IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 November 2023, accepted 12 December 2023, date of publication 14 December 2023,
date of current version 11 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3342915

RESEARCH ARTICLE

MGRBA: Gas Recognition With Biclustering and Adaboost

RUN ZHOU[1], JIANHAO WANG[1], KANG HUANG[1], HUI WANG[2], AND ZHIJUN CHEN[1]
[1]School of Mechanical and Electrical Information, Yiwu Industrial and Commercial College, Jinhua 322000, China
[2]School of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321000, China

Corresponding author: Run Zhou (zhourun@mail.nwpu.edu.cn)

This work was supported in part by the General Project of Zhejiang Provincial Department of Education: Research on Multi-User Multiple Access Protocol for QoS in the Next Generation WLAN under Project 2021JYTYB02, in part by the Zhejiang Provincial Department of Education 2023 "Domestic University Visits" Project, and in part by the 2021 Zhejiang Province Industry-Academia Cooperation Collaborative Education Project "Practice and Innovation of 'Big Data+X' Talent Cultivation under the Digital Economy" under Grant 101.

ABSTRACT Gas recognition has been widely used in many fields such as air quality monitoring in dangerous areas. However, existing recognition methods suffer from two limitations: first, the recognition accuracy is not high. Due to the stochastic nature of air turbulence, gas features are not steady. The global features are sensitive to feature variations. Existing methods are based on global similarity, ignoring local similarity. Samples may be dissimilar in respect of global similarity, but are similar in terms of local similarity; Second, most existing recognition methods are based on the closed-set assumption that the gases categories in the train and test set are same. However, in real world applications, the test set may have non-overlapping gas category with the train set. To address above limitations, biclustering is used to extract local similarity. However, original biclustering method is not suitable for extraction. Since original biclustering method is used to find all kinds of biclusters, here we just want to find column nearly constant bicluster Therefore, a modified biclustering method is proposed. The local similarity can be used to construct classifier to recognize gas with adaboost. However, original adaboost cannot be used for open-set recognition. Thus a modified adaboost that uses two thresholds is proposed to recognize the unknown gases. To assess the efficacy of the proposed method, it is tested on public dataset. Experiment results demonstrate that the proposed method outperforms several state-of-the-art methods in respect of several evaluation measures on both closed-set and open-set cases.

INDEX TERMS Gas sensors, open-set recognition, local similarity, biclustering, ensemble learning, adaboost.

I. INTRODUCTION

Gas recognition [1], [2], [3] has been widely employed in many fields such as gas leakage detection [4], food recognition [5] and so on. Gas recognition can be divided into static gas recognition and mobile gas recognition [6]. The main difference between mobile gas recognition and static recognition is that the recognition system hardware is installed on mobile platform instead of being fixed. Mobile gas recognition system can be used in larger space and are more convenient. Compared with static gas recognition, mobile gas recognition is more important.

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

Mobile gas recognition system is composed of four parts: mobile platform, gas sensor array, feature extraction and recognition. The mobile platform can be one people, a robot or other mobile devices. There are gas sensor array and other edge computing devices on the mobile platform. When gas meets gas sensor array, gas responses such as voltages or currents can be obtained. Then gas features can be extracted from gas responses. Finally, with recognition methods, the category of gas can be predicted. The recognition methods such as SVM, CNN and Transform are based on global similarity between samples. In this study, a novel recognition method based on local similarity is proposed. Recognition is the core part of gas recognition system, the focus of the methods in the literature mainly lies in recognition. When

gases meet the mobile gas recognition system, gas label can be predicted.

Many gas recognition methods have been proposed in the literature. These methods can be grouped into two categories: deep learning based recognition method and conventional machine learning based recognition method. Deep learning based methods usually transform gas sample to be an matrix and then deals with the matrix with deep learning methods [2], [7], [8]. Conventional machine learning based method uses support vector machine, K-Nearest neighbor, etc [9], [10] to recognize gas. These methods achieved success to some extents. However, recognition accuracy is not satisfying. In a review of existing gas recognition methods, it can be found that these recognition methods are based on gas ***global similarity(feature)*** instead of gas ***local similarity(feature)***. Global feature means the whole features of the gas feature vector. Local feature denotes partial features of the gas feature vector. To improve recognition accuracy, we propose a recognition model based on gas local feature. Our model is guided by answering following questions:

(1).Why and how to use local features for mobile gas recognition?

Because gas feature is produced by gas sensor's response, due to the stochastic nature of air turbulence, sensor responses will change if the gas contact with the sensor unsteadily [12], [13]. The final step of many existing recognition methods is sorting the membership percentages of each class, finally assign the class label that has the biggest percentage to the test sample. When calculating the membership percentage, existing methods take the whole gas features into consideration, ignoring much local information. The loss of local information may reduce recognition accuracy. Global features may be redundant and may contain misleading information and fail to capture finer semantic details [14], [15]. Therefore, in this study local feature is used for recognition.

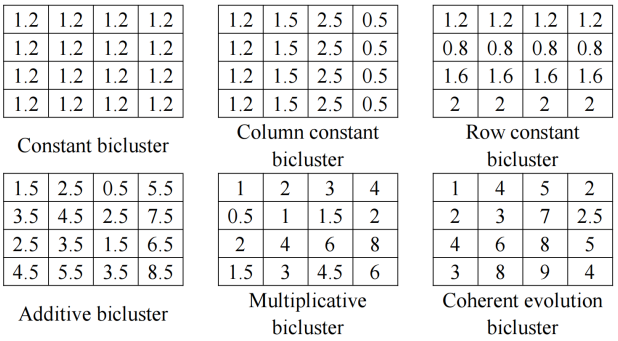

FIGURE 1. Bicluster type.

Biclustering [16], [17] can find much hidden local information ignored by traditional clustering, we plan to extract local information(local partial features) with biclustering. Biclustering can simultaneously cluster from both row(sample) dimension and column(feature) dimension, generating many biclusters. The samples in biclusters are similar with respect to partial features instead of the whole features. One sample can be grouped into multiple biclusters. With biclustering, much local information can be extracted to improve recognition accuracy. Biclustering has been used in many other fields such as recommendation system [18], missing value prediction [19] and gene expression data analysis [20]. Typical methods include greedy search based methods [20], and model based methods [21]. As shown in Fig. 1, bicluster can be divided into six types [22]: constant bicluster, column constant bicluster, row constant bicluster, additive bicluster, multiplicative bicluster and coherent evolution bicluster. In this context, only column constant bicluster is used because the samples in bicluster are similar in respect of partial features. However, most existing biclustering methods are designed for finding all kinds of biclusters. Therefore, a novel column constant bicluster mining method is proposed in this study. To the best of our knowledge, biclustering is the first time to be used in gas recognition.

Biclusters cannot be used to recognize gas directly. This study intends to recognize gas with bicluster and adaboost [23]. Biclusters can be converted to many weak classifiers. Considering that ensemble learning can combine multiple base classifiers to yield higher recognition accuracy than single classifier, we plan to combine bicluster and adaboost to achieve high accuracy. Adaboost is an ensemble learning method that can recognize gas through combining the recognition results of team members(two or more other methods) to improve recognition accuracy. With adaboost, the weak classifiers built by biclusters can be combined to construct strong classifier to recognize gas.

(2).In real application, unknown gases may coexist with known gases, how to recognize the known gases and reject the unknown gases?

Existing methods are based on the closed-set assumption that the gas categories in test set completely overlap with that in the train set. However, in real applications the test set may have non-overlapping gases with the train set. For example, if there are only two kinds of gases (carbon dioxide and oxygen) in the train dataset, in real application in the air there must be nitrogen which does not exist in the train dataset. Recognizing the new unknown gas as one of the old known gases will greatly reduce recognition accuracy. Even though open-set gas recognition has seldomly been studied, in other internet of things fields related works have been reported [24], [25]. Open-set recognition [26], [27] is different from few-shot learning and zero-shot learning [28], open-set recognition needs to recognize the known gases and reject the unknown gases simultaneously. Suppose in train set there are m kinds of gases, in test set there are $n(n>m)$ kinds of gases. Gases $1, 2, \ldots, m-1, m$ are deemed as known gases, and gases $m+1, m+2, \ldots n-1, n$ are deemed as unknown gases. Open-set recognition method should not only accurately recognize the m known gases, but also reject the $n-m$ unknown gases. In the literature, the open-set recognition in computer vision field is based on the similarity

between test/train sample. If the biggest similarity is greater than predefined threshold, the test sample is recognized as known gas, otherwise rejected as unknown gas. Such strategy effectively differentiates the known and unknown gases. In this study, a modified adaboost is proposed for open-set gas recognition. As shown in Eq. 1, the principle of adaboost is as follows: if the value of $SC(V_{SC})$ is greater(less) than zero, the test sample is assigned the label of $+1(-1)$. The larger absolute V_{SC} is, we have more confidence in that the predicted label is correct. In the literature the threshold for determing sample's label is 0, we think that 0 may not be the optimal threshold, the optimal threshold should be changed. Since adaboost is designed for binary classification, two thresholds(positive threshold T_+ and negative threshold T_-) should be designed. If V_{SC} is larger than T_+, corresponding sample is more similar with positive sample, it is assigned positive label, if V_{SC} is less than T_-, the sample is more similar with negative sample, it is assigned negative label. If $T_- \leq V_{SC} \leq T_+$, the sample should be assigned neither positive nor negative label, it is assigned the "unknown label".

$$SC = \sum_{i=1}^{k} w_i * WC_i \tag{1}$$

where k denotes the number of weak classifiers, WC_i denotes the ith weak classifier, w_i is the weight of WC_i.

In the proposed method, specifically, firstly the train dataset is divided into many binary subsets that contains only two kinds of gases. Then many biclusters are mined from each subset to construct the binary strong classifier, each binary classifier can output a label(negative, positive or unknown). Combining the labels of the whole binary classifiers, the support of each label can be obtained. Finally, majority voting is adopted to the labels, the label that has biggest support is assigned to the test sample. The contribution of the proposed Gas Recognition with Biclustering and Adaboost(MGRBA) method contains three parts:

- We early notice that due to the stochastic nature of air turbulence, mobile gas features are not steady. Existing methods are based on global similarity and are sensitive to feature variation. Existing methods ignore local information and are hard to achieve high recognition accuracy. Besides, we early notice that the problem of open-set mobile gas recognition has been seldomly investigated in the literature.
- Motivated by that local similarity is insensitive to feature variation, a biclustering and modified adaboost based method is proposed to improve recognition accuracy. A novel column nearly constant biclustering method is proposed. The core motivation is that biclustering can mine local features to construct weak classifiers, adaboost can combine the weak classifiers to construct strong classifiers to boost recognition accuracy. To solve the open-set recognition problem, a modified adaboost algorithm where two new thresholds T_- and T_+ instead of single 0 are adopted to discriminate the known and unknown gases is proposed.
- We implement the MGRBA framework and apply it to gas dataset on embedded platform to validate its performance. Experiment results demonstrate its superiority to state-of-the-art methods in terms of three evaluation measures on both closed-set and open-set cases.

The remaining parts are organized as follows: Section II describes the related work, motivation and approach overview are given in Section III and IV, the detailed description of the proposed MGRBA method is shown in Section V, experiment results are displayed in Section VI, conclusion are drawn in Section VII.

II. RELATED WORK

In this section, related work is described. Firstly, closed-set gas recognition methods are presented. Subsequently, open-set recognition in other fields are given.

A. CLOSED-SET GAS RECOGNITION

Since in the literature, seldomly open-set gas recognition is reported, only closed-set gas recognition related methods are introduced. Closed-set gas recognition methods can be grouped into three categories, namely classical machine learning based gas recognition methods, Artificial neural network(ANN) based gas recognition methods and Spiking neural network(SNN) based gas recognition methods. Classical machine learning based methods include LDA(linear discriminant analysis), DT(Decision tree), SVM(Support Vector Machine), NB(Naive Bayesian) etc. ANN based methods firstly transform each gas sample feature vector into a gray image(row denotes sensor, column denote feature), then deal with image with convolutional layer, pooling layer, fully connected layer and softmax layer. Spiking neural network (SNN) is proposed as a biologically realistic algorithm with lower computing cost and is arguably the only viable option at the neuronal description level. Towards linking biologically plausible learning methods and better recognition capability, a number of SNN-based gas recognition algorithms have been developed [29]. Besides, recognition with non sensor-based method also has been proposed. For example, in [30] ultrasound is applied to identify gases. Its recognition principle is that when different gases meet the same ultrasonic, different ultrasonic effects can be produced.

B. OPEN-SET RECOGNITION

Open-set recognition can also be divided into traditional machine learning based methods) [31] and deep learning based methods [32], [33]. For example, traditional machine learning based method CV(Class Verification) [31] stores all training samples, compares the distances that are determined in pairs of the selected test sample and each training sample, sorts the distances, selects two nearest neighbors. If the two

nearest neighbors share same class label, corresponding label is assigned to test sample. If two nearest neighbors have different class labels, test sample is assigned as unknown label. CV can be deemed as the extension of conventional k-Nearest-Neighbor(KNN) classification method. Deep learning based method ST(Softmax Threshold) [32]. In ST, a threshold is added to the maximal probability of the softmax layer. If maximal probability is smaller than the threshold, corresponding sample is assigned the unknown label, otherwise it is assigned the known label. The limitation of these methods lies in that they take all features into account when calculating sample similarity, ignoring local information.

	F_1	F_2	...	F.	F.	F.	...	F_{128}	Label
S1	1.1	0.7	...	3.2	1.5	0.1	...	0.9	1
S2	1.2	1.8	...	3.1	2.2	1.3	...	0.8	-1
S3	0.2	1.7	...	3.3	0.7	0.2	...	0.8	-1
S4	2.4	1.8	...	3.2	1.4	2.1	...	0.7	1
S5	1.1	1.7	...	3.3	2.1	1.4	...	1	-1

FIGURE 2. An example for illustrating local similarity.

III. MOTIVATION

In this part, we conduct two experiments on a public real-world gas dataset(batch 10 of GSADD [34]). Two observations are analyzed to motivate the approach of MGRBA.

A. OBSERVATION 1: IMPROVE RECOGNITION ACCURACY WITH LOCAL SIMILARITY

To explain the use of local similarity instead of global similarity for gas recognition, closed-set experiment is conducted. Through observing the features of samples, it can be found that some samples are different in terms of global features, but alike in terms of partial features. For example, as shown in Fig. 2, five gas samples are selected for illustration. In Fig. 2, row denotes sample, column denotes feature and label. Sample 2, 3 and 5 share identical label. Sample 2 and sample 5 are similar with regards to the whole features because their euclidean distance is small. Sample 2 and 3 are not similar with regards to the whole features. However, they are similar in terms of the three green-marked features(local features). Therefore, local similarity cannot be ignored, biclustering should be used to calculate local similarity in this study. To some extents, whole features can be deemed as the subset of local features.

B. OBSERVATION 2: MODIFY THE THRESHOLD OF ADABOOST FOR OPEN-SET RECOGNITION

To show the necessity of modifying adaboost for open-set gas recognition, close-set experiment is conducted on For ease of presentation, only two kinds of gases, Ethanol and Ethylene, are used, adaboost is used to recognize the two gases. 80% gases are used as train dataset, the remaining gases are used as test dataset. Record the outputs of the binary strong classifier, the minimum of positive outputs is 0.039 and the maximum of negative outputs is -0.039. Among the 240 test gas samples, 39 gas samples are wrongly recognized. Among the 39 wrongly recognized samples, 20 of the 39 outputs fall in the range of [-0.039 0.039]. This result demonstrates that the larger the absolute value of strong binary classifier output, the more possible the recognization result is.

Motivated by this result, we propose the modified adaboost that incorporates positive threshold T_+ and negative threshold T_- to improve recognition accuracy. If the output falls in the range of $[T_-, T_+]$, the sample may be unseen gas.

IV. APPROACH OVERVIEW

A. SYSTEM ARCHITECTURE

System architecture of MGRBA is displayed in Fig. 3. Firstly, biclustering is applied to dataset to mine biclusters. Subsequently, employ column average to the biclusters to generate discrimination rules. Then the weak classifiers can be built by combining the discrimination rules in pair. Finally, with adaboost the weak classifiers can construct a strong classifier to recognize gas.

B. APPLICATION SCENARIOS

MGRBA can be used in many scenarios where gases need recognizing. For example, the common gas categories in the air are known, a soldier can use this device to detect whether the enemy released unknown poison gases on the battlefield.

V. METHOD

In this section, the three steps of MGRBA preprocessing, training of MGRBA and test of MGRBA are introduced in detail.

A. PREPROCESSING

Suppose there are C_{tr} kinds of gases in train dataset M_{tr}, and the number of gas samples is N_{tr}, there are $C_{ts}(C_{tr} \leq C_{ts})$ kinds of gases in the test dataset M_{ts}, the number of gas samples is N_{ts}.

Because different features are in different scales, in order to eliminate the scale influence between features, max-min normalization is employed to each column of M_{tr} to map all feature values in the range of [0, 1].

As discussed above, multiple classifier is divided into the combination of many binary classifiers. In the train dataset, there are C_{tr} kinds of gases. The normalized train dataset M_{tr}^n is divided into C_{tr} subsets according to gas sample label. Then combine the C_{tr} subsets in pairs to construct $N_{bd}(N_{bd} = \frac{C_{tr}*(C_{tr}-1)}{2})$ new datasets D. In each new dataset, there are only two kinds of gas samples. Following step is to mine binary classifier from each new dataset $D_i(i \in 1, 2, 3, \ldots, N_{bd} - 1, N_{bd})$.

B. TRAINNING OF MGRBA

1) BICLUSTER MINING

The frequent used bicluster quality measure is MSR(Mean Square Residule) [20] that is used to mine additive bicluster,

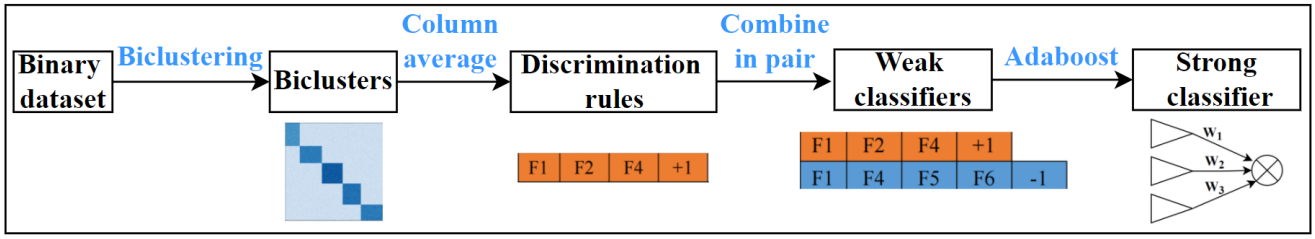

FIGURE 3. Architecture of MGRBA.

since this study aims to mine column nearly constant bicluster, MMSR(modified MSR) is adopted. The modification lies in that mean column variance is added to original MSR. The mining of biclusters contains two parts: finding similar clusters from each column and expanding the clusters. These clusters are found with Gap [11]. The numbre of clusters can be automatically determined. The elements in the mined clusters are similiar, these clusters can be deemed as bicluster seeds. After finding bicluster seeds, the size of the bicluster is enlarged by adding remaining row/column to the clusters iteratively untill MMSR of biclustr is less than predefined threshold δ_B.

$$MSR(B) = \frac{\sum_{i\in R, j\in C}(b_{ij} - b_{iC} - b_{Rj} + b_{RC})^2}{|R| * |C|} \tag{2}$$

$$MMSR(B) = MSR(B) + \frac{1}{n}\sum_{i=1}^{|C|} var(B(:, i)) \tag{3}$$

$$b_{iC} = \frac{\sum\limits_{j\in C} b_{ij}}{|C|}, b_{Rj} = \frac{\sum\limits_{i\in R} b_{ij}}{|R|}, b_{RC} = \frac{\sum\limits_{i\in R, j\in C} b_{ij}}{|R| * |C|} \tag{4}$$

where b_{iC} is the average of the ith row in B, b_{Rj} is the average of the jth column in B, b_{RC} is the average of the whole elements in B.

2) RULE CONSTRUCTION

Having mined many biclusters from above step, following step is to transform these biclusters into discriminative rules. A rule is composed of two parts: features(precondition) and label(postcondition). The rule's format is like this: if f_1 is 0.15, f_5 is 0.25, f_{47} is 0.75, f_{100} is 0.05, then sample label is 2. The precondition is constructed by averaging the columns of bicluster, the post condition is constructed with majority voting.

3) WEAK CLASSIFIER CONSTRUCTION

From above rule construction step, two kinds of rules can be obtained. In this step, the rules mined from above step are combined in pair to construct many weak classifiers whose classification accuracy is not high. If there are p positive rules and q negative rules, combination in pair can construct $p * q$ weak classifiers. The principle of the weak classifier is that if the euclidean distance between test sample and the positive features is smaller than the euclidean distance between test sample and the negative features in weak classifier, the test sample is set the positive label, else set the negative label. It is noted that when calculating the euclidean distance, only partial features of the test sample is considered.

The working principle of weak classifier is shown in Fig.4. If D_p, the distance between test sample and the positive rule, is smaller than D_n, positive label is assigned to test sample, else negative label is assigned to test sample.

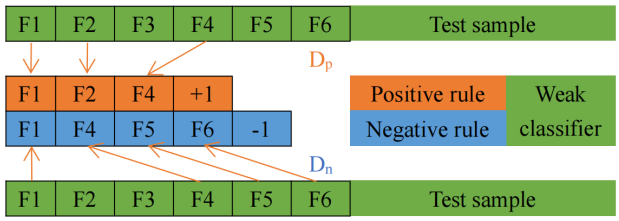

FIGURE 4. Illustration of weak classifier's working principle.

4) STRONG CLASSIFIER CONSTRUCTION

Generally, the classification accuracy of the weak classifiers is not high. To boost accuracy, in this step, adaboost is employed to combine these weak classifiers to build a strong binary classifier SC_i with Eq. 1.

For closed-set gas recognition, SC works as Eq.5. The threshold is 0.

$$y = \begin{Bmatrix} -1 \ if \quad SC \le 0 \\ +1 \ if \quad SC \ge 0 \end{Bmatrix} \tag{5}$$

As described above, for the open-set gas recognition, each binary strong classifier should have two thresholds, T_- and T_+. Input each train sample to binary strong classifier, then separate the output values of binary classifiers into two parts: negative part and positive part. Set the minimum of positive part as T_+ and set the maximum of negative part as T_-. SC works as Eq.6.

$$y = \begin{Bmatrix} -1 & if & SC \le T- \\ +1 & if & SC \ge T_+ \\ unknown & if & T- \le SC \le T+ \end{Bmatrix} \tag{6}$$

C. TEST OF MGRBA

From above training step, N_{bd} binary strong classifiers and $2 * N_{bd}$ thresholds have been available. The test dataset M_{ts}^n contains two parts: known gases and unknown gases. During

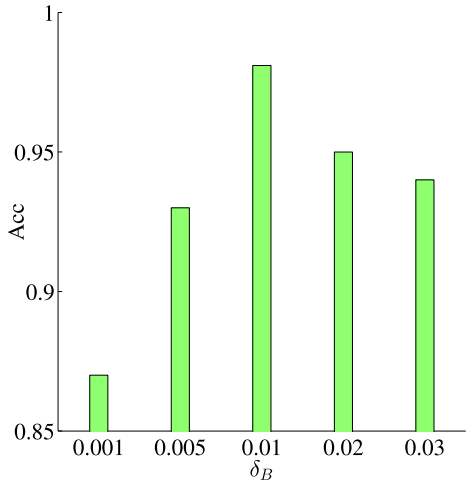

FIGURE 5. Recognition result of MGRBA with different values of δ_B.

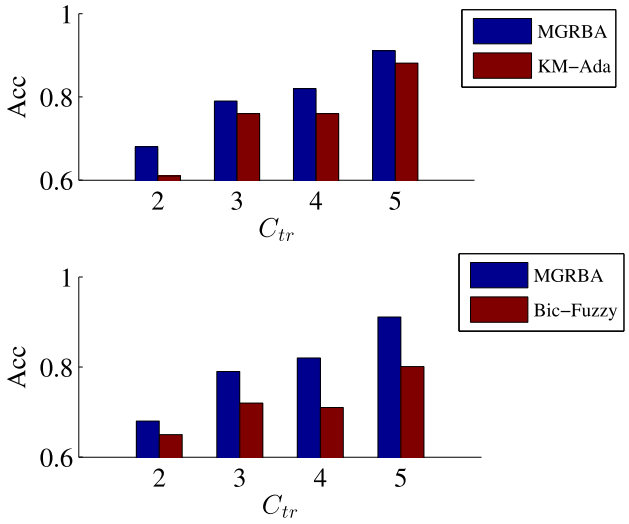

FIGURE 6. Ablation experiment result.

test phase, each gas sample in M_{ts}^n is input to each of the N_{bd} binary strong classifiers. Each binary strong classifier can output a value, if this value is larger(smaller) than $T_+(T_-)$, positive(negative) label is assigned to the sample, else unseen label is assigned to the sample. N_{bd} weak classifiers can output N_{bd} labels, finally majority voting [35], [36] is adopted to select the label that has the biggest support as the final label of the sample. For example, if there are three kinds of gases (labels are 1, 2 and 3) in train datasets, then 3 binary strong classifiers can be obtained. One binary classifier can output label 1, label 2 or label unknown. One binary classifier can output label 1, label 3 or label unknown. One binary classifier can output label 2, label 3 or label unknown. If a sample is input to the three binary classifiers, corresponding labels are unknown, 3 and 3, respectively, label 3 has maximal support, label 3 assigned to the sample. If a sample is input to the three binary classifiers, corresponding labels are unknown, 1 and unknown, respectively, label unknown has maximal support, label unknown is assigned to the sample.

D. SUMMARIZATION

During the train process, firstly the multiple-class dataset is divided into many two-class subsets, multiple classifier is divided into many binary classifiers. Subsequently, construct binary classifiers from each binary subset with following steps:biclustering, constructing discrimination rules, constructing weak classifiers and strong classifier. Finally, many binary strong classifiers each of which has two thresholds T_+ and T_- can be outputed. During test phase, each test gas sample is input to these binary strong classifiers, producing many outputs, each of which is compared with T_+ and T_- to generate many labels. Finally, majority voting is applied to the labels to generate final label(one of the seen gas or unseen gas). When finding biclustets, complexity is $O(n)$ When building binary strong classifiers, complexity is $O(n^2)$. Thus, the total complexity of MGRBA is $O(n^2)$.

VI. EVALUATION

In this section, the performance of MGRBA is evaluated over public dataset on one embedded platform under closed-set and open-set cases.

A. EVALUATION SETUP

1) DATASET

We quantitatively evaluate the performance of the proposed MGRBA on public gas dataset GSADD [34]. There are six kinds of gases: Ethanol, Ethylene, Ammonia, Acetaldehyde, Acetone and Toluene. GSADD is composed of ten batches, there are totally 13910 gas samples, the detailed information is shown in Fig.1. This gas dataset is collected with 16 metal-oxide gas sensors(4 TGS2600, 4 TGS2602,4 TGS2610 and 4 TGS2620). From each sensor's response(resistance fluctuation) curve, 8 features(2 steady-state features and 6 transient features) are extracted. Each gas sample is represented by a 1*128 dimension vector.

2) EVALUATION MEASURE

In closed-set recognition the performance evaluation measures used are Acc(Accuracy) and confusion matrix [37]. Confusion matrix is a common measure for visually evaluating the performance of supervised learning algorithm, from confusion matrix the recognition statistics of each category can be seen in detail. The mathematical formulation of Acc is shown in Eq. 7.

$$Acc = \frac{N_{ck} + N_{cuk}}{N_k + N_{uk}} \tag{7}$$

where $N_k(N_{uk})$ denote the number of the whole known (unknown) gas samples in test dataset, $N_{ck}(N_{cuk})$ denote the number of the correctly recognized known(unknown) gas samples. The higher Acc is, the better algorithm's recognition ability is.

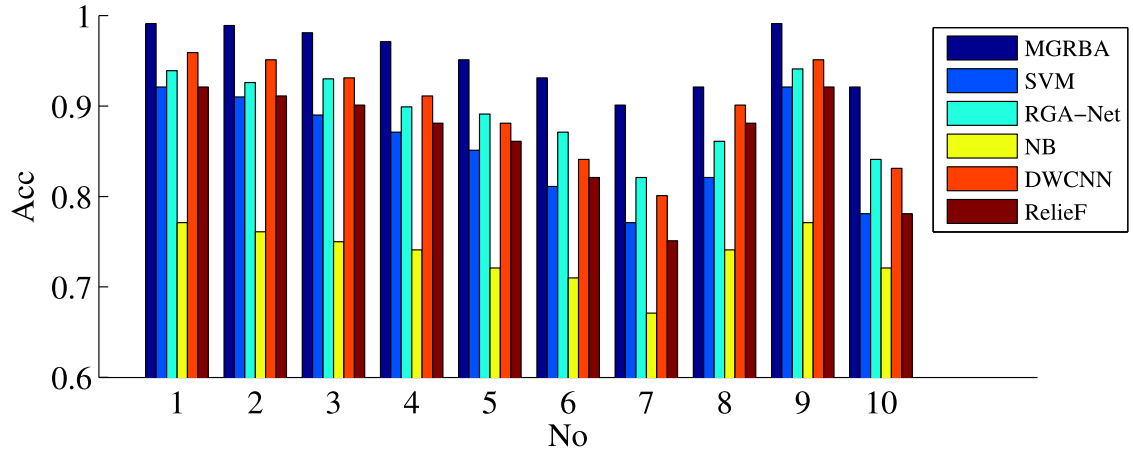

FIGURE 7. The closed-set recognition accuracy of the six closed-set classifiers on ten batches of GSADD.

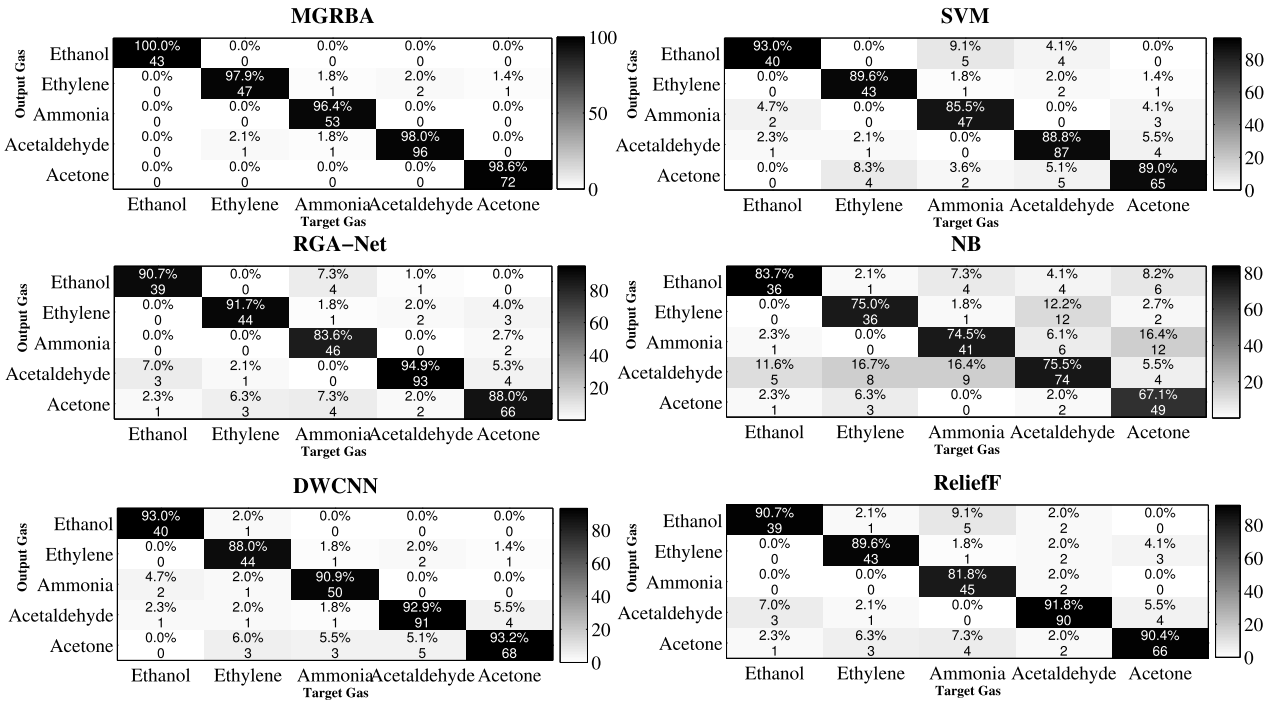

FIGURE 8. Confusion matrixes of six methods on closed-set case under batch 3.

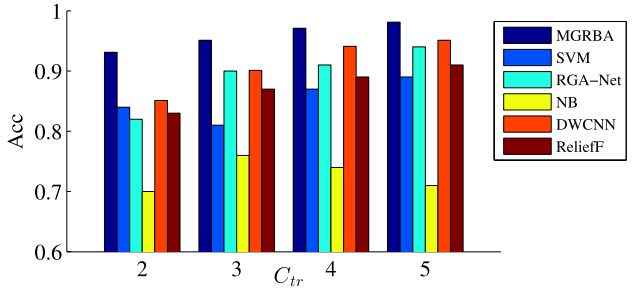

FIGURE 9. Closed-set gas recognition results under different C_{tr}.

For open-set recognition, two measures *Acc* and *AUROC* [37] are used. AUROC is a calibration-free measurement that is suitable for the open-set gas recognition. In the Receiver Operating Characteristic (ROC) curve, the horizontal coordinate of each point is the false positive rate (FPR, which recognize the known classes as the unknown classes), and the vertical coordinate is the true positive rate (TPR, which recognize the known classes as the known classes).

3) MODELS IN COMPARISON

To demonstrate the superiority of MGRBA, it is compared with several alternative state-of-the-art methods in closed-set and open-set recognition fields. For closed-set recognition, MGRBA is compared with three methods: SVM(Support Vector Machine) [38], RGA-Net [39], NB(Naive Bayesian) [40], DWCNN(Dynamic Wavelet Convolutional Neural Network) [41], and ReliefF [42]. For open-set recognition, MGRBA is compared with the following methods: CV (Class

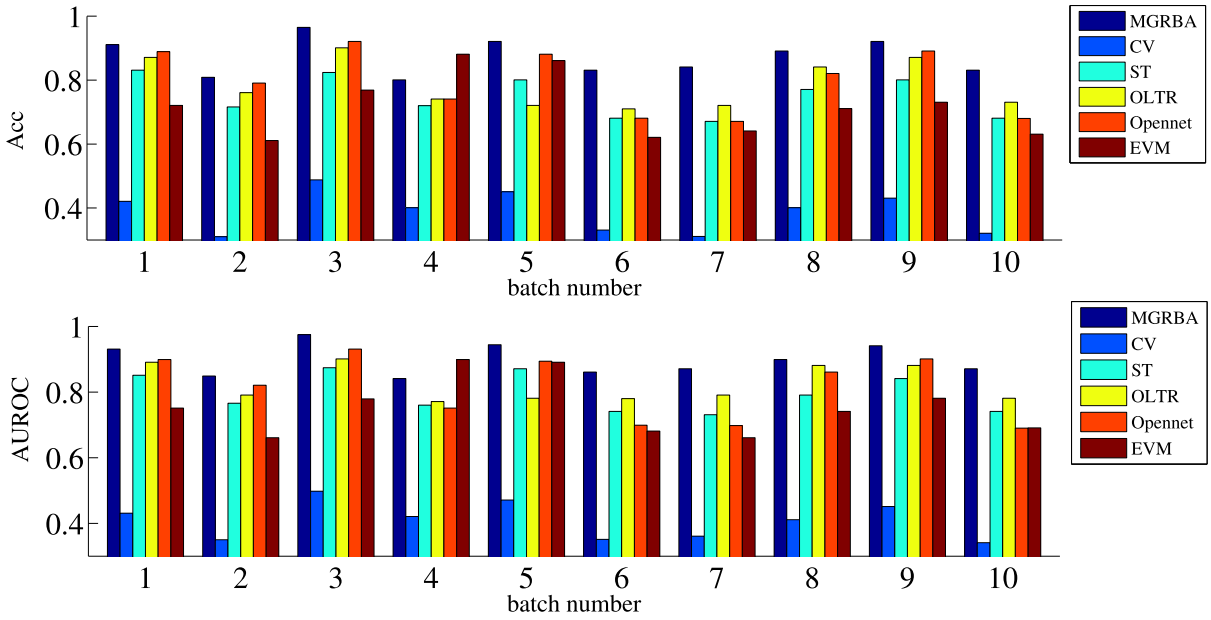

FIGURE 10. *Acc* and *AUROC* of six methods on open-set case.

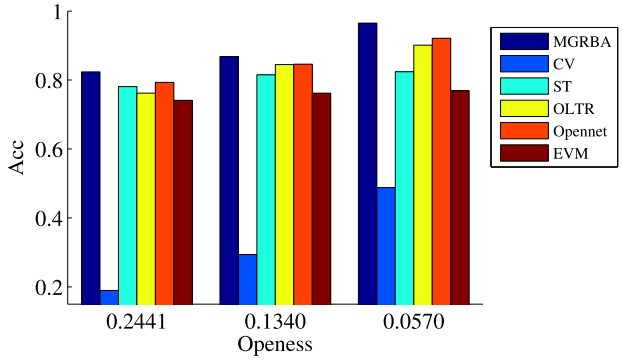

FIGURE 11. Recognition results of six methods on batch 3 under different *Openess*.

Verification) [31], OLTR(Open Long-Tailed Recognition) [32] and ST(Softmax Threshold) [33], Opennet [43], EVM [44]. All comparison methods recognize gas with global similarity.

B. PARAMETER IMPACT

To determine the optimal values of parameters, grid search is used on closed-set recognition over batch 3 of GSADD dataset. For δ_B, five values(0.001, 0.005, 0.01, 0.02 and 0.03) are empirically set. Different recognition results of MGRBA with different δ_B is shown in Fig. 5. It can be found that 0.01 can generate highest recognition accuracy, therefore δ_B is set as 0.01. The parameter values of the comparison methods are set as default in the publications.

C. ABLATION STUDY

Since MGRBA is composed of biclustering and adaboost, it is necessary to investigate the contribution of biclustering and adaboost separately. Ablation study is conducted on closed-set recognition over batch 3 of GSADD dataset.

1) IMPACT OF LOCAL SIMILARITY

To demonstrate the superiority of local similarity, MGRBA is compared with k-means+adaboost(KM-Ada) under four closed-set cases. The difference between MGRBA and K-Ada only lies in that biclustering is replaced by k-means to extract global similarity. The comparison result is shown in the above subfigure of Fig.6. On the whole cases, recognition accuracy of MGRBA is higher than that of KM-Ada, demonstrating the positive contribution of biclustering based local similarity.

2) IMPACT OF ADABOOST

Similarly, to investigate the contribution of adaboost, comparison experiment between MGRBA and bicluster+fuzzy(Bic-Fuzzy) under four closed-set cases. The result is shown in the below subfigure of Fig.6. It can be found that MGRBA greatly outperforms Bic-Fuzzy on all cases, the positive contribution of adaboost is verified.

D. CLOSED-SET GAS RECOGNITION

For closed-set gas recognition, following experiments are conducted: (1) Test on each of the ten batches. From each batch, we randomly select 80% of each kind of gas samples as train dataset, the remaining 20% are used for constructing the test dataset. Experiment results(*ACC*) is displayed in Fig. 7. It can be seen that MGRBA outperform SVM, LeNet-5, NB, DWCNN and Relief in respect of *ACC*, improving 11% on average. The five comparison methods are based on global similarity, from the experiment results the advantage

TABLE 1. Detailed information of GSADD dataset.

No	Ammonia	Acetaldehyde	Acetone	Ethylene	Ethanol	Toluene	Total
1	83	30	70	98	90	74	445
2	100	109	532	334	164	5	1244
3	216	240	275	490	365	0	1586
4	12	30	12	43	64	0	161
5	20	46	63	40	28	0	197
6	110	29	606	574	514	467	2300
7	360	774	630	662	649	568	3613
8	40	33	143	30	30	18	294
9	100	75	78	55	61	101	470
10	600	600	600	600	600	600	3600

of biclustering based local similarity is demonstrated. The confusion matrixes of the six methods on batch 3 are shown in Fig. 8. We can see that for the five kinds of gases, MGRBA achieves higher accuracy than five comparison methods. (2)Test the effect of the number of gas kinds(C_{tr}) on batch 3 of GASDD. The experiment result is shown in Fig. 9. We can see MGRBA outperforms the five comparison methods on all cases, demonstrating the robustness of MGRBA.

E. OPEN-SET GAS RECOGNITION

Openess is a measure for evaluating the difficulty in open-set recognition field [45]. The definition of *Openess* is shown in Eq. 8. The larger *Openess* is, the less information is contained in the train dataset, the more difficult it is to recognize the unknown gases.

$$Openess = 1 - \sqrt{\frac{2 * C_{tr}}{C_{tr} + C_{ts}}} \tag{8}$$

Fig. 10 shows the *Acc* and *AUROC* on ten batches of GSADD when there is only one unknown gas category in test dataset. *Openess* of batch 3,4,5 is $0.0572(1 - \sqrt{(2 * 4)/(4 + 5)})$ and *Openess* of the remaining seven batches is 0.0465. From Fig. 10 it can be found that on ten batches MGRBA outperforms CV, ST, OLTR, Opennet and EVM, improving over 10%(*ACC*) in terms of both *ACC* and *AUROC*. To investigate the performance of MGRBA under different *Openess*, open-set experiment on batch 3 is conducted. C_{tr}, C_{ts} and corresponding *openess* is shown in Table. 2. Similar with closed-set recognition, 80% of the known samples are selected to construct train dataset, remaining 20% known samples and the whole unknown samples are combined to construct the test dataset. The open-set gas recognition results of the six methods under three openesses are displayed in Fig. 11. MGRBA outperforms the other five methods under all openesses. With the increase of openess, the accuracy of all methods decrease.

Take the following case as an example to describe MGRBA in detail: training subset contains two kinds of gases(Ethanol and Ethylene), test dataset contains the whole five kinds of gases. 32 biclusters are outputed. When constructing classification rules with column averaging and majority voting, six rules are deleted because their label confidence is less than predefined threshold. Finally, seven Ethylene rules are found, 19 Ethanol rules are found. Two thresholds of adaboost classifier are 0.0524(T_+) and -0.0524(T_-).

TABLE 2. Openess under different C_{tr} and C_{ts}.

No.	Ctr	Cts	Openness
1	4	5	0.057
2	3	5	0.134
3	2	5	0.2441

VII. CONCLUSION

In this paper, we early notice the open-set gas recognition problem and propose a local similarity based ensemble learning method MGRBA. Through reviewing the literature, we find that existing gas recognition methods take the whole features into account when calculating sample similarity, ignoring local information. We extract partial features with biclustering. We find adaboost's limitation that 0 is not the optimal threshold for determining sample's label. We propose a modified adaboost method with two thresholds. Combing biclustering and adaboost can accurately determine the label of gas sample. Experiment results demonstrate that MGRBA outperforms many state-of-the-art methods on both closed-set and open-set gas recognition cases.

In this study, MGRBA is investigated in gas datasets, we have confidence in that MGRBA works well in other fields. One future work is to apply MGRBA to other fields such as human activity recognition [46] to demonstrate its generalization performance. Besides, the gas sensor drift problem is a significant problem, MGRBA will be improved to solve gas sensor drift problem in the future.

REFERENCES

[1] I.-S. Chang, S.-W. Byun, T.-B. Lim, and G.-M. Park, "A study of drift effect in a popular metal oxide sensor and gas recognition using public gas datasets," *IEEE Access*, vol. 11, pp. 26383–26392, 2023, doi: 10.1109/ACCESS.2023.3257414.

[2] X. Pan, H. Zhang, W. Ye, A. Bermak, and X. Zhao, "A fast and robust gas recognition algorithm based on hybrid convolutional and recurrent neural network," *IEEE Access*, vol. 7, pp. 100954–100963, 2019, doi: 10.1109/ACCESS.2019.2930804.

[3] X. Zhao, Z. Wen, X. Pan, W. Ye, and A. Bermak, "Mixture gases classification based on multi-label one-dimensional deep convolutional neural network," *IEEE Access*, vol. 7, pp. 12630–12637, 2019, doi: 10.1109/ACCESS.2019.2892754.

[4] D.-S. Lee, S.-W. Ban, M. Lee, and D.-D. Lee, "Micro gas sensor array with neural network for recognizing combustible leakage gases," *IEEE Sensors J.*, vol. 5, no. 3, pp. 530–536, Jun. 2005, doi: 10.1109/JSEN.2005.845186.

[5] B. Mahata, K. Dixit, S. Acharyya, P. Banerji, and P. K. Guha, "Adulterated fish recognition employing SnO_2 nanostructure-based chemiresistive sensor," *IEEE Sensors Lett.*, vol. 7, no. 8, pp. 1–4, Aug. 2023, doi: 10.1109/LSENS.2023.3300807.

[6] A. Loutfi and S. Coradeschi, "Odor recognition for intelligent systems," *IEEE Intell. Syst.*, vol. 23, no. 1, pp. 41–48, Jan./Feb. 2008, doi: 10.1109/MIS.2008.11.

[7] Y. Tian, J. Yan, Y. Zhang, T. Yu, P. Wang, D. Shi, and S. Duan, "A drift-compensating novel deep belief classification network to improve gas recognition of electronic noses," *IEEE Access*, vol. 8, pp. 121385–121397, 2020, doi: 10.1109/ACCESS.2020.3006729.

[8] C. Zhang, W. Wang, Y. Pan, and S. Zhai, "Curriculum learning-based approaches for end-to-end gas recognition," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3205901.

[9] W. Xia, T. Song, Z. Yan, K. Song, D. Chen, and Y. Chen, "A method for recognition of mixed gas composition based on PCA and KNN," in *Proc. 19th Int. Conf. Opt. Commun. Netw. (ICOCN)*, Qufu, China, Aug. 2021, pp. 1–3, doi: 10.1109/ICOCN53177.2021.9563663.

[10] A. Bermak and S. B. Belhouari, "Bayesian learning using Gaussian process for gas identification," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 3, pp. 787–792, Jun. 2006, doi: 10.1109/TIM.2006.873804.

[11] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, Jul. 2001.

[12] T. Yoshida, R. Takemura, H. Matsukura, and H. Ishida, "Determination of gas source existence in a specified area by active airflow generator robots," in *Proc. IEEE SENSORS*, Busan, South Korea, Nov. 2015, pp. 1–4, doi: 10.1109/ICSENS.2015.7370339.

[13] H. Ishida, Y. Wada, and H. Matsukura, "Chemical sensing in robotic applications: A review," *IEEE Sensors J.*, vol. 12, no. 11, pp. 3163–3173, Nov. 2012, doi: 10.1109/JSEN.2012.2208740.

[14] M. A. Alzubaidi, M. Otoom, and H. Jaradat, "Comprehensive and comparative global and local feature extraction framework for lung cancer detection using CT scan images," *IEEE Access*, vol. 9, pp. 158140–158154, 2021, doi: 10.1109/ACCESS.2021.3129597.

[15] D. Srivastava, S. S. Singh, B. Rajitha, M. Verma, M. Kaur, and H.-N. Lee, "Content-based image retrieval: A survey on local and global features selection, extraction, representation, and evaluation parameters," *IEEE Access*, vol. 11, pp. 95410–95431, 2023, doi: 10.1109/ACCESS.2023.3308911.

[16] S. Qian, H. Liu, X. Yuan, W. Wei, S. Chen, and H. Yan, "Row and column structure-based biclustering for gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 1117–1129, Mar. 2022, doi: 10.1109/TCBB.2020.3022085.

[17] M. Vandromme, J. Jacques, J. Taillard, L. Jourdan, and C. Dhaenens, "A biclustering method for heterogeneous and temporal medical data," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 506–518, Feb. 2022, doi: 10.1109/TKDE.2020.2983692.

[18] J. Sun and Y. Zhang, "Recommendation system with biclustering," *Big Data Mining Anal.*, vol. 5, no. 4, pp. 282–293, Dec. 2022, doi: 10.26599/BDMA.2022.9020012.

[19] F. Meng, C. Cai, and H. Yan, "A bicluster-based Bayesian principal component analysis method for microarray missing value estimation," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 3, pp. 863–871, May 2014, doi: 10.1109/JBHI.2013.2284795.

[20] Y. Cheng and G M. Church, "Biclustering of expression data," *Ismb*, vol. 8, pp. 93–103, Dec. 2000.

[21] S. Chen, L. Zhang, L. Lu, J. Meng, and H. Liu, "FBCwPlaid: A functional biclustering analysis of epi-transcriptome profiling data via a weighted plaid model," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 3, pp. 1640–1650, May 2022, doi: 10.1109/TCBB.2021.3049366.

[22] Q. Huang, D. Tao, X. Li, and A. Liew, "Parallelized evolutionary learning for detection of biclusters in gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 2, pp. 560–570, Mar. 2012, doi: 10.1109/TCBB.2011.53.

[23] E. Lodhi, F.-Y. Wang, G. Xiong, A. Dilawar, T. S. Tamir, and H. Ali, "An AdaBoost ensemble model for fault detection and classification in photovoltaic arrays," *IEEE J. Radio Freq. Identificat.*, vol. 6, pp. 794–800, 2022, doi: 10.1109/JRFID.2022.3212310.

[24] Z. Xia, P. Wang, G. Dong, and H. Liu, "Surrounding prototype loss for radar HRRP open set target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4027305, doi: 10.1109/LGRS.2022.3211623.

[25] G. Bhavanasi, L. Werthen-Brabants, T. Dhaene, and I. Couckuyt, "Open-set patient activity recognition with radar sensors and deep learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 3501005, doi: 10.1109/LGRS.2023.3235243.

[26] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013, doi: 10.1109/TPAMI.2012.256.

[27] J. Huang, D. Prijatelj, J. Dulay, and W. Scheirer, "Measuring human perception to improve open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11382–11389, Sep. 2023, doi: 10.1109/TPAMI.2023.3270772.

[28] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5652–5667, Nov. 2018, doi: 10.1109/TIP.2018.2861573.

[29] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural Netw.*, vol. 111, pp. 47–63, Mar. 2019.

[30] S. Su and J. Hu, "Gas identification by a single metal-oxide-semiconductor sensor assisted by ultrasound," *ACS Sensors*, vol. 4, no. 9, pp. 2491–2496, Sep. 2019.

[31] P. R. M. Júnior, R. M. de Souza, R. D. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. B. Penatti, R. D. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Mach. Learn.*, vol. 106, no. 3, pp. 359–386, Mar. 2017.

[32] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2532–2541, doi: 10.1109/CVPR.2019.00264.

[33] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1563–1572, doi: 10.1109/CVPR.2016.173.

[34] W. Dang, J. Guo, M. Liu, S. Liu, B. Yang, L. Yin, and W. Zheng, "A semi-supervised extreme learning machine algorithm based on the new weighted kernel for machine smell," *Appl. Sci.*, vol. 12, no. 18, p. 9213, Sep. 2022.

[35] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[36] M. F. Wahid, R. Tafreshi, and R. Langari, "A multi-window majority voting strategy to improve hand gesture recognition accuracies using electromyography signal," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 427–436, Feb. 2020, doi: 10.1109/TNSRE.2019.2961706.

[37] D. Krstinić, L. Šerić, and I. Slapničar, "Comments on 'MLCM: Multi-label confusion matrix,'" *IEEE Access*, vol. 11, pp. 40692–40697, 2023, doi: 10.1109/ACCESS.2023.3267672.

[38] G. Wang, K.-S. Choi, J. Y. Teoh, and J. Lu, "Deep cross-output knowledge transfer using stacked-structure least-squares support vector machines," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3207–3220, May 2022, doi: 10.1109/TCYB.2020.3008963.

[39] J. Chang and D. Song, "RGA-Net: An effective deep learning method combined with an electronic nose system to identify the pork quality," *IEEE Sensors J.*, vol. 23, no. 19, pp. 22161–22168, Oct. 2023, doi: 10.1109/JSEN.2023.3305292.

[40] D. R. Wijaya, R. Sarno, and A. F. Daiva, "Electronic nose for classifying beef and pork using Naïve Bayes," *Int. Seminar Sensors, Instrum., Meas. Metrology (ISSIMM)*, Surabaya, Indonesia, 2017, pp. 104–108, doi: 10.1109/ISSIMM.2017.8124272.

[41] W. Zhang, L. Wang, J. Chen, W. Xiao, and X. Bi, "A novel gas recognition and concentration detection algorithm for artificial olfaction," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 2509514, doi: 10.1109/TIM.2021.3071313.

[42] J. Wen, Y. Zhao, Q. Rong, Z. Yang, J. Yin, and Z. Peng, "Rapid odor recognition based on reliefF algorithm using electronic nose and its application in fruit identification and classification," *J. Food Meas. Characterization*, vol. 16, no. 3, pp. 2422–2433, Jun. 2022.

[43] M. Hassen and P. K. Chan, "Learning a neural-network-based representation for open set recognition," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2020, pp. 154–162.

[44] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018, doi: 10.1109/TPAMI.2017.2707495.
[45] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021, doi: 10.1109/TPAMI.2020.2981604.
[46] O. Saidani, M. Alsafyani, R. Alroobaea, N. Alturki, R. Jahangir, and L. Jamel, "An efficient human activity recognition using hybrid features and transformer model," *IEEE Access*, vol. 11, pp. 101373–101386, 2023, doi: 10.1109/ACCESS.2023.3314492.

RUN ZHOU was born in Hunan, China, in 1982. He received the B.S. degree in communication engineering from Xidian University, Xi'an, China, in 2008, the M.S. degree from Xi'an Shiyou University, Xi'an, in 2011, and the Ph.D. degree in communication engineering from Northwestern Polytechnical University, Xi'an, in 2019. His current research interest includes artificial intelligence.

JIANHAO WANG was born in Zhejiang, China, in 1995. He received the B.S. and M.S. degrees from the College of Electrical Engineering, Fuzhou University, Fuzhou, China, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering. His current research interests include power electronics and current transformers.

KANG HUANG was born in Zhejiang, China, in 1994. He received the joint B.Eng. degree (Hons.) from The University of Sydney, Sydney, Australia, and the Harbin Institute of Technology, Harbin, China, in 2016, and the Ph.D. degree from The University of Sydney, in 2021. His research interests include wireless networked control systems and artificial intelligence.

HUI WANG received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 2008. From 2008 to 2009, he was a Postdoctoral Fellow with the Department of Computer Science, Evry University, France. From 2009 to 2011, he was a BK21 Postdoctoral Fellow with the Department of IT Convergence, POSTECH, South Korea. He is currently an Associate Professor with the Department of Computer Science, Zhejiang Normal University. His research interests include compressive sensing, edge/fog computing, distributed optimization, online optimization, edge intelligence, and self-adaptive IoT.

ZHIJUN CHEN was born in Zhejiang, China, in 1982. He received the master's degree from Hangzhou Dianzi University, Hangzhou, China, in 2007. His current research interests include big data technology applications and vocational education.