https://doi.org/10.31436/iiumej.v24i2.2832

FEATURE EXTRACTION AND SUPERVISED LEARNING FOR VOLATILE ORGANIC COMPOUNDS GAS RECOGNITION

NOR SYAHIRA MOHD TOMBEL[1], HASAN FIRDAUS MOHD ZAKI[2,3]* AND HANNA FARIHIN BINTI MOHD FADGLULLAH[3]

[1]Dept. of Science (Computational and Theoretical), Kulliyyah of Science, International Islamic University of Malaysia, Kuantan, Pahang
[2]Centre for Unmanned Technologies (CUTe), Kulliyyah of Engineering, International Islamic University of Malaysia, Gombak, Kuala Lumpur
[3]Dept. of Mechatronic, Kulliyyah of Engineering, International Islamic University of Malaysia, Gombak, Kuala Lumpur

**Corresponding author: hasanzaki@iium.edu.my*

(Received: 11 April 2023; Accepted: 13 June 2023; Published on-line: 4 July 2023)

ABSTRACT: The emergence of advanced technologies, particularly in the field of artificial intelligence (AI), has sparked significant interest in exploring their potential benefits for various industries, including healthcare. In the medical sector, the utilization of sensing systems has proven valuable for diagnosing pulmonary diseases by detecting volatile organic compounds (VOCs) in exhaled breath. However, the identification of the most informative and discriminating features from VOC sensor arrays remains an unresolved challenge, essential for achieving robust VOC class recognition. This research project aims to investigate effective feature extraction techniques that can be employed as discriminative features for machine learning algorithms. A preliminary dataset was used to predict VOC classification through the application of five supervised machine learning algorithms: k-Nearest Neighbors (kNN), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Networks (ANN). Ten feature extraction methods were proposed based on changes in sensor response as inputs to classify three types of gases in the dataset. The performance of each model was evaluated and compared using k-Fold cross-validation (k=10) and metrics derived from the confusion matrix. The results demonstrate that the RF model achieved the highest mean accuracy and standard deviation, with values of 0.813 ± 0.035, followed closely by kNN with 0.803 ± 0.033. Conversely, LR, SVM (kernel=Polynomial), and ANN exhibited poor performances when applied to the VOC dataset, with accuracies of 0.447 ± 0.035, 0.403 ± 0.041, and 0.419 ± 0.035, respectively. Therefore, this paper provides evidence that classifying VOC gases based on sensor responses is feasible and emphasizes the need for further research to explore sensor array analysis to enhance feature extraction techniques.

ABSTRAK: Perkembangan teknologi canggih, khususnya dalam bidang kecerdasan buatan (AI), telah mencetuskan minat yang ketara dalam menerokai manfaatnya untuk pelbagai industri, termasuk bidang kesihatan. Dalam sektor perubatan, penggunaan sistem penderiaan telah terbukti bernilai untuk mendiagnosis penyakit paru-paru dengan mengesan sebatian organik meruap (VOC) dalam nafas yang dihembus manusia. Walau bagaimanapun, pengenalpastian ciri yang paling bermaklumat dan mendiskriminasi daripada penderia VOC kekal sebagai cabaran yang tidak dapat diselesaikan, penting untuk mencapai pengiktirafan kelas VOC yang kukuh. Projek penyelidikan ini bertujuan untuk menyiasat teknik pengekstrakan ciri yang berkesan yang boleh digunakan sebagai ciri diskriminatif untuk

https://doi.org/10.31436/iiumej.v24i2.2832

algoritma pembelajaran mesin. Set data awal digunakan untuk meramalkan klasifikasi VOC melalui aplikasi lima algoritma pembelajaran mesin yang diselia: k-Nearest Neighbors (kNN), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), dan Artificial Neural Networks (ANN). Sepuluh kaedah pengekstrakan ciri telah dicadangkan berdasarkan perubahan dalam tindak balas penderia sebagai input untuk mengklasifikasikan tiga jenis gas dalam set data. Prestasi setiap model telah dinilai dan dibandingkan menggunakan pengesahan silang *k-Fold* (k=10) dan metrik yang diperoleh daripada *confusion* matriks . Keputusan menunjukkan bahawa model RF mencapai ketepatan minima tertinggi dan sisihan piawai, dengan nilai 0.813 ± 0.035, diikuti oleh kNN dengan 0.803 ± 0.033. Sebaliknya, LR, SVM (kernel=Polinomial), dan ANN mempamerkan prestasi yang lemah apabila digunakan pada dataset VOC, dengan ketepatan masing-masing 0.447 ± 0.035, 0.403 ± 0.041 dan 0.419 ± 0.035. Oleh itu, kertas kerja ini memberikan bukti bahawa mengklasifikasikan gas VOC berdasarkan tindak balas penderia adalah boleh dilaksanakan dan menekankan keperluan untuk penyelidikan lanjut untuk meneroka analisis tatasusunan penderia untuk meningkatkan teknik pengekstrakan ciri.

KEYWORDS: *Supervised machine learning; Volatile Organic Compound; VOC Sensor; Gas classification; feature extraction*

1. INTRODUCTION

Volatile organic compounds (VOC) have been used as preclinical biomarkers in breath analysis to monitor health and diagnose various pulmonary diseases such as asthma and lung cancer [1] [2] [3][4]. An array of sensors, or electronic nose (e-nose) is known as the alternative for a non-invasive method of detecting volatile organic compounds (VOC). E-nose is a device inspired by the olfactory system of humans or mammals (sense of smell), composed of a collection of an array of gas sensors with a pattern recognition system designed to detect and differentiate a wide variety of gas compounds [5].

The advancement of nanosensor arrays with pattern recognition involving pre-processing, feature extraction and machine learning algorithms makes it a powerful tool for the detection and recognition of gas samples with concentration estimation. Feature extraction is an essential technique used to extract significant information from the sensor response signal [6] [7] to optimize the performance of pattern recognition algorithms for gas classification [6] [8].

However, the detection of VOC using nanosensor technologies still has some constraints in its detection system. The VOC sensor as a sensing unit faced a few limitations such as lack of sensitivity and selectivity [9] [10]. Besides, it is still not clear which type of features from VOC sensor arrays are the most descriptive and discriminative leading to a robust recognition of the VOC classes. Data collection from a gas sensor array can also be cumbersome and time-consuming which poses a nuisance in employing data-hungry machine learning algorithms.

Therefore, this paper proposes employing supervised machine learning algorithms to classify the preliminary data of the individual sensors for VOC recognition. The VOC detection was performed on a chemiresistive sensor from various functionalised reduced Graphene Oxide (rGO) as a sensing layer. The targeted VOC gases used are acetone, toluene and isoprene which have been suggested as pulmonary disease-related biomarkers [11] with concentration levels ranging from 1 to 6 ppm.

Concretely, we explore 10 feature sets that were extracted from the sensor's original response curve. Then, we analyse the effect of these features towards VOC classification with five benchmark machine learning algorithms including K-Nearest Neighbours (kNN), Random

https://doi.org/10.31436/iiumej.v24i2.2832

Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Artificial Neural Network (ANN). The recognition models were then put into comparison to determine the one which provides the best evaluation and high accuracy in performing the classification of the targeted VOC gases using k-Fold Cross Validation (k=10) and Confusion Matrix.

2. GAS SENSING MECHANISM

Sensing mechanism of the sensor is first studied, to understand the VOC detection on the sensor. The thin film is comprised of rGO which is one of the most promising materials for detecting low VOC concentrations at room temperature [12]. Graphene is a two-dimensional building block made up of a one-atom-thick sheet of a carbon atom.

Graphene can work well at room temperature because it has enormously high mobility [13]. Researchers are interested in modifying graphene into reduced Graphene Oxide as a sensing element because of its excellent electrical, high thermal conductivity, and mechanical properties [14] [15]. The functionalisation of rGO with nanoparticles and plasma treatment can improve sensor functionality and selectivity in distinguishing different vapours [11]. Different functionalisation of sensing elements is a good technique to improve the gas sensor's sensitivity and characteristics.

In this research, the sensing layer was deposited on Ti/Pt Interdigitated Electrode (IDE). The electrode was used to supply current flow from the power source to the device, which improved the sensing material's catalytic properties towards a specific gas [16]. Furthermore, the VOC sensor employed is a resistive type, which produces a signal based on a change in resistance in response to gas exposure. In general, VOC gas detection on a sensor is caused by the adsorption and desorption processes that occur between analytes and the sensor surface [17].

Oxygen ion species were absorbed on the sensor surface in the presence of air (humidity) and lowered the electron from the conduction band [18]. The electron density is falling off and forming an electron depletion layer and barrier potential on the surface. Electron removal causes an increase in the depletion layer. The related equation for chemisorbed oxygen at temperatures less than 100°C [19] is as follows:

$$O_2\ (gas) \rightarrow O_2\ (adsorption) \quad (1)$$

$$O_2\ (gas) + e^-\ (surface) \leftrightarrow O_2^-\ (adsorption)\ (<100°C) \quad (2)$$

When VOC gas was introduced into the chamber, the gas molecules started to react with the absorbed oxygen ions and released electrons back into the conduction band. The predominant carrier in the sensors was modified by the reaction of the VOC gas (oxidizing or reducing agent) with the molecules in the sensing layer, resulting in an increasing or decreasing in the resistance measurement as the output [19].

Reduced Graphene Oxide has been reported to exhibit p-type behaviour [20]. However, the functionalised sensor was shown to be an n-type semiconductor in this VOC test, and the VOC analytes acted as reducing gases [18]. The sensor experienced an electron carrier majority, causing a decrease in depletion width and potential barrier. As a result, the sensor resistance decreased in the presence of VOC gas [21].

https://doi.org/10.31436/iiumej.v24i2.2832

3. RELATED WORKS

3.1. Feature Extraction

Feature extraction is a technique that is used to extract significant information from the sensor response graph [6] to ensure better performance of machine learning algorithms in pattern recognition [8]. The information is deemed relevant when the derived value extracted from the measured data is non-redundant, not correlated with other features and projects the decisive features [22]. Other than that, feature selection is also related to the dimensionality reduction process of transforming high dimensional data into a low dimensional feature [8].

Detection of VOCs using gas sensors commonly used real-time analysis and discrimination of "breath prints" to perform the gas classification process [2]. In 2012, Vergara and his team applied 8 feature extraction from the time-series sensor, which are the change of the maximal resistance change (ΔR), the normalized resistance change ($||\Delta R||$), minimum and maximum exponential moving average(ema) with a value of α = 0.001,0.01,0.1 each [10].

On the other hand, many features can be extracted from raw signals and applied in electronic nose applications. Commonly extracted features from gas original response curves such as maximum response, the response of special time, time of special response, area, integral, derivative, difference and second derivative [6]. Table 1 shows a few lists of feature extraction from electronic nose sensor data for wound detection [23].

Table 1: List of Feature Extraction for Wound Detection based Electronic Nose [23]

Feature	Description
Normalization	Preprocessing the sensor data for features from the steady-state response, eliminate the effect of a concentration difference on recognition.
Integral and derivative methods	Integrals may represent the accumulative total of the reaction degree change and derivatives may represent the rate at which the sensor reacts to the odour.
The Root Mean Square Error (RMSE) of curve fitting	Depends on the type of model and the number of parameters in the model.
Fourier transform and wavelet transform	Fourier transform decomposes the original response curve into a superposition of the DC component and different harmonic components.

3.2. Supervised Machine Learning

There were few studies which implemented the detection of different gases by Supervised learning models such as k-Nearest Neighbour (kNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF) and Logistic Regression. The findings were summarised in Table 2.

There are two gaseous flows in the system: for carrier gas and VOC gas. Clean Dry Air (CDA) was used as carrier gas, while isoprene, toluene, and acetone as the targeted VOC gas. The temperature of the gas and temperature chuck in the sensor chamber were controlled using a Cellkraft Humidifier P-10 and a Nextron Temperature controller module (Nextron Microprobe Station, with platier heater and 4 probe needles). Agilent SMU 34410A was used to drive the voltage and input current. A data acquisition (DAQ) system that is used to convert the output/measured signal from the sensor system into the computer is via a user interface software that is programmed using the LabVIEW program, provided by MIMOS.

Table 2: Summary of Implementation of Supervised Learning Algorithms on Different Types of Gas Detection.

Model	Description	References
k-Nearest Neighbour (Knn)	kNN is widely used in the classification of mixed gas and for gas discrimination systems. The kNN model is advantageous because it is comprehensible, insensitive to noise, low cost for retaining and good combination with other algorithms. However, this model is sensitive to sample distribution, it has a slow speed for recognition, high spatial complexity, heavy calculation burden and poor interpretability.	[24]
Support Vector Machines (SVM)	SVM of classifiers can cope well with gas sensor drift and perform better than the baseline competing methods on the extensive dataset. However, the SVM model requires a long learning time and poor application for larger data. Choice of kernel function is important as it is the key for feature space in SVM.	[10],[24]
Artificial Neural Network (ANN)	ANN is the frequently used method in predicting and analysing complex gas (Hashoul & Haick, 2019). It has good learning ability, good parallel processing capability and detecting compatibility error. However, this model has poor interpretability for output, long time learning and is easy to overfit. Therefore, weight, activation function and the number of hidden layers are important to develop an ANN algorithm in performing the classification of targeted output.	[24],[25]
Random Forest (RF)	RF model is used in a lot of feature datasets as it can prevent overfitting from a decision tree algorithm. In the Random Forest algorithm, the number of trees affected the accuracy of the model, as each tree has a classification result and the final result is based on the majority decision trees vote	[26], [36]
Logistic Regression	LR is a classification algorithm that calculates linear output and statistical function through the regression output. Logistic regression can perform multiclass classification problems by using one-vs-rest or one-vs-one wrapper models. The algorithm can be applied to a non-linear classification problem with a proper feature selection. LR model can produce high accuracy as it is a good signal to noise ratio.	[27]

4. EXPERIMENTAL SETUP

As illustrated in Fig 1, the gas sensing system for this study comprises a gas supply system, a sensor chamber, a temperature and humidity controller module, and data collection system [28].

Fig 2 showed the DAQ board of the LabVIEW that contain all the controller variables for the test measurement such as flow of the CDA, flow of the VOC gas, input current, input voltage, temperature inside the chamber, temperature of the sensor's heater, relative humidity, and system ramp rate. Then, the sensor was tested individually with the targeted VOC gas and the sensor’s responses were recorded to study performance of the individual sensor.

https://doi.org/10.31436/iiumej.v24i2.2832

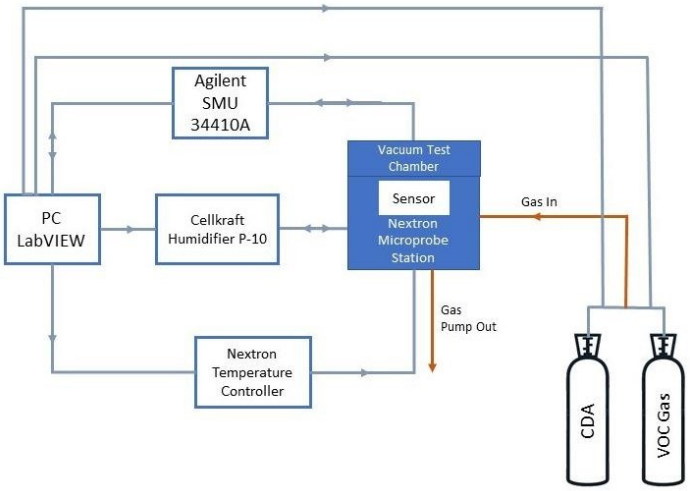

Fig. 1. Schematic Diagram of the Experimental Setup for Gas Detection System.

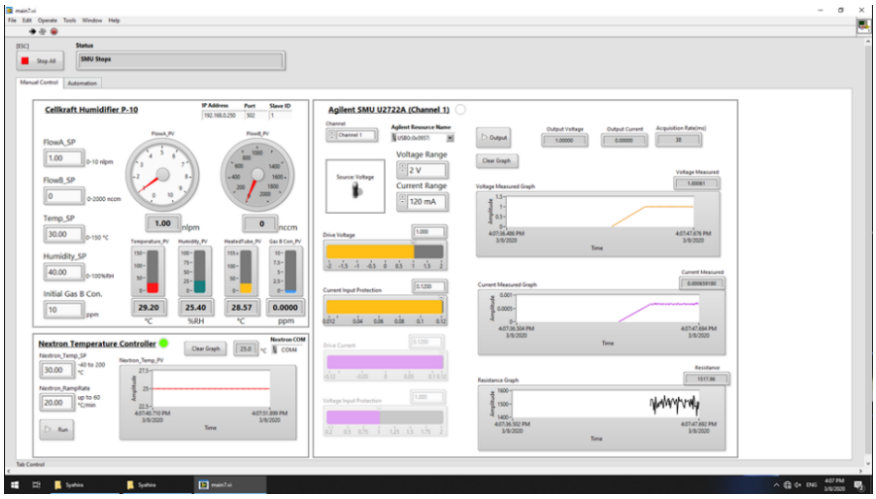

Fig. 2. Manual control of LabVIEW Gas System.

4.1. VOC Sensor

The gas sensor used in this study is called a VOC sensor, which is prepared, fabricated and functionalised by the engineering team at MIMOS Bhd. Reduced Graphene Oxide (rGO) as a sensing membrane was deposited on the Platinum-titanium interdigitated electrode (Pt/Ti IDE) on a silicon and silicon dioxide (Si/SiO_2) substrate. The rGO was functionalised with nanoparticles such as; gold (Au), silver (Ag) and platinum (Pt) and plasma treatment such as; hydrogen (H_2) and Octafluorocyclobutane (C_4F_8).

The sensor was fabricated using a standard semiconductor process using Chemical Vapor Deposition (CVD) by a standard lithography process for the functionalisation with different recipes. The rGO was functionalised with nanoparticles at a different duration of sputtering and Relative Frequency (RF) power, while functionalisation with plasma treatment at a variety of plasma power and temperature. Therefore, there are 21 individual VOC gas sensors used in this study and the details are according to Table 3. Next, the pre-processed signal proceeded with a feature extraction method to extract pertinent information to be input for supervised machine learning at classifying the gas components into targeted gas output. The features were

https://doi.org/10.31436/iiumej.v24i2.2832

decided to extract from the original gas response involving measured resistance in the absence and presence of the VOC gas.

Table 3: List of Functionalisation and Recipe of VOC Sensors

Sample no.	Nanoparticles	Power [W_{RF}]	Time [sec]	Remarks
1		Reference rGO film		Bare Rgo
2	Au	30	15	rGO/Au (30W 15s)
3	Au	30	75	rGO/Au (30W 75s)
4	Au	70	15	rGO/Au (70W 15s)
5	Au	70	75	rGO/Au (70W 75s)
6	Pt	30	15	rGO/Pt (30W 15s)
7	Pt	30	75	rGO/Pt (30W 75s)
8	Pt	70	15	rGO/Pt (70W 15s)
9	Pt	70	75	rGO/Pt (70W 75s)
10	Ag	30	15	rGO/Ag (30W 15s)
11	Ag	30	75	rGO/Ag (30W 75s)
12	Ag	70	15	rGO/Ag (70W 15s)
13	Ag	70	75	rGO/Ag (70W 75s)
Sample no.	**Plasma Treatment**	**Power (W_{RF})**	**Temperature (℃)**	**Remarks**
14	H_2	-	RT	rGO/H_2 RT ℃
15	H_2	-	200	rGO/H_2 200 ℃
16	H_2	-	400	rGO/H_2 400 ℃
17	H_2	-	700	rGO/H_2 700 ℃
18	C_4F_8	150	-	rGO/C_4F_8 150 ℃
19	C_4F_8	200	-	rGO/C_4F_8 200 ℃
20	C_4F_8	250	-	rGO/C_4F_8 250 ℃
21	C_4F_8	300	-	rGO/C_4F_8 300 ℃

4.2. Data Collection

The sensors were tested individually with each of the selected VOC gas. The sensor was placed in a chamber with 30℃ of temperature and presence of 40% relative humidity (RH). The voltage and current input were set at 1V and 1.2A respectively. CDA was maintained at 1 L/min for 5 minutes to stabilize the baseline reading. Then, the VOC gas was purged into the chamber with a gradual increase of concentrations, from 1 to 6 ppm in 12 minutes (2 minutes for each concentration). The sensor responses were analysed from the resistance changes of individual sensors that undergo the VOC test.

5. DATA PRE-PROCESSING AND FEATURE EXTRACTION

In this phase, the analytes of the VOC gas were reacting with the sensing element, thus leading to a change in resistance. The sensor response was determined by analysing the measured resistance as a signal output from each sensor. However, the parameter setup was not in optimal condition and the output signal contained unexpected noise from the SMU system. A typical sensor response could not be seen clearly from the graph of resistance versus time.

As a result, the signal was pre-processed by applying filter and smoothing methods to denoising the signal and reduce the influence of random variation caused by instrumental conditions and atmospheric effects [29]. The data was filtered using a moving average (MA

https://doi.org/10.31436/iiumej.v24i2.2832

length = 3) and smoothed with Minitab software using a single exponential method with a constant = 0.02 value. The sensor response was determined by using the formula [30][31]:

$$\Delta R = R_g - R_a \quad (3)$$

$$S = \frac{\Delta R}{R_a} = \frac{R_g - R_a}{R_a} \times 100(\%) \quad (4)$$

Where,

R_a = resistance in clean dry air, without VOC gas

R_g = resistance with the exposure of VOC gas

Next, the pre-processed signal proceeded with a feature extraction method to extract pertinent information as input for supervised machine learning at classifying the gas components into targeted gas output. The features were decided to extract from the original gas response involving measured resistance in the absence and presence of the VOC gas. The 10 selected features as listed in Table 4.

Table 4: List of Features and Formula for the 10 Selected Features

No.	Feature	Formula
1	R_{gas}	Resistance at steady-state phase
2	R_0	Baseline resistance
3	Sensor Response, S	$\frac{\Delta R = R_{gas} - R_{air}}{R_{air}}$
4	Difference, ΔR	$\Delta R = R_{gas} - R_{air}$
5	Relative difference	$\frac{R_{gas}}{R_{air}}$
6	Log relative resistance value	$log(\frac{R_{gas}}{R_{air}})$
7	Normalisation	$\left\|\frac{R_{gas} - R_{air}}{R_{air}}\right\|$
8	G R_{gas}	$\frac{1}{R_{gas}}$
9	G R_0	$\frac{1}{R_{air}}$
10	Conductance difference	G_{gas} - G_{air}

The VOC dataset comprises ten feature extraction values and is organised into three categories. In summary, there are 918 total samples, with 252, 324, and 342 each for acetone, toluene, and isoprene gas, respectively.

6. SUPERVISED LEARNING FOR VOC GAS CLASSIFICATION

Five supervised learning models including k-Nearest Neighbour (kNN), Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were benchmarked for the VOC gas classification. The model was implemented using the Python and Scikit-learn library. Each model's parameter settings are described in Table 5.

To avoid bias in the analysis, the dataset was first standardised in the range 0 to 1 to uniform the values with different scales by using the min-max normalisation technique [32].

https://doi.org/10.31436/iiumej.v24i2.2832

Following that, the dataset was divided into 70% for the training set and 30% for the testing set. The performances of each model are then evaluated from confusion matrix-based measures in terms of accuracy, precision and using k-fold cross-validation technique, where k =10. K-fold Cross Validation is a cross validation technique used to evaluate the performance of a machine learning model by the resampling procedure. The training of the models proceeds using the k-1 parts and validation or testing errors from the remaining part [33].

Table 5: Parameter Setting of the Approached Supervised Machine Learning

Model	Parameter
K-nearest Neighbour	The K-value is decided as one, (k=1) and distance between two points is calculated by applying the distance metric formula (2), (mentioned in chapter 2), with p = 2, to manipulate the generalised distance to Euclidean Distance.
Random Forest	Grid search for the setting parameter, with n-estimator:100.
Artificial Neural Network	A shallow Neural Network was implemented, with a standard three-layer feed-forward network. For the hidden layer, the size was set up to 50 and used the ReLU activation function. While Softmax activation function for output layer with learning rate 0.001 and 50 epochs.
Support Vector Machine (RBF kernel)	Radial basis function (RBF) was selected as a kernel function for this SVM model, as defined in equation (6) (in Chapter 2) The kernel function, σ and regularisation, C used GridSearchCV from sci-kit- learn library to perform grid search for parameter setting.
Logistic Regression	The model used 'l2' for regularisation (penalty) and solver 'lbfgs'.

7. RESULT

Figures 3 a) to e) showed the results of the confusion matrices for each proposed supervised learning method. Meanwhile, Table 6 below showed the accuracies from the 10-fold cross validation from each model.

The diagonal values in the confusion matrix denoted the accuracy values of the gas classification to the targeted output [34]. Figure 3 shows that the kNN and RF models performed well in classifying each of the targeted VOCs. The kNN model correctly predicted all three gases with greater than 80% accuracy, while the Random Forest model predicted them with greater than 70% accuracy. On the other hand, Logistic Regression, Support Vector Machine (kernel = RBF) and Artificial Neural Network performed poor classification on the 3 VOCs gases. The SVM and ANN models misclassified the gas more to isoprene gas while LR model misclassified the toluene and isoprene gas.

It is noticeable in Table 6, the RF model showed the highest mean accuracy with 0.813 ±0.035, followed by the kNN model with 0.803 ±0.033. RF model are known to have advantages in the process of random sampling which can ensure randomness and avoid overfitting. Besides, this model is also robust to noise [5] and it is good at handling missing data and imbalance classes [4].

Whereas, the kNN model has limitations in understanding the relationship between the features and the class (output) thus easily producing the wrong classification for a multiclass problem [4]. Therefore, the highest accuracy achieved by kNN in this study showed that the features were related well to the output class.

https://doi.org/10.31436/iiumej.v24i2.2832

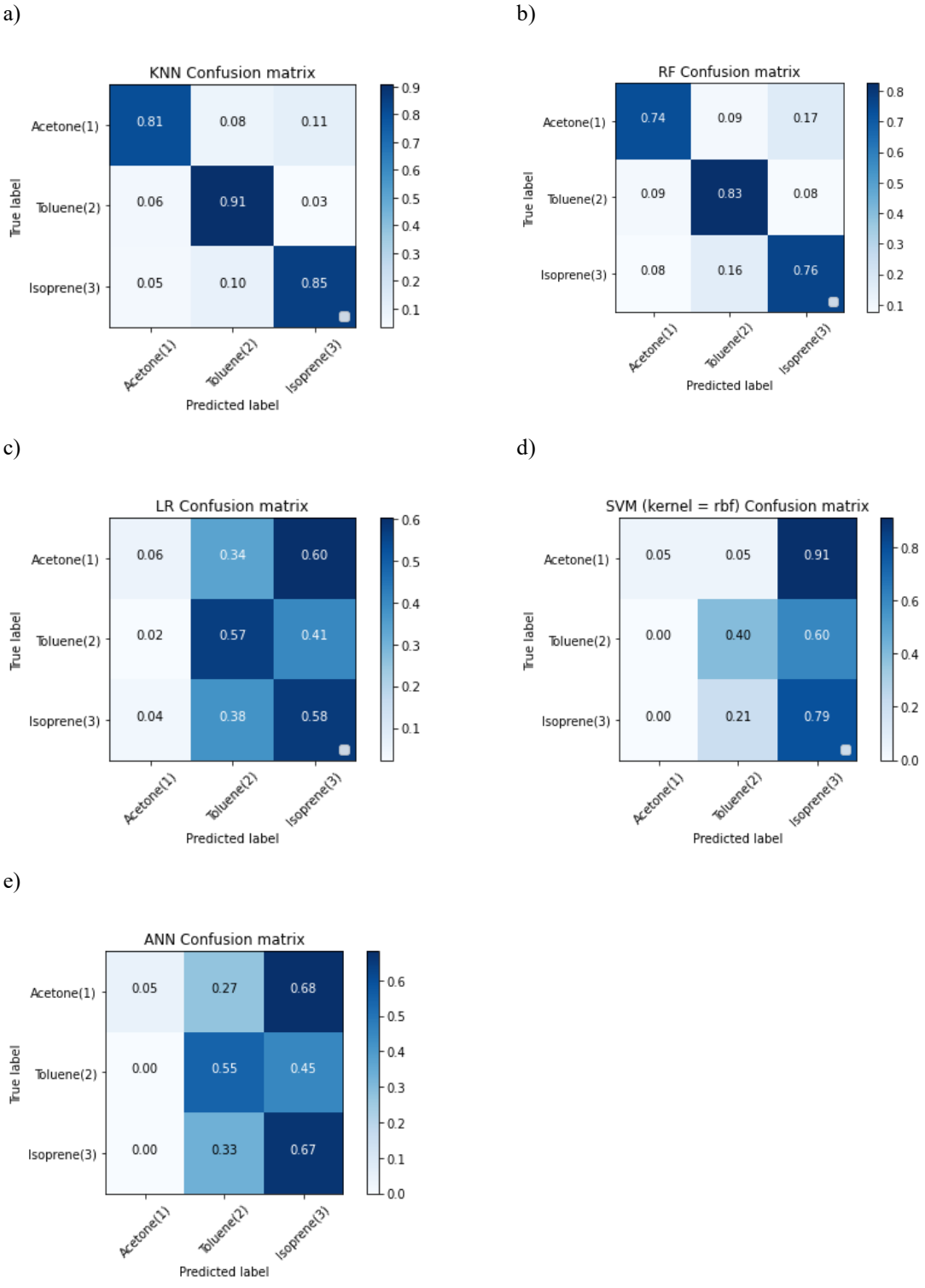

Fig. 3. Comparison of Normalised Confusion Matrix of a) kNN, b) RF, c) LR, d) SVM (kernel= RBF) and e) ANN.

However, the ANN, LR and SVM (Polynomial kernel) models had poor performance compared with the RF and kNN model with accuracies of 0.447 ± 0.035, 0.403 ± 0.041 and 0.419 ± 0.035 respectively. Meanwhile, the SVM model with Grid search parameters showed

https://doi.org/10.31436/iiumej.v24i2.2832

high accuracy by using a Polynomial kernel at degree = 3, compared with other kernels such as Linear and RBF.

Table 6: Model Performance based on 10-Fold Cross-Validation Technique.

Model	10-Fold Cross-Validation (Mean Accuracy ± Standard Deviation)
Random Forest	0.813 ± 0.035
K-Nearest Neighbors	0.803 ± 0.033
Logistic Regression	0.403 ± 0.041
Support Vector Machine (Kernel = Polynomial)	0.419 ± 0.035
Support Vector Machine (Kernel = Linear)	0.401 ± 0.039
Support Vector Machine (Kernel = RBF)	0. 408 ± 0.055
Artificial Neural Network	0.447 ± 0.035

The poor performance from the ANN, LR and SVM (Polynomial kernel) models might be due to their weakness, in which they are very prone to overfitting training data [33] and they required testing with various kernels and model parameters [4]. The ANN model also is a learning-based algorithm and is more complex in architecture. Thus, it has more hyperparameters required to be tuned [33] and it needs enough samples for training.

Other than that, the ROC curve and AUC value are another way of visualising the output performances from the computed confusion matrix. Evaluation of the Receiver Operating Characteristics (ROC) and Area Under Curve (AUC) was done to analyse the performance of the classifiers. The highest value of the AUC showed good value prediction of the model to assign a larger probability to a random positive example than a random negative example [35]. The AUC value should be between 0.5 and 1.0. The ROC curves for each classifier are illustrated in Figure 4. As the minimum is 0.541, it can be said that the SVM classifier does not predict our dataset very well and could not differentiate the classes, while the highest value of AUC goes to the KNN classifier which is equal to 0.886 and 0.885 for the RF model. As a result, in this research, the kNN and RF are two models that can deal with the selected features in the VOC dataset as they obtained the highest accuracy for the gas classification.

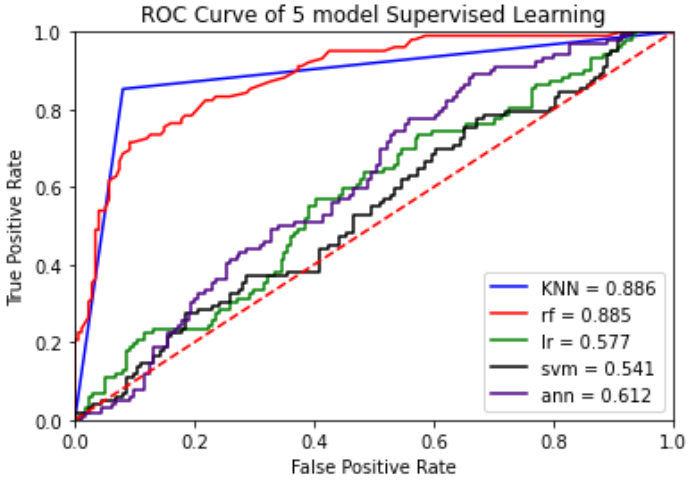

Fig. 4. ROC Curve of the 5 Supervised Learning Model

8. CONCLUSION

The gas sensor data was collected at a preliminary stage and has been used for the machine learning part, which involved pre-processing, feature extraction and classification algorithm. Each sensor was performing well at low operating temperatures and in the presence of humidity. The sensors response on different targeted VOC gas from 1 to 6 ppm were collected. Then, feature extracted were performed on the resistance-based data.

Then, feature extracted were performed on the resistance-based data. Ten featured were proposed as inputs to five supervised learning algorithms to accurately recognise and classify the selected VOC gas based on the labelled output. The confusion matrix and 10-Fold Cross Validation were used to evaluate each model's performance. As a result, the RF and kNN models have higher accuracy with $0.813 \pm 0.035$and 0.803 ± 0.033, compared with LR, SVM and ANN with the accuracy of 0.447 ± 0.035, 0.403 ± 0.041 and 0.419 ± 0.035 respectively. The two highest accuracies achieved by RF and kNN models demonstrated that they distinguished the gas well from the VOC dataset.

Despite the gas sensor's shortcomings, such as low sensitivity, selectivity, and noise in the sensor signal output, the findings of this study can be utilised as a guide for selecting the optimum algorithm for dealing with a gas sensor array. The performance of the kNN and RF models is the Proof of Concept that the algorithm can perform gas classification tasks from the simplest feature selected from the steady-state phase. The feature extraction approach, on the other hand, can be discovered more from the raw signal to build a dataset with more significant features and relevant information to improve the algorithm's performance.

ACKNOWLEDGEMENT

This project is a collaboration between the International Islamic University Malaysia (IIUM) with the Centre of Unmanned Technologies, Kulliyyah of Engineering and Department of R&D, MIMOS Bhd. This research was partially sponsored by the Fundamental Research Grant Scheme (FRGS19159-0768) and MIMOS Berhad (SPG21-015-0015). Great thanks to research team at Department of Research & Development, MIMOS Berhad; Dr. Ismahadi Syono, Firzalaila Syarina Md Yakin and Siti Aishah Mohamad Badaruddin for their guidance and supervision during data collection and for device preparation.

REFERENCES

[1] Krisher S, Riley A, Mehta K. (2014) Designing breathalyser technology for the developing world: How a single breath can fight the double disease burden. Journal of Medical Engineering and Technology, 38(3), 156–163. doi:10.3109/03091902.2014.890678.

[2] Dragonieri S, Pennazza G, Carratu P, and Resta O. (2017) Electronic Nose Technology in Respiratory Diseases. Lung, 195 (2):157–165. doi:10.1007/s00408-017-9987-3.

[3] Thriumani R, Zakaria A, Hashim YZH, Jeffree AI, Helmy KM, Kamarudin LM, Omar MI, Shakaff AYM, Adom AH, Persaud KC. (2018) A study on volatile organic compounds emitted by in-vitro lung cancer cultured cells using gas sensor array and SPME-GCMS. BMC Cancer, 18:362 doi:10.1186/s12885-018-4235-7.

[4] Chen C, Lin W, Yang H. (2020) Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research. Respiratory Research, 21:45. doi:10.1186/s12931-020-1285-6.

[5] Hu W, Wan L, Jian Y, Ren C, Jin K, Su X, Bai X, Haick H, Yao M, Wu W. (2018) Electronic Noses: From Advanced Materials to Sensors Aided with Data Processing. Advanced Materials Technologies, 4(2),1-38. doi: 10.1002/admt.201800488.

[6] Yan J, Guo X, Duan S, Jia P, Wang L, Peng C, Zhang S. (2015) Electronic Nose Feature Extraction Methods: A Review. Sensors, 15, 27804-27831. doi:10.3390/s151127804 27804–27831.
[7] Zulkhairi MA, Mustafah YM, Abidin ZZ, Zaki HFM, Rahman HA. (2019) Car Detection Using Cascade Classifier on Embedded Platform. 7th International Conference on Mechatronics Engineering (ICOM), Putrajaya, Malaysia, pp. 1-3, doi: 10.1109/ICOM47790.2019.8952064.
[8] Ansari AQ, Khusro A, Ansari MR. (2016) Performance evaluation of classifier techniques to discriminate odors with an E-Nose. 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON, 2325-9418. doi:10.1109/INDICON.2015.7443838.
[9] Yi Z, Li C. (2019) Anti-Drift in Electronic Nose via Dimensionality Reduction: A Discriminative Subspace Projection Projection Approach. IEEE Access, 7, 170087–170095. doi:10.1109/ACCESS.2019.2955712.
[10] Vergara A, Vembu S, Ayhan T, Ryan MA, Homer ML, Huerta R. (2012) Chemical gas sensor drift compensation using classifier ensembles. Sensors and Actuators, B: Chemical, 166–167, 320–329. doi: 10.1016/j.snb.2012.01.074.
[11] Liu B, Huang Y, Kam KW, Cheung WF, Zhao N, Zheng B. (2019) Functionalized graphene-based chemiresistive electronic nose for discrimination of disease-related volatile organic compounds. Biosensors and Bioelectronics: X, 1, 100016. doi: 10.1016/j.biosx.2019.100016.
[12] Gargiulo V, Alfano B, Capua RD, Alfè M, Vorokhta M, Polichetti T, Massera E, Miglietta ML, Schiattarella C, Francia GD. (2018) Graphene-like layers as promising chemiresistive sensing material for detection of alcohols at low concentration. Journal of Applied Physics, 123, 024503. doi:10.1063/1.5000914.
[13] Lu G, Ocola LE, Chen J. (2009) Reduced graphene oxide for room-temperature gas sensors. Nanotechnology, 20, 445502, 1-9. doi:10.1088/0957-4484/20/44/445502.
[14] Lee K, Yoo YK, Chae MS, Hwang KS, Lee J, Kim H, Hur D, Jeong HL. (2019) Highly selective reduced graphene oxide (rGO) sensor based on a peptide aptamer receptor for detecting explosives. Sci Rep, 9, 10297. Scientific Reports, 9. doi:10.1038/s41598-019-45936-z.
[15] Tian W, Liu X, Yu W. (2018) Research progress of gas sensor based on graphene and its derivatives: A review. Applied Sciences (Switzerland), 8(7). doi:10.3390/app8071118.
[16] Lee SP. (2017) Electrodes for Semiconductor Gas Sensors. Sensors, 17, 683; doi:10.3390/s17040683.
[17] Wang C, Yin L, Zhang L, Xiang D, Gao R. (2010) Metal Oxide Gas Sensors: Sensitivity and Influencing Factors. Sensors, 10, 2088-2106. doi:10.3390/s100302088.
[18] Baharuddin AA, Ang BC, Hoong Y, Haseeb ASMA, Wong YC. (2019) Materials Science in Semiconductor Processing Advances in chemiresistive sensors for acetone gas detection. Materials Science in Semiconductor Processing, 103, 104616. doi: 10.1016/j.mssp.2019.104616.
[19] Amiri V, Roshan H, Mirzaei A, Neri G, Ayesh AI. (2020) Nanostructured Metal Oxide-Based Acetone Gas Sensors: A Review. Sensors, 20, 3096. doi:10.3390/s20113096.
[20] Norizan MN, Zulaikha S, Demon N, Halim NA. (2021) The frontiers of functionalized graphene-based nanocomposites as chemical sensors. Nanotechnology Reviews, 10: 330-369. doi:10.1515/ntrev-2021-0030.
[21] James F, Fiorido T, Bendahan M, Aguir K. (2017) Comparison between MOX sensors for low VOCs concentrations with interfering gases ALLSENSORS, pp.39-40.
[22] Phillips CO, Syed Y, Parthaláin NM, Zwiggelaar R, Claypole TC, Lewis KE. (2012) Machine learning methods on exhaled volatile organic compounds for distinguishing COPD patients from healthy controls. Journal of Breath Research, 6(3). doi: 10.1088/1752-7155/6/3/036003.
[23] Yan J, Tian F, He Q, Shen, Y. (2012) Feature Extraction from Sensor Data for Detection of found Pathogen Based on Electronic Nose. Sensors and Materials, 24(2), 57–73.
[24] Feng S, Farha F, Li Q, Wan Y, Xu Y, Zhang T, Ning H. (2019) Review on smart gas sensing technology. Sensors (Switzerland), 19(17), 1–22. doi:10.3390/s19173760.
[25] Hashoul D, Haick H. (2019) Sensors for detecting pulmonary diseases from exhaled breath. European Respiratory Review, 28(152). doi:10.1183/16000617.0011-2019.

[26] Xu Y, Zhao X, Chen Y, Yang Z. (2019) Research on a mixed gas classification algorithm based on extreme random trees. Applied Sciences (Switzerland), 9(9). doi:10.3390/app9091728.
[27] Thorson J, Collier-Oxandale A, Hannigan M. (2019) Using A Low-Cost Sensor Array and Machine Mixtures and Identify Likely Sources. Sensors, 19, 3723. doi:10.3390/s19173723.
[28] Tombel NSM, Badaruddin SAM, Yakin FSM, Zaki HFM, Syono MI (2021) Detection of low PPM of volatile organic compounds using nanomaterial functionalized reduced graphene oxide sensor. AIP Conference Proceedings, 2368 vol. 020004. doi.org/10.1063/5.0057775.
[29] Smolinska A. (2014) Current breathomics - A review on data pre-processing techniques and machine learning in metabolomics breath analysis. Journal of Breath Research, 8, 027105: 22. doi:10.1088/1752-7155/8/2/027105.
[30] Das S, Jayaraman V. (2014) Progress in Materials Science SnO2: A comprehensive review in structures and gas sensors. Progress in Materials Science, 66, pp. 112-255. doi: 10.1016/j.pmatsci.2014.06.003.
[31] Maity A, Raychaudhuri AK, Ghosh B. (2019) High sensitivity NH3 gas sensor with electrical readout made on paper with prevoskite halide as sensor material. Scientific Report, 9, 7777. doi: 10.1038/s41598-019-43961-6.
[32] Webb AR (2002) Statistical Pattern Recognition Statistical Pattern Recognition Second Edition. John Wiley & Sons,Ltd, vol. 9.
[33] Bastuck M. (2019) Improving the Performance of Gas Sensor Systems with Advanced Data Evaluation, Operation and Calibration Methods. Linköping University Electronic Press, 298.
[34] Kong C, Zhao S, Weng X, Liu C, Guan R, Chang Z. (2019) Weighted Summation: Feature Extraction of Farm Pigsty Data for Electronic Nose. IEEE Access, vol. 7, pp. 96732-96742. doi:10.1109/access.2019.2929526.
[35] Allwright S. (2022) How to interpret AUC score. Retrieved from https://stephenallwright.com/interpret-auc-score/.
[36] Nyssa SSC, XueVZ, Justin TN, V. R. Saran KC, Raia CF, Michael JL, Thomas LW, Mike F, Gregory JS, Philippe B, Christopher JH, Steven JK. (2022) Machine Learning-Based Rapid Detection of Volatile Organic Compounds in a Graphene Electronic Nose. ACS Nano 2022 16 (11), 19567-19583. doi: 10.1021/acsnano.2c10240.