Journal of Engineering Technology and Applied Sciences, 2022 *e-ISSN:* 2548-0391
Received 2 December 2021, Accepted 17 June 2022 *Vol. 7, No. 2, 97-107*
Published online 20 August 2022, Research Article *doi: 10.30931/jetas.1030981*

Citation: Aksoy, S., Özavşar, M., Altındal, A., "Classification of VOC Vapors Using Machine Learning Algorithms". Journal of Engineering Technology and Applied Sciences 7 (2) 2022 : 97-107.

CLASSIFICATION OF VOC VAPORS USING MACHINE LEARNING ALGORITHMS

Serra Aksoy[a]*, Muttalip Özavşar[a], Ahmet Altındal[b]

[a] *Department of Mathematics, Faculty of Arts and Sciences, Yildiz Technical University, Istanbul, Turkey*
*serra-aksoy@hotmail.com (*corresponding author), mozavsar@yildiz.edu.tr*

[b] *Department of Physics, Faculty of Arts and Sciences, Yildiz Technical University, Istanbul, Turkey, altindal@yildiz.edu.tr*

Abstract

Detection of volatile organic compound (VOC) vapors, which are known to have carcinogenic effects, is extremely important and necessary in many areas. In this work, the sensing properties of a cobalt phthalocyanine (CoPc) thin film at six different VOC vapors (methanol, ethanol, butanol, isopropyl alcohol, acetone, and ammonia) concentrations from 50 to 450 ppm are investigated. In this sense, it is observed that the interaction between the VOC vapors and the CoPc surface is not selective. It is shown that using machine learning algorithms the present sensor, which is poorly selective, can be transformed into a more efficient one with better detection ability. As a feature, 10 seconds of responses taken from the steady state region are used without any additional processing technique. Among classification algorithms, k-nearest neighbor (KNN) reaches the highest accuracy of 96.7%. This feature is also compared with the classical steady state response feature. Classification results indicate that the feature based on 10 seconds of responses taken from the steady state region is much better than that based on the classical steady state response feature.

Keywords: Volatile organic compound (VOC), machine learning, classification, k-nearest neighbor (KNN)

1. Introduction

The detection of volatile organic compound (VOC) vapors selectively has a great importance in wide range of areas from indoor air quality control to early diagnosis of certain diseases [1-5]. In this sense, gas sensors offer many advantages including low cost, high response and recovery times, and low power consumption [6]. Among the various types of gas sensors based on different operating principles such as acoustic wave-based, calorimetric, capacitive, optical,

etc. resistive type gas sensors are of special importance. In its simplest form, a resistive type gas sensor consists of a sensing element and a transducer which is capable of converting energy of one kind into energy of another kind. High sensitivity exhibited by phthalocyanines (Pcs) and their derivatives in the form of measurable changes in their conductivity when exposed to reducing or oxidizing gases makes them natural choices for VOC vapor sensing. Pcs as sensing elements in resistive type gas sensors have been studied intensively because of their open coordination sites for axial ligation [7]. As a result of their conductivity, Pcs exhibit excellent sensing characteristics of various gases such as NO_2 [8], CO_2 [9], SO_2 [10], VOCs [11], etc.

The main drawback of Pc based sensing devices is indeed their lack of selectivity towards VOC vapors which strongly limits their use in sensing applications. This is a problem that must be solved, especially if it is aimed to manufacture high-tech devices such as an electronic nose. One essential part of an electronic nose is a system that finds a relationship between the sensor response and the gas type in order to detect gases selectively. This can be achieved by utilizing machine learning approach where feature extraction and classification are two important steps. As the feature plays a key role in the performance of classification, it should represent the characteristic of original high dimensional gas sensor data set efficiently. When this is made, finding a successful classification algorithm is usually easier. Since the maximum value represents the final steady-state feature of the entire dynamic response process in the final balance, which reflects the maximum reaction degree change of sensors responding to vapors, it is usually used as the most widely used and simple electronic nose feature [12]. Therefore, many previous works have made contributions based on the steady state response feature in various gas sensor applications [13-16].

In this work, for the detection of six different VOC vapors (methanol, ethanol, butanol, isopropyl alcohol, acetone, and ammonia) of nine different concentrations from 50 ppm to 450 ppm, the experiment was carried out using a resistive type gas sensor based on CoPc (cobalt phthalocyanine) surface. It is observed that the interaction between the vapors and the surface is not selective enough. In order to solve this problem, it is aimed that the system can distinguish vapors with the use of machine learning algorithms. As many studies did for extracting robust information from the sensor response curve, we also used the steady state response. However, results show that classification accuracy was very low when only the steady state response was used in our work. Hence, it is thought that not only one response but also a few seconds of responses from the steady state region could be used in order to utilize more information from the response curve. Without any additional feature processing technique, the 10s data extracted from the response curve after the sensor reaches 90% of its maximum value were directly used without taking into account the other information in the whole response curve. The performance of this feature was tested by various machine learning classification algorithms such as Decision Tree, Support Vector Machine (SVM), K-nearest Neighbor (KNN), and Ensemble Method.

The structure of this paper is as follows: In Section 2 and 3, the measurement and the sensing results are presented. Section 4 introduces the preprocessing stage. Section 5 describes the extracted feature from the gas sensor data set, and Section 6 introduces the classification results of the algorithms. Finally, Section 7 presents the conclusion of our work.

2. VOC sensing measurement

Spin coated thin film of 5'-6'-Bis(17',25',32'-trinitro-phthalocyaninyl) (1,4,7,10-tetrathia-12-crown) dicobalt(II) on interdigital array (IDAT) of Au electrodes was used as sensing element. The synthesis details of the sensing layer, shown in Figure 1, were described in [17]. After the

spin coating of the sensory layer, coated IDA was installed in a home-made detection cell with a capacity of 5×10^{-4} liters. During the VOC sensing experiments, the carrier gas was dry nitrogen with a purity of 99.8% and the desired level of relative humidity was obtained by bubbling the carrier gas through liquid VOC. The experimental setup used during VOC vapor sensing studies is shown in Figure 2. For studying the sensing behaviour of organic vapours with 5'-6'-Bis(17',25',32'-trinitro-phthalocyaninyl) (1,4,7,10-tetrathia-12-crown)dicobalt(II) (CoPc) thin film, it was exposed to six different VOC vapours (methanol, ethanol, butanol, isopropyl alcohol, acetone, and ammonia) and the variations in the sensor current with time were recorded under the applied constant voltage of 0.5 V.

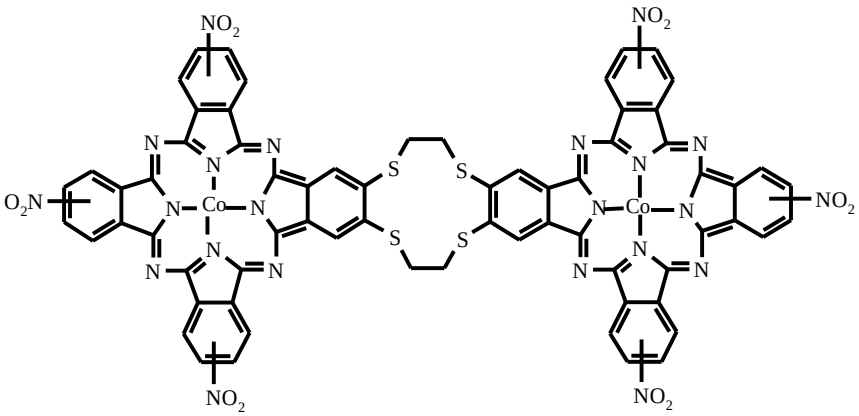

Figure 1. Structural formula for 5'-6'-Bis(17',25',32'-trinitro-phthalocyaninyl) (1,4,7,10-tetrathia-12-crown)dicobalt(II)

Well-defined concentrations of VOC vapors were prepared by mixing the carrier gas with the target vapors. The concentration of the target vapors was varied from 50 to 450 ppm by using mass flow controllers (Alicat Scientific Inc.). In a typical sensing experiment, the sensor surface was exposed to VOC vapor for 20 min. and then purged with carrier gas for another 20 min. to reset the baseline. Total flow rate of the carrier gas was adjusted as 100 standard cubic centimetre (sccm) during the purging experiments. All sensing experiments were performed at a cell temperature of ~ 28 °C.

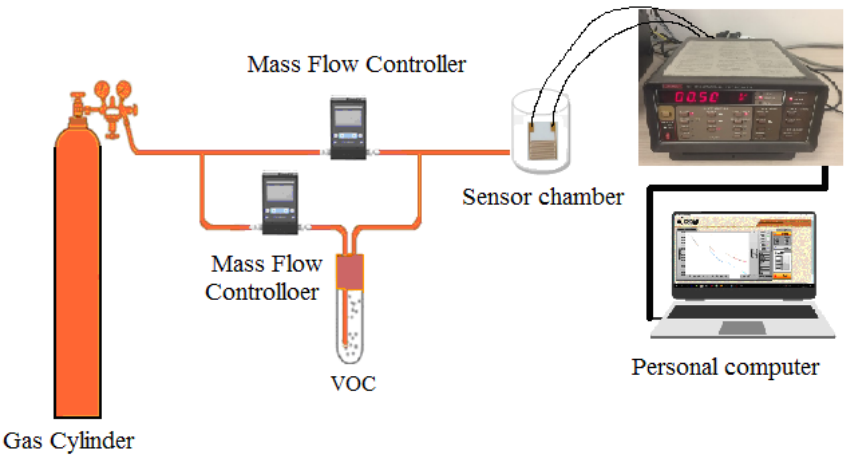

Figure 2. Schematic of the VOC vapor sensing configuration

3. Sensing results

The response recovery characteristic of CoPc thin film with exposure of six organic vapours methanol, ethanol, isopropyl alcohol, butanol, acetone, and ammonia is shown in Figure 3 at different concentrations as labeled in this figure. We observed that CoPc film exhibits an increase in current after the exposure of all the VOC vapors even during ammonia exposure. We also observed during sensing studies of these six organic vapors that after nine cycles of exposure-purging stages, the sensor response recovers to 90% of its initial value. This finding indicates that the interaction between CoPc and VOC vapors is reversible. The maximum increase in sensor current has been observed for methanol vapours, followed by acetone for all concentrations of investigated VOC vapors.

The sensitivities (S) of the CoPc thin film towards VOC vapors have been calculated from the measured dynamic characteristics of the sensor using the following equation

$$S = \frac{1}{C_V}\frac{\Delta I}{I_0} \quad (1)$$

where C_v is the concentration of the VOC vapor under investigation, ΔI is the change in sensor current and I_0 is the baseline current of CoPc thin film before the exposure of the VOC vapours. As is clear from Figure 3, for the same concentrations of VOC vapors, maximum sensitivity (0.0232 ppm^{-1}) has been observed for methanol vapor while minimum sensitivity (0.0024 ppm^{-1}) is observed for isopropylalcohol vapor. It should be mentioned here that the sensitivity values towards methanol is nearly the same for acetone vapor. The similar trend of sensitivity has been observed for ethanol (0.0083 ppm^{-1}) and butanol (0.0077 ppm^{-1}), and for isopropylalcohol (0.0024 ppm^{-1}) and ammonia (0.003 ppm^{-1}) vapors, respectively. This reveals that the interaction between the VOC vapors and the CoPc surface is not selective.

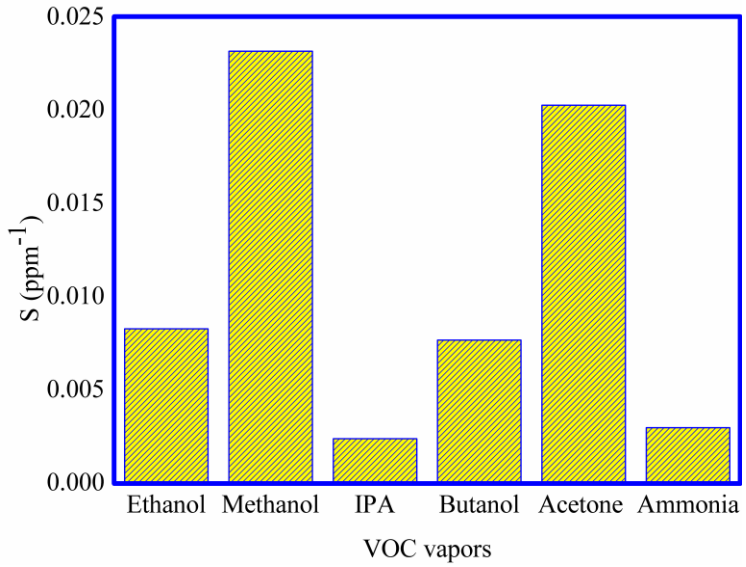

Figure 3. Sensitivities of CoPc thin film towards VOC vapors investigated

As it can be understood by Figure 3, the present sensor is poorly selective and it is not possible to provide a real identification among considered vapors with traditional methods. Hence, we utilized machine learning approach in order to give the system a way to identify the vapors by itself, as explained in the next sections.

4. Preprocessing

The collected data should be preprocessed before the feature extraction step. Since data processing in a machine learning system directly determines the input, it is an indispensable step before training a model. There are no general guidelines to determine the appropriate data preprocessing technique, so the technique to be used varies from application to application. Among different techniques, in this research, in order to reduce the effect of the baseline on the data and ensure the reliability of the data, the baseline subtraction method was applied to the raw data first. In order to eliminate the impact of the baseline on the data, baseline values were subtracted from sensor responses.

The method is as shown in Equation (2), where $R(t)$ value is the dynamic response value, $R(0)$ is the baseline value, which is the minimum sensor response value when exposed to a reference gas. That is, the preprocessed data is equal to the difference between the response value and the baseline value.

$$Y_S = R(t) - R(0) \tag{2}$$

5. Feature extraction

Feature extraction step helps to reduce the amount of redundant data from the original high dimensional gas sensor data set. The steady state region reflecting the sensing dynamics at the sensor surface is of special interest in this work. In other words, we followed the idea of utilizing the steady-state feature, which is the "gold-standard" for chemo-sensory feature extraction [18]. The steady state response feature, which is one of the most traditional features, only samples one data from the original response curve of each concentration cycle. But in our approach, we use 10 s response values of the steady state region, in order to utilize more information. The data used for classification is displayed in Figure 4.

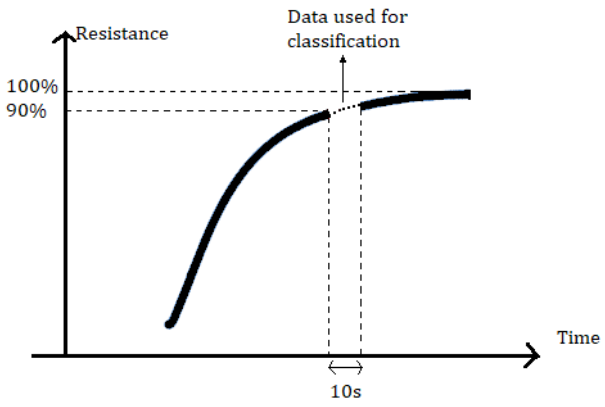

Figure 4. The feature extracted from each response curve

In order to obtain successful classification results, only a few seconds of sensor data after the response curve reaches 90% of its maximum value were directly used without taking into account the other information in the whole response curve. Since the data acquisition device recorded data with 2 seconds period, 6 data were taken from each concentration region during

10 seconds. Thus, for i-th gas type and its j-th concentration cycles taken data can be denoted as

$$R_{ij} = [r_{ij1}, r_{ij2}, r_{ij3}, r_{ij4}, r_{ij5}, r_{ij6}] \quad (3)$$

As there are 9 different concentrations varied from 50 ppm to 450 ppm for each gas, totally the number of samples in sub-data set for classification is $6 \times 6 \times 9 = 324$ (54 samples per class for six classes). Besides, another data set which consists of the maximum value of each concentration region was also generated for a comparison of classification results. Since there are 9 concentrations, 9 maximum values were used for each gas, and totally the data set has $6 \times 9 = 54$ samples.

6. Classification

Classification step addresses the problem of finding a relation between the sensor responses and the gas types using the 10 s responses in the steady state region. For each type of gas, the dataset consisted of 54 samples is randomly split into 60% training and 40% test sets. So, the training set contains 192 samples for six gas types, and the test set contains the remaining samples that were not used during the training process.

Due to the small amount of data, k-fold cross validation method was used to avoid overfitting. According to the method, the classifier is trained with k−1 splits and validated on the missing split. The method is performed k times until all of the data is used for validation. In our work, we have used 5-fold cross valdiation. In order to compare the performance of classification algorithms, classification accuracy was utilized as a performance metric. It is obtained by dividing the number of correctly recognized samples into the total number of samples. Furthermore, since the training set and test set are randomly selected from the original gas sensor dataset, we performed train-test procedure 10 times to avoid bias in the classification process. Then, the final classification accuracy of each classifier was calculated by averaging of ten iterations. Each classification algorithm was learned from the same training set and test its classification accuracy on the same test data. Hence, the best classification method is clearly the one with the highest accuracy.

The classifiers used in this work and their descriptions are given in Table 1, and the performance of these algorithms is evaluated through their accuracy, reported in Table 2. Results indicate that KNN algorithm achieves the highest classification accuracy among the considered classification algorithms. KNN algorithm is based on the idea of classifying a sample with unknown class by a plurality vote of its k nearest neighbor classes. The choice of the value for the parameter k and the distance metric are two parameters that affect the performance of the algorithm. In this work, 96.7% accuracy is obtained with KNN algorithm when the distance metric is chosen as euclidean (The value of k can be chosen as 1 or 10). Besides, an ensemble tree method also gave high classification accuracy, but KNN is faster, simpler and easier to implement.

Table 1. Classification algorithms used in this work.

Classifier	Description
KNN - 1	Preset: Weighted KNN Number of neighbors: 10 Distance metric: Euclidean Distance weight: Squared inverse Standardize data: true
KNN - 2	Preset: Fine KNN Number of neighbors: 1 Distance metric: Euclidean Distance weight: Equal Standardize data: true
Tree-1	Preset :Fine Tree Maximum number of splits: 100 Split criterion: Gini's diversity index Surrogate decision splits: Off
Tree-2	Preset: Coarse Tree Maximum number of splits: 4 Split criterion: Gini's diversity index Surrogate decision splits: Off
SVM-1	Preset: Fine Gaussian SVM Kernel function: Gaussian Kernel scale: 0.25 Box constraint level: 1 Multiclass method: One-vs-One Standardized data: true
SVM-2	Preset: Quadratic SVM Kernel function: Quadratic Kernel scale: Automatic Box constraint level: 1 Multiclass method: One-vs-One Standardized data: true
Ensemble - 1	Preset: Bagged Trees Ensemble method: Bag Learner type: Decision Tree Number of learners: 30
Ensemble - 2	Preset: Boosted Trees Ensemble method: AdaBoost Learner Type: Decision Tree Maximum number of splits: 20 Number of learners: 30 Learning rate: 0.1

Table 2. Classification performance of the algorithms.

Classifier	Classification Accuracy (%) (with data set including only steady state response)	Classification Accuracy (%) (with data set including 10 s responses from steady state region)
KNN - 1	9.1	96.2
KNN - 2	8.3	96.7
Tree -1	21.2	61.5
Tree - 2	25.4	38.1
SVM - 1	17	44.4
SVM - 2	22	30.6
Ensemble - 1	8.3	95.6
Ensemble - 2	8.3	73.6

The obtained classification results point out that using 10 s responses from the steady state region of the sensor curve gives much better accuracy when compared with traditional steady state response feature. As can be seen from Table 2, while steady state feature achieves maximum accuracy rate of 25.4%, the proposed feature method achieves maximum accuracy rate of 96.7%. The performance of KNN – 2 model with confusion matrices for train and test data is illustrated in Figure 5 and Figure 6.

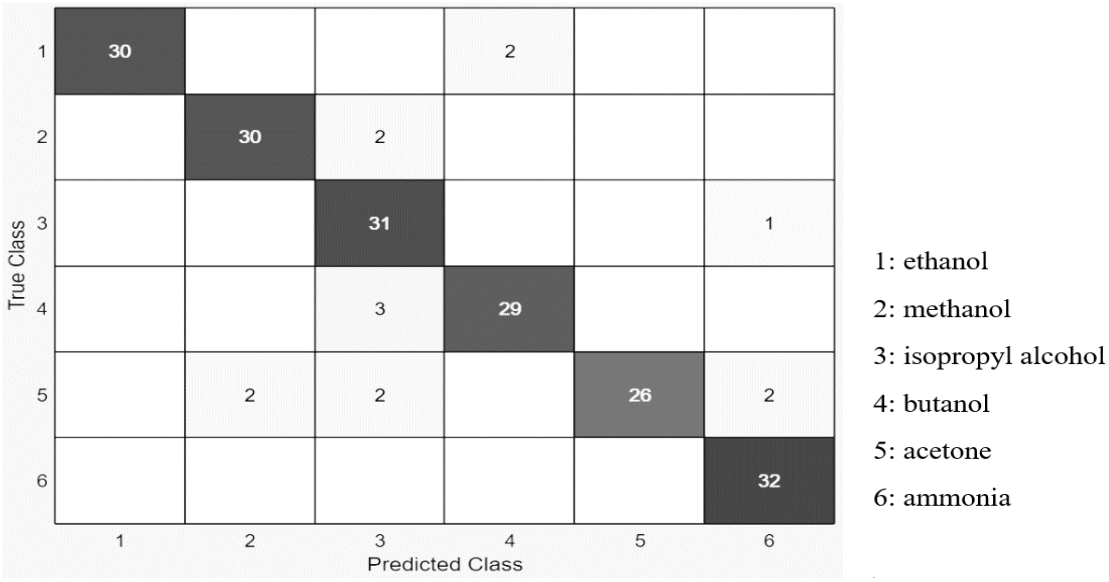

Figure 5. Confusion matrix of KNN - 2 model for train data

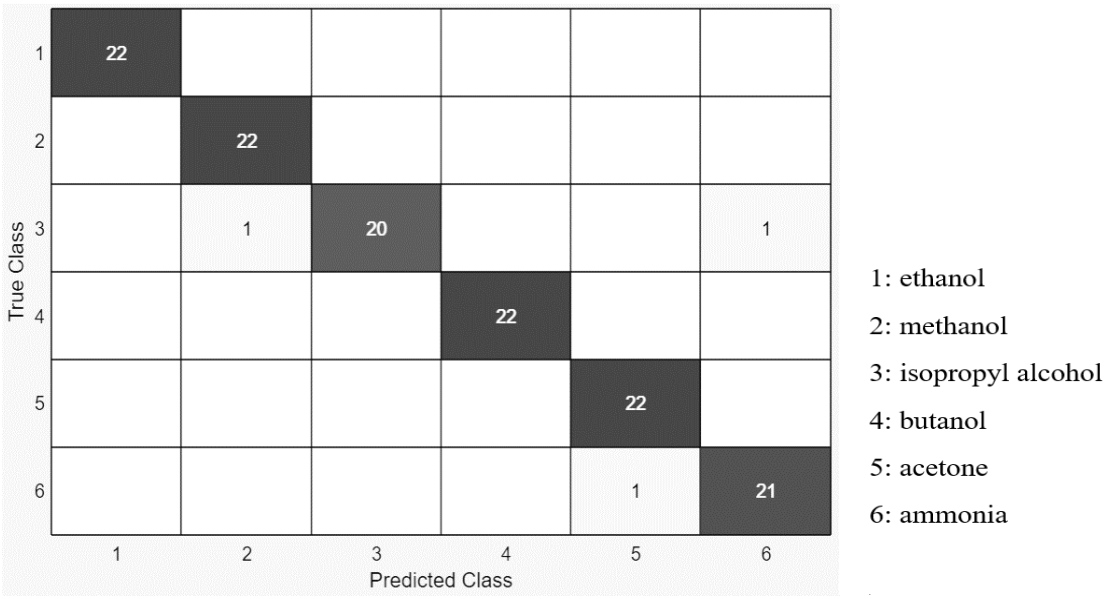

Figure 6. Confusion matrix of KNN - 2 model for test data

Besides, in order to obtain the best classification accuracy with KNN algorithm, the optimum k value should be found. For this purpose, the above mentioned KNN-1 and KNN-2 algorithms were run with different k values from 1 to 10. Since KNN-1 model is based on weighting, the value of k does not affect the accuracy of the algorithm. Classification accuracy results for different k values for KNN-2 model, which is based on euclidean metric without weightening, are illustrated in Figure 7. It is observed that as the number of k values increases, the accuracy value decreases.

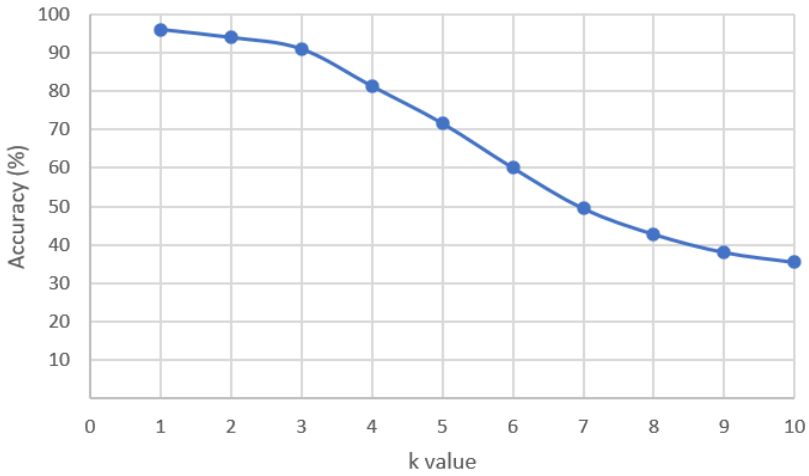

Figure 7. Accuracy of the KNN model for different k values

7. Conclusion

In this work, the sensing performance of a resistive type gas sensor based on CoPc thin film was investigated with respect to six different VOC vapors (methanol, ethanol, butanol, isopropyl alcohol, acetone, and ammonia) concentrations from 50 to 450 ppm. In order to solve

the issue of selectivity while using a single gas sensor, machine learning algorithms were applied to the gas sensor data set. Results indicate that fast and high classification accuracy is obtained by using only 10 s data from the steady state region of the sensor curve. Among various classification algorithms; KNN, which is one of the fast and easy to implement algorithms, achieved the highest accuracy (96.7%), which is a notable achievement with a single resistive gas sensor. Besides, the performance of the feature proposed in this study was compared with the traditional steady state response feature, and it was found that the proposed feature provides much better classification accuracy. Therefore, it is shown that the most necessary information that helps to distinguish between the different types of VOC vapors can be gained not using only one response from the steady state region but also a few seconds of responses taken from the steady state region. As KNN is easy to implement and works very fast, it is one advantage of the proposed method in this study. In addition to this, whereas in many applications a single feature can not fully reflect the characteristics of sensor responses, in this work using only one feature based on a single sensor, high classification accuracy was obtained. In addition to this, though there are many researches based on acoustic wave sensors for the detection of the vapors investigated in the literature, the number of resistive based sensor studies we have introduced in this study is quite limited. As a result, our model has great potential in practical applications for solving the issue of selectivity while using a single gas sensor.

References

[1] Göl E. Y., Karabudak E., "Mini-review: “Ball-Type Phthalocyanines”: similarities and differences from mono phthalocyanines", Mini Reviews in Organic Chemistry 16 (2019) : 410-421.

[2] Van Keulen K. E., Jansen M.E., Schrauwen R. W. M., Kolkman J. J., Siersema P.D., "Volatile organic compounds in breath can serve as a non-invasive diagnostic biomarker for the detection of advanced adenomas and colorectal cancer", Alimentary Pharmacology and Therapeutics 51(3) (2020) : 334-346.

[3] Tripathi K. M., Kim T. Y., Losic D., Thanh Tung T., "Recent advances in engineered graphene and composites for detection of volatile organic compounds (VOCs) and non-invasive diseases diagnosis", Carbon 110 (2016) : 97-129.

[4] Saalberg Y., Wolff M., "VOC breath biomarkers in lung cancer", Clinica Chimica Acta 459 (2016) : 5-9.

[5] Fend R., Bessant C., Williams A. J., Woodman A. C., "Monitoring haemodialysis using electronic nose and chemometrics", Biosensors and Bioelectronics 19 (2003) : 1581-1590.

[6] Singh Bhati V., Hojamberdiev M., Kumar M., "Enhanced sensing performance of ZnO nanostructures-based gas sensors: A review", Energy Reports 6 (2020) : 46–62.

[7] Ridhi R., Saini G. S. S., Tripathi S. K., "Sensing of volatile organic compounds by copper phthalocyanine thin films", Materials Research Express 4 (2017) : 025102.

[8] Wanga B., Li Z., Zuoa X., Wu Y., Wang X., Chen Z., He C., Duan W., Gao J., "Preparation, characterization and NO2-sensing properties of octa-iso-pentyloxyphthalocyanine lead spin-coating films", Sensors and Actuators B 149 (2010) : 362–367.

[9] Altındal A., Kurt Ö., Şengül A., Bekaroğlu Ö., "Kinetics of CO2 adsorption on ball-type dicopper phthalocyaninethin film", Sensors and Actuators B 202 (2014) : 373–381.
[10] Ağırtaş M. S., Altındal A., Salih B., Saydam S., Bekaroğlu Ö., "Synthesis, characterization, and electrochemical and electrical properties of novel mono and ball-type metallophthalocyanines with four 9,9-bis(4-hydroxyphenyl)fluorine", Dalton Trans. 40 (2011) : 3315–3324.
[11] Yang R. D., Gredig T., Colesniuc C. N., Park J., Schuller I. K., Trogler W. C., Kummel A. C., "Ultrathin organic transistors for chemical sensing", Applied Physics Letters 90 (2007) : 263506.
[12] Yan J., Guo X., Duan S., Jia P., Wang L., Peng C., Zhang S., "Electronic nose feature extraction methods: A Review", Sensors 15 (2015) : 27804-27831.
[13] Li H., Luo D., Sun Y., GholamHosseini H., "Classification and identification of industrial gases based on electronic nose technology", Sensors 19 (2019) : 5033.
[14] Jia P., Tian F., He Q., Fan S., Liu J., Yang S. X., "Feature extraction of wound infection data for electronic nose based on a novel weighted KPCA", Sensors and Actuators B : Chemical 201 (2014) : 555–566.
[15] Kong C., Zhao S., Weng X., Liu C., Guan R., Chang Z., "Weighted Summation: Feature extraction of farm pigsty data for electronic nose", IEEE Access 7 (2019) : 96732–96742.
[16] Ngo K. A., Lauque P., Aguir K., "Identification of toxic gases using steady-state and transient responses of gas sensor array", Sensors and Materials 18 (2006) : 251-260.
[17] Abdurrahmanoğlu Ş., Altındal A., Bulut M., Bekaroğlu Ö., "Synthesis and electrical properties of novel supramolecular octa-phthalocyaninato-dicobalt(II)-hexazinc(II) and dicobalt(II)-dimeric-phthalocyanine with six ferrocenylimin pendant groups", Polyhedron 25 (2006) : 3639–3646.
[18] Llobet E., Brezmes J., Vilanova X., Sueiras J. E., Correig X., "Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick-film tin oxide gas sensor array", Sensors and Actuators B: Chemical 41 (1–3) (1997) : 13–21.