classes.

TABLE I: Information about the three datasets in RP1312 [2]

Datasets	# of samples		
	normal samples	fault samples	fault types
2007 summer	18720	18720	13
2008 spring	27360	27360	19
2008 winter	14400	14400	10

The performances of the proposed methods are mainly evaluated by accuracy. The model is trained using datasets with different proportions of fault samples according to [2]. The datasets are divided through random sampling, allocating 50% of the data as the training set, while the remaining 50% is designated as the testing set.

In N-IG, The numbers of selected features are set to 20, 25, and 30. The baseline methods used in the following experiments include EKF-CS-D-ELM [2], stacking classifier (SC) [27], logistic regression (LR) [38], k-nearest neighbors (KNN) [39], naive Bayes (NB) [40], and support vector machine (SVM) [41].

B. Experiment on 2007 Summer Dataset

Table II records the mean accuracy values of the fault types obtained by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs with random initialization on the 2007 Summer dataset, where the best and second-best results are highlighted in bold and underlined, respectively. As shown in Table II, AP-ANOVA-SC, and N-IG-SC outperform the baselines in terms of the accuracy values on most of the fault types. Table III records the mean values and standard derivations of the accuracy of the results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs with random initialization on the 2007 Summer Dataset. As shown in Table III, N-IG-SC and AP-ANOVA outperform the baselines, where N-IG-SC achieves accuracies ranging from 99.37% to 99.55%, and AP-ANOVA-SC achieves an accuracy of 99.43%, and the baselines achieve accuracies ranging from 66.20% to 98.92%. The standard deviation of the results by using N-IG-SC and AP-ANOVA-SC is the lowest compared to the baselines, indicating a more stable and consistent performance of the proposed methods.

C. Experiment on 2008 Spring Dataset

Table IV records the mean accuracy values of the fault types obtained by using AP-ANOVA-SC, N-IG-SC with different selected feature numbers, and the five baselines over 30 runs with random initialization on the 2008 Spring dataset. As shown in Table IV, N-IG-SC outperforms the baselines in terms of accuracy for most fault types. N-IG-SC with $s = 30$ achieves the highest accuracy for 15 out of 19 fault types.

Table V records the mean values and standard derivations of the accuracy of the results by using AP-ANOVA-SC, N-IG-SC with different selected feature numbers, and the five baselines over 30 runs with random initialization on the 2008 Spring dataset. As shown in Table V, N-IG-SC with $s = 30$ achieves the highest mean accuracy (i.e., 99.49%), indicating its effectiveness in fault classification.

D. Experiment on 2008 Winter dataset

Table VI records the mean accuracy values of the fault types obtained by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the six baselines over 30 runs with random initialization on the 2008 Winter dataset. As shown in Table VI, AP-ANOVA-SC outperforms the baselines in terms of accuracy for most fault types, and achieves the highest accuracy for 7 out of 10 fault types. Table VII records the mean values and standard derivations of accuracies of the results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs with random initialization on the 2008 Winter dataset. As shown in Table VII, AP-ANOVA-SC achieves the highest mean accuracy (i.e., 99.71%), indicating its effectiveness in fault classification.

E. Discussion of Results

Fig. 4 summarizes the mean values of accuracies resulting from AP-ANOVA-SC, ND-SC and the five baselines on the three datasets. As shown in Fig. 4, N-IG-SC outperforms SC and the other baselines on all datasets in terms of accuracy. AP-ANOVA-SC outperforms SC on the 2007 Summer and 2008 Winter datasets, and it outperforms the baselines on the three datasets except for SC in terms of accuracy.

It is worth noting that the accuracy of the algorithms decreases with the decrease in the number of samples. In particular, the dataset of 2008 Winter has the lowest number of samples (i.e., 14,400) and the dataset of 2008 Spring has the

TABLE II: The mean accuracy values of the fault types obtained using AP-ANOVA-SC, N-IG-SC with different numbers of selected features s, and the six baselines over 30 runs on the 2007 Summer Dataset.

Fault types	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	SC	LR	KNN	NB	SVM	EKF-CS-D-ELM [2]
F0	99.2	99.2	**99.3**	99.1	99.0	27.8	98.2	59.8	28.6	91.3
F1	98.9	**99.4**	98.8	98.5	98.4	0.1	97.2	41.4	39.2	95.2
F2	99.9	**100**	99.9	99.9	99.1	**100**	99.2	99.9	**100**	84.1
F3	99.5	99.5	**99.6**	99.3	99.3	40.3	99.1	40.0	40.2	93.1
F4	98.9	**99.0**	98.7	98.3	97.3	0.1	97.1	23.6	65.7	94.8
F5	97.7	98.9	**99.0**	98.9	98.2	41.9	97.6	48.9	58.7	91.3
F6	98.7	98.8	99.0	98.9	98.3	68.1	98.5	**99.2**	62.4	96.4
F7	97.3	**98.4**	98.2	98.1	96.8	0.8	95.5	0.7	33.9	95.9
F8	98.5	99.0	**99.2**	98.7	97.5	47.5	96.0	94.9	51.7	96.7
F9	99.5	99.7	**99.8**	**99.8**	94.6	62.6	95.0	59.5	69.9	99.1
F10	**99.3**	98.8	98.4	98.4	98.2	10.0	96.7	14.4	66.2	92.4
F11	**99.4**	99.3	98.8	97.0	97.2	0.0	96.6	7.6	39.8	94.4
F12	**98.9**	98.3	98.6	98.7	98.4	30.1	98.5	11.0	50.2	97.6

TABLE III: The mean values and standard derivations of accuracies of the results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the six baselines over 30 runs on the 2007 Summer dataset.

2007 summer dataset	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	SC	LR	KNN	NB	SVM
mean values	99.43	**99.55**	99.51	99.37	98.92	66.20	98.59	73.15	76.83
standard deviations	0.000417	**0.000196**	0.000264	0.000245	0.000954	0.006241	0.001013	0.010439	0.012373

TABLE IV: The mean accuracy values of the fault types results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs on the 2008 Spring dataset.

Fault types	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	SC	LR	KNN	NB	SVM	EKF-CS-D-ELM [2]
F0	91.8	**98.8**	98.7	98.1	92.9	0.1	94.0	2.2	12.5	90.4
F1	97.7	**98.3**	98.2	97.6	93.7	1.4	93.8	20.6	7.3	92.7
F2	97.5	**99.5**	99.4	99.4	98.9	59.3	99.3	95.8	51.9	95.3
F3	96.9	**99.6**	**99.6**	**99.6**	99.1	83.5	98.5	4.7	63.5	92.8
F4	93.1	99.3	**99.4**	99.1	94.8	0.1	97.2	38.7	59.6	92.1
F5	94.0	**99.4**	99.3	**99.4**	98.7	48.5	98.9	0.3	69.8	88.8
F6	96.9	**99.4**	99.3	**99.4**	98.9	23.8	99.3	0.6	69.6	89.9
F7	99.2	99.3	99.3	**99.5**	97.8	39.6	96.5	58.3	39.0	96.3
F8	97.7	**99.9**	99.7	99.7	96.9	39.4	97.8	26.2	47.7	87.6
F9	97.4	**99.3**	99.0	**99.3**	96.6	2.9	97.3	14.2	66.9	93.5
F10	97.2	**99.1**	99.0	98.0	98.9	35.7	98.3	0.4	70.4	87.5
F11	99.1	**100**	**100**	99.1	98.9	66.2	99.1	46.1	73.0	97.4
F12	99.3	99.6	**99.7**	99.1	99.0	51.4	99.1	67.3	67.8	91.8
F13	96.7	98.0	97.8	**98.2**	**98.2**	45.2	97.7	21.2	59.4	89.0
F14	90.5	**99.0**	98.3	97.9	94.5	12.7	93.4	7.9	18.1	96.4
F15	91.3	**97.6**	97.1	96.6	88.1	0	85.5	1.9	17.0	90.6
F16	92.5	**99.2**	99.1	98.7	96.6	0.6	93.9	16.9	29.6	89.6
F17	92.6	**97.6**	97.4	96.8	94.3	0.1	93.5	4.0	10.6	93.3
F18	92.9	**98.7**	**98.7**	97.6	96.1	20.0	94.8	93.5	52.4	95.7

TABLE V: The mean values and standard derivations of accuracy of the results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs on the 2008 Spring dataset

2008 spring dataset	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	SC	LR	KNN	NB	SVM
mean values	97.74	**99.49**	99.42	99.26	98.22	63.91	98.04	46.81	73.06
standard deviations	0.000589	**0.000174**	0.000229	0.000241	0.0012	0.005329	0.000727	0.007848	0.007871

TABLE VI: The mean accuracy values of the fault types results obtained by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the six baselines over 30 runs on the 2008 winter dataset

Fault types	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	EKF-CS-D-ELM [2]	SVM	KNN	NB	LR	SC
F0	99.1	**99.2**	**99.2**	98.6	96.4	63.9	96.8	97.6	72.6	97.2
F1	**98.8**	98.4	98.5	98.6	95.1	66.9	97.4	45.7	56.6	97.3
F2	**99.6**	99.3	99.3	99.1	94.8	46.9	98.7	49.7	34.7	99.1
F3	**99.8**	99.2	99.2	98.5	91.3	88.1	98.5	87.2	41.2	99.1
F4	**99.6**	99.2	99.2	99.3	96.6	97.7	98.9	95.2	61.6	98.8
F5	**99.8**	99.5	99.6	99.6	99.5	58.0	98.3	44.9	9.8	98.5
F6	99.3	**99.5**	99.3	99.4	92.8	70.7	99.0	22.0	58.8	98.8
F7	99.3	**99.4**	**99.4**	99.3	93.2	69.2	99.0	27.3	17.2	98.7
F8	**99.3**	**99.3**	99.2	98.8	93.0	33.3	97.8	58.2	11.9	98.7
F9	**99.7**	99.3	99.1	99.2	94.4	98.6	98.4	73.1	38.7	98.7

TABLE VII: The mean values and standard derivations of accuracy of the results by using AP-ANOVA-SC, N-IG-SC with different numbers of selected features, and the five baselines over 30 runs on the 2008 Winter Dataset

2008 winter dataset	AP-ANOVA-SC	N-IG-SC ($s = 30$)	N-IG-SC ($s = 25$)	N-IG-SC ($s = 20$)	SC	LR	KNN	NB	SVM
mean values	**99.71**	99.57	99.56	99.47	99.24	70.15	99.10	80.06	84.65
standard deviations	**0.000164**	0.000281	0.000223	0.000382	0.000886	0.009807	0.000981	0.011404	0.005179

highest number of samples (i.e., 27,360), and the algorithms achieve better results on the 2007 Summer dataset than on 2008 Spring in terms of accuracy. As shown in Tables III, V, and VII, the standard deviations of accuracies resulting from AP-ANOVA-SC and N-IG-SC are smaller than the five baselines, the mean values of accuracies resulting from AP-ANOVA-SC and N-IG-SC are higher than the five baselines. This shows that AP-ANOVA-SC and N-IG-SC achieve more consistent and efficient results than the baselines. Additionally, the decrease in accuracy of the results using AP-ANOVA is more pronounced if the number of samples increases compared to N-IG-SC. It shows that N-IG-SC performs more robustly than AP-ANOVA.

F. Ablation studies

Tables III, V, and VII reveal that SC achieves better performance than LR, KNN, NB, and SVM in terms of mean values of accuracies on the three datasets. By adopting N-IG for feature selection, N-IG-SC achieves better performance than SC on the three datasets. By introducing AP-ANOVA, AP-ANOVA-SC achieves better performance than SC statistically on most of the three datasets.

V. CONCLUDING REMARKS

In this paper, HVAC fault diagnosis is based on two feature selection algorithms and a stacking classifier. N-IG and the proposed AP-ANOVA algorithm are used for feature selection, and the stacking classifier is utilized for fault classification by using the selected features. Compared to the neurodynamic optimization algorithm based on information gain, the AP-ANOVA feature selection method does not require setting the number of selected features. The experimental results show that N-IG-SC and AP-ANOVA-SC outperform the six baselines in terms of most of the mean accuracy values on three datasets. Future investigations may aim at improving the efficiency of the algorithms for feature selection by using labeling information.

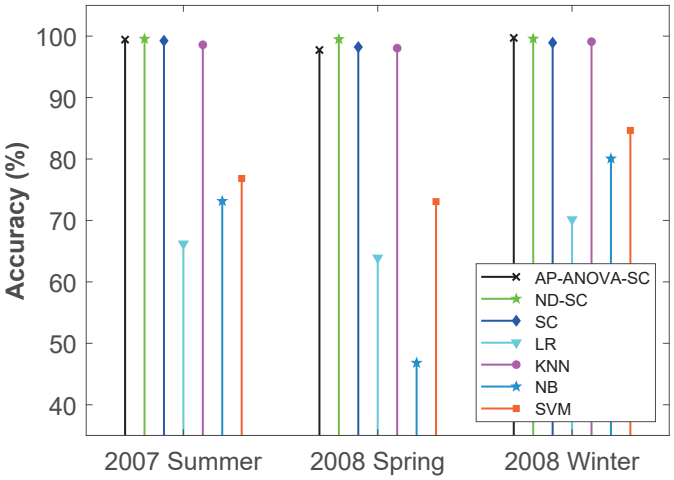

Fig. 4: Mean values of accuracies resulting from AP-ANOVA-SC, ND-SC and the five baselines on the three datasets.

REFERENCES

[1] J. Solano, E. Caamaño-Martín, L. Olivieri, and D. Almeida-Galárraga, "HVAC systems and thermal comfort in buildings climate control: An experimental case study," *Energy Reports*, vol. 7, pp. 269–277, 2021.

[2] K. Yan, Z. Ji, H. Lu, J. Huang, W. Shen, and Y. Xue, "Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1349–1356, 2019.

[3] F. Guo, A. P. Rogers, and B. P. Rasmussen, "Multivariate fault detection for residential HVAC systems using cloud-based thermostat data, part I: Methodology," *Science and Technology for the Built Environment*, vol. 28, no. 2, pp. 109–120, 2022.

[4] J. Schein and S. T. Bushby, "A hierarchical rule-based fault detection and diagnostic method for HVAC systems," *HVAC&R Research*, vol. 12, no. 1, pp. 111–125, 2006.

[5] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House, "A rule-based fault detection method for air handling units," *Energy and Buildings*, vol. 38, no. 12, pp. 1485–1492, 2006.

[6] H. Wang, Y. Chen, C. W. 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.

[7] M. S. Mirnaghi and F. Haghighat, "Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review," *Energy and Buildings*, vol. 229, p. 110492, 2020.

[8] S. Pan, Z. Ye, and J. Zhou, "Fault detection filtering for a class of non-homogeneous Markov jump systems with random sensor saturations," *International Journal of Control, Automation and Systems*, vol. 18, pp. 439–449, 2020.

[9] Y. Shen and K. Khorasani, "Hybrid multi-mode machine learning-based fault diagnosis strategies with application to aircraft gas turbine engines," *Neural Networks*, vol. 130, pp. 126–142, 2020.

[10] D. Xiao, B. T. Le, Z. Yu, C. Liu, H. Li, Q. He, H. Xie, and J. Wang, "A method of fault monitoring and diagnosis for the thickener in hydrometallurgy," *IEEE Access*, vol. 7, pp. 142 317–142 324, 2019.

[11] D. Xiao, H. Li, B. T. Le, S. Zhang, J. Wang, D. He, and X. Fu, "Research on a method of gross error elimination for slope monitoring data based on machine learning," *IEEE Access*, vol. 7, pp. 164 682–164 695, 2019.

[12] M. Schmid, E. Gebauer, C. Hanzl, and C. Endisch, "Active model-based fault diagnosis in reconfigurable battery systems," *IEEE Transactions on Power Electronics*, vol. 36, no. 3, pp. 2584–2597, 2021.

[13] M. Mansouri, M.-F. Harkat, H. N. Nounou, and M. N. Nounou, *Data-driven and model-based methods for fault detection and diagnosis*. Elsevier, 2020.

[14] Q. Zhou, S. Wang, and Z. Ma, "A model-based fault detection and diagnosis strategy for HVAC systems," *International Journal of Energy Research*, vol. 33, no. 10, pp. 903–918, 2009.

[15] T. Mulumba, A. Afshari, K. Yan, W. Shen, and L. K. Norford, "Robust model-based fault diagnosis for air handling units," *Energy and Buildings*, vol. 86, pp. 698–707, 2015.

[16] S. Gharsellaoui, M. Mansouri, M. Trabelsi, M.-F. Harkat, S. S. Refaat, and H. Messaoud, "Interval-valued features based machine learning technique for fault detection and diagnosis of uncertain HVAC systems," *IEEE Access*, vol. 8, pp. 171 892–171 902, 2020.

[17] S. Frank, M. Heaney, X. Jin, J. Robertson, H. Cheung, R. Elmore, and G. Henze, "Hybrid model-based and data-driven fault detection and diagnostics for commercial buildings," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2016.

[18] I. Matetić, I. Štajduhar, I. Wolf, and S. Ljubic, "A review of data-driven approaches and techniques for fault detection and diagnosis in HVAC systems," *Sensors*, vol. 23, no. 1, pp. 1–37, 2022.

[19] R. J. P. Silvio Simani, Cesare Fantuzzi, "Model-based fault diagnosis in dynamic systems using identification techniques," 2003.

[20] Y. Wang, X. Li, and J. Wang, "A neurodynamic optimization approach to supervised feature selection via fractional programming," *Neural Networks*, vol. 136, pp. 194–206, 2021.

[21] H. Li and J. Wang, "Collaborative annealing power k-means++ clustering," *Knowledge-Based Systems*, vol. 255, p. 109593, 2022.

[22] ——, "CAPKM++ 2.0: An upgraded version of the collaborative annealing power k-means++ clustering algorithm," *Knowledge-Based Systems*, p. 110241, 2023.

[23] ——, "Capacitated clustering via majorization-minimization and collaborative neurodynamic optimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, 2023, in press.

[24] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[25] S. M. Ross, *Introduction to probability and statistics for engineers and scientists*. Academic press, 2020.

[26] B. Jiang, C. Zhang, Y. Zhong, Y. Liu, Y. Zhang, X. Wu, and W. Sheng, "Adaptive collaborative fusion for multi-view semi-supervised classification," *Information Fusion*, vol. 96, pp. 37–50, 2023.

[27] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] V. O. Khilwani, V. Gondaliya, S. Patel, J. Hemnani, B. Gandhi, and S. K. Bharti, "Diabetes prediction, using stacking classifier," in *International Conference on Artificial Intelligence and Machine Vision*. IEEE, 2021, pp. 374–379.

[31] L. Guo, S. Wang, and Z. Cao, "An ensemble classifier based on stacked generalization for predicting membrane protein types," in *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*. IEEE, 2017, pp. 1–6.

[32] S. S. Ahmadi and H. Khotanlou, "A hybrid of inference and stacked classifiers to indoor scenes classification of RGB-D images," in *International Conference on Machine Vision and Image Processing*. IEEE, 2022, pp. 1–6.

[33] E. S. Lee, "Exploring the performance of stacking classifier to predict depression among the elderly," in *IEEE International Conference on Healthcare Informatics*. IEEE, 2017, pp. 13–20.

[34] K. Leartpantulak and Y. Kitjaidure, "Music genre classification of audio signals using particle swarm optimization and stacking ensemble," in *7th International Electrical Engineering Congress*. IEEE, 2019, pp. 1–4.

[35] C. Sheng and H. Yu, "An optimized prediction algorithm based on XGBoost," in *International Conference on Networking and Network Applications*. IEEE, 2022, pp. 442–447.

[36] G. D. Kumar, V. Deepa, N. Vineela, and G. Emmanuel, "Detection of parkinson's disease using LightGBM classifier," in *6th International Conference on Computing Methodologies and Communication*. IEEE, 2022, pp. 1292–1297.

[37] J. Wen and S. Li, "Tools for evaluating fault detection and diagnostic methods for air-handling units," ASHRAE Research Project 1312 Final Report, Atlanta, GA: ASHRAE, Tech. Rep., 2011.

[38] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 135, no. 3, pp. 370–384, 1972.

[39] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
[40] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM New York, 2001, pp. 41–46.
[41] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.