Classified Volatile Organic Compound Detection using Data Classification Algorithms

Jaya Prakash Chennoju
Department of ECE, Koneru Lakshmaiah Education Foundation Deemed to be University, India | Department of ECE, Sir C. R. Reddy College of Engineering, India
jp.chennoju@gmail.com (corresponding author)

Nalluri Siddiah
Department of ECE, Koneru Lakshmaiah Education Foundation Deemed to be University, India
nalluri.siddu@kluniversity.in

Received: 16 October 2023 | Revised: 5 November 2023 and 18 November 2023 | Accepted: 29 November 2023

 DOI: https://doi.org/10.48084/etasr.6531

ABSTRACT

Sensors are becoming smaller and less expensive, sparking interest in assessing vast volumes of sensor data. Meanwhile, the emergence of machine learning has led to the development of technologies that have a substantial impact on our lives. Machine learning models are often used to produce accurate, real-time predictions even in the presence of noisy sensed data. In this study, a Volatile Organic Compound (VOC) categorization system based on sensor data collected from a sensor array was developed. The most difficult challenge posed in the sensor array was the detection of the type of VOC. It is feasible to categorize VOCs brought on by applying data classification algorithms to data collected from sensor devices. In this work, we used data from the classification algorithms Decision Tree (DT), Naive Bayes (NB), and Linear Regression (LR) on a developed linear sensor array and their classification accuracy was compared. Four different VOCs were evaluated: acetone (C_3H_6O), benzene (C_6H_6), ethanol (C_2H_5OH), and toluene ($C_6H_5CH_3$). The acquired classification accuracy reached 95.65% with the LR algorithm.

Keywords-cantilever; classified detection; MEMS; machine learning; sensor; VOC

I. INTRODUCTION

Sensor system applications growth in a variety of fields, including conservation and development, climate motoring, and anthropogenic detection, has been made possible, to a substantial degree, by developments that have taken place over the years, specifically in recent times. It is common knowledge that sensor-based systems are application-dependent [1]. This also comprises the top part, which is responsible for handling the data in an effective and constructive approach. As the number and scale of deployed sensor networks rises, so does the amount of data gathered, necessitating the use of specialized methodologies that can handle this scale while still meeting application requirements. In this research, the authors demonstrate how sophisticated Machine Learning (ML) algorithms [2] might be utilized to analyze autonomously obtained sensor data in conjunction with manually gathered data in order to make predictions for a diversity of occurrences. Our presentation is based on the data acquired from an array-modeled sensor that was designed and developed for detection of most hazardous Volatile Organic Compounds (VOCs) [3]. This research concentrates on ML and data mining for the interpretation of sensor data as a component of comprehensive system integration. This implementation ranges from hardware at the bottom level all the way up to data-driven ML algorithms at the topmost level. The proposed system is an example of applying ML methods on sensor data, which can provide high-level guidelines for similar applications and involve predictions based on sensor data.

II. RELATED WORK

The number of vertical system implementations is not very large, but a similar research was conducted in [4]. The authors describe a networked sensor architecture designed to enable human interaction. This architecture is made up of equipment often found in an office setting, such as personal computers, Personal Digital Assistants (PDAs), telephones, etc. In order to construct the training set, each activity observed by the monitoring devices must first be manually labeled. This strategy is different from the one we've been using in terms of the sensors involved. Instead of a vertical system, the focus is given on the Bayesian learning approach. Table I shows a review of the recent literature on VOC detection. The research in [12] pertains to technology implementation in sensor nodes. Sensor data are modeled for queries in a variety of contexts via the use of RDF (Resource Description Framework) and RDQL (RDF Data Query Language) query languages, with a few tweaks here and there. In spite of this, the dataset was acquired by simulating a sensor network, which highlights the system performance as well as the programming language. Because of this, the system performance in a real-time environment may

differ. An alternative technique is presented in [13, 14], in which all the detection systems are modeled employing Dynamic Conditional Random Fields (DCRFs) to evaluate the real-world, in which sensor information may be distorted, impacted by noise, or lost. This allows for a more accurate representation of the environment [15, 16].

TABLE I. LITERATURE REVIEW

Year	Reference	Methodology	Application	Limitation
1998	[5]	MEMS cantilever arrays with functionalized coatings	Environmental monitoring, industrial safety	Limited selectivity, cross-sensitivity to other compounds
2007	[6]	MEMS cantilevers combined with gas chromatography	Air quality assessment, chemical process control	Complex setup, expensive instrumentation
2010	[7]	MEMS cantilever arrays with nanomaterial functionalization	Indoor air quality, breath analysis	Short functionalization lifespan, drift over time
2013	[8]	Multiple MEMS cantilevers with differential readout	Indoor air quality, leak detection	Cross-sensitivity to humidity, temperature variations
2016	[9]	MEMS cantilever arrays with integrated microheaters	Environmental monitoring, explosive detection	Energy consumption, slow response time
2019	[10	Compact MEMS cantilever array-based sensor	Personal exposure monitoring, wearable devices	Limited dynamic range, potential interference from background odors
2022	[11]	MEMS cantilevers with machine learning algorithms	Smart buildings, precision agriculture	Data interpretation challenges, need for continuous calibration

III. METHODOLOGY

Figure 1 demonstrates the system components, which are: the sensors and actuators, a server for collecting sensor data, a humanoid element that is instrumental in developing extra data, a database that contains the additional data, data preparation tools, and a ML toolbox. In order to collect the dataset, we must first get the raw data from the sensor array that are located on the sensor node, and then send this data to a computer so as to be stored. In the next step, we will combine the automatically collected data with the data labeled by hand so that the ML algorithms can be trained using the pre-processing techniques.

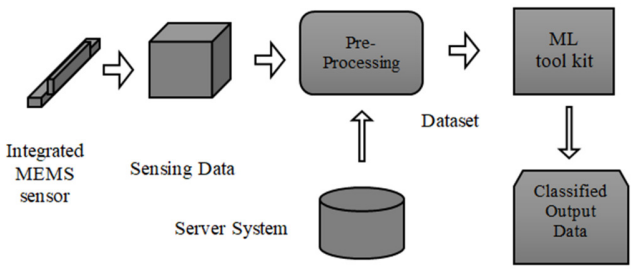

Fig. 1. Basic components of the proposed system.

As per the characteristics of fabricated sensor [17], Figure 2 shows that the computer is wired to the RS232 to USB converter, which is a part of the integrated sensor network. The sensor node also includes a 3 V power supply, display screen, and an ATmega328P microcontroller. When the sensor is exposed to respective reactants, a reaction takes place on cantilever surface and provides output voltage. Then, the microcontroller collects data from the integrated sensors at a sample rate of 500 ms, packs the data into a vector, and then transmits them to a server through the serial connection so that it will be possible for them to be retained. The data may also be observed via a serial monitor.

During the course of the research, the collected data were kept in a word document, with each sample including a time stamp along with the numerical values obtained from the sensor array. Each time sensor array is able to identify one of the 4 VOCs but the user does know the exact classifier from the range of the output values. This issue is tackled by introducing data classifiers between the sensor output and the system display. Minimum, maximum, and mean values of the sensor data are shown in Table II, for all 4 VOCs in terms of resistance change of the cantilever surface. The values have not significant differences. Since there are no extreme results, it is possible to say that the observations are correct.

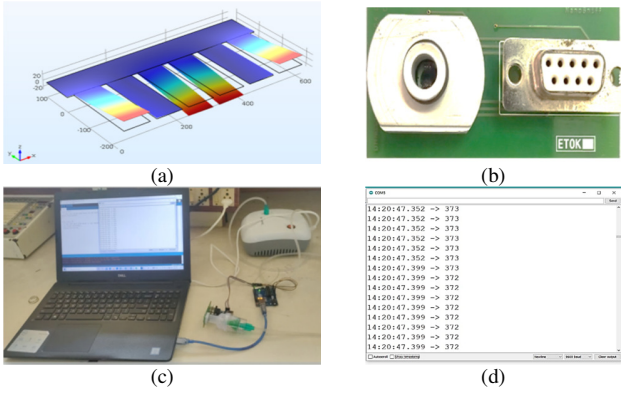

Fig. 2. Experimental setup to identify resistance range of the array cantilever: (a) Cantilever deflection when its is exposed to VOC, (b) encapsulated cantilever, (c) practical setup interfaced with the system, (d) terminal output.

TABLE II. MINIMUM AND MAXIMUM VALUES OF RESISTANCE CHANGE FOR EACH VOC

VOC	Min.	Max.	Mean
Acetone (C_3H_6O)	53.3 KΩ	53.6 KΩ	53.4 KΩ
Benzene (C_6H_6)	55.2 KΩ	55.8 KΩ	55.6 KΩ
Ethanol (C_2H_5OH)	56.4 KΩ	56.6 KΩ	56.5 KΩ
Toluene ($C_6H_5CH_3$)	55.5 KΩ	55.9 KΩ	55.7 KΩ

The illustration of a scenario showing the way sensor data progressed in the laboratory during the period of experimentation is shown in Figure 3. It can be seen that when the integrated sensor was exposed to the reactants, after the

transient state, the values increased with time and became steady after a few seconds. This can be noticed in Figure 3(a), representing the sensor response in the idle case. A negative transient indicates fluctuation due to surge voltage when the power is ON. With respect to time, it becomes stable in the steady state.

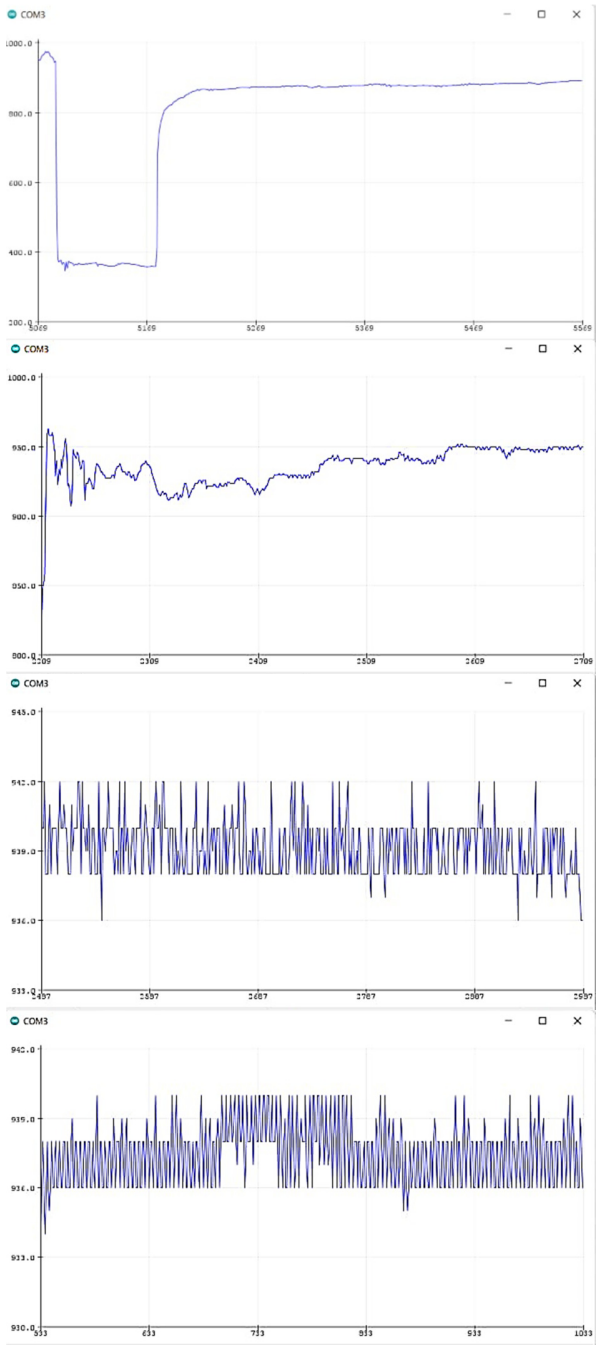

Fig. 3. Response of the array cantilever from the initial state to the steady state. (a) Idle sensor, (b) sensor exposed to a proportional reactant, (c) response of the sensor from the transient to the steady state, (d) steady state response of a sensor with a significant change in amplitude.

Figure 3(b) depicts the result of a sensor exposed to a proportional reactant. Significant reaction takes place between the analyte and the reactant, and as the sensor is more sensitive in nature, a few glitches occurred in the transient state. Figure 3(c) illustrates the sensor responce from the transient to the steady state. The observed glitches have low frequencies. Similarly, Figure 3(d) shows the steady state sensor response with significant change in amplitude.

A. Data Processing and Learning

In the first stage of data processing, sensor information and physically obtained data were aligned using time. The initial sampling rate for the sensor information was set to 500 ms, whereas the time stamp for the physically obtained data was simply included with the hours and minutes of the input. We allocated the matching instance from the extra acquired sensor data occurred within the same minute for each cantilever sensor of the array modelled sensor. So, given that time stamps on the manually gathered data are only in HH:MM:SS format, the second step was to reduce the dimensionality problem. The first thing we did was to choose a sample rate of 100 ms. However, several sensor data characteristics, such as the non-clean room ambient environment (light intensity, air pressure, and dust particulates), show significant fluctuations over the instant of experimentation, and these variations cannot be accurately correlated with any other data. The time stamps from the two different data sources may differ by a few seconds due to the fact that extra data were added by a greater number of samples and no time synchronization was performed. We also removed data based on the difference of successive samples collected throughout the experimentation, since this enabled us to avoid having an excessive number of redundant samples. To be more specific, the data collected throughout the experiment did not include any repeated values, and thus incorporating all these data would have resulted in an extremely imbalanced dataset. Due to the fact that it is impossible to precisely correlate sensor data with manually gathered data and that humans make mistakes, the produced dataset includes some inaccurate occurrences. On the other hand, there are no gaps in the data and no missing values in the dataset.

The learning process utilized a dataset that contained a total of 20286 measurements, each of which has 6 attributes: Deflection, stress, change in resistance, temperature, moisture, ambient pressure. We used two different learning methods, i.e. classification and regression. The first technique is applied for the prediction of classified class labels, while the second approach approximates the target variable by modeling a continuous-valued function (class attribute). Since the target variable in our context might be interpreted both as a discrete-valued variable and as a numerical variable, we came to the conclusion that it would be best to evaluate both approaches. DT and the NB classifiers were used as classification techniques. The first one offers a clear visual explanation of the findings, whereas the second one makes more effective use of the characteristics of the dataset as a whole. We decided to go with the standard method of LR for regression analysis [18].

IV. RESULT INTERPRETATION AND EVALUATION

Aiming to test the predictive capability of each method, we ran experiments on two different datasets: one with just sensor data characteristics (deflection Voltage, change in resistance, stress, and ambient pressure) and the other with a class attribute with extra arbitrarily inserted input. Two distinct ML algorithms were chosen to classify data: the J48 algorithm for learning DTs and NB allowed us to evaluate how various learning approaches performed on the datasets. For the regression data analysis, we used a procedure known as conventional LR. The WEKA toolbox [19] provided the execution of these computations. It is necessary to examine how well the algorithms worked with both the basic and its supplemented dataset. The Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) were engaged to measure how accurate the forecasts are. Although MAE and RMSE give more emphasis on how significant the disparity between the expected and observed values is, the classification accuracy was additionally provided for each classification method.

A. Decision Tree

After conducting a number of initial experiments, we decided to implement the limitation that representation must include at least 10% of the total number of instances. All these DTs [20] feature the cantilever deflection response measurements property in the cluster head. This allows for simple classification of a huge observations set when sensors exposed to other than the specified VOCs lead to a low value (less than or equal to 56.7 K). The sensor's cantilever resistance change and output voltage are the two parameters utilized for categorization when looking at the basic dataset. Additionally, the topology of the DT was altered as a result of new data input. In this case, only one piece of the new data is used, which is the total number of computers that are working. Considering data overfitting, we chose the Naive Bayesian learning approach [21, 22]. This method employs probability distribution over a set of variables. If the minimum number of occurrences in a branch is raised, it could lead to data overfitting.

B. The Bayesian Network

In an attempt to implement the Naive Bayesian method, we configured the WEKA toolkit with the settings of simple estimator and K2 search algorithm. Tables III-IV provide the confusion matrices for two distinct scenarios involving simple and enhanced datasets. In the case of the supplemented dataset, there is a clearer contrast between occurrences that include zero deflection response and the instances that contain more than non-zero deflection response. When considering the class property to have only two possible values—Zero (0) and Non-Zero (!=0) (in terms of the response), we are able to determine whether or not our system is suitable for VOC detection. We carried out Cost-Benefit Analysis (CBA). The results demonstrated increased predictive performance on the supplemented dataset up to 86.52%, in contrast to the 84% of the cases properly identified when there were 4 categories (class attribute values). The findings are shown in Table V, which also includes a representation of the cost and confusion matrices.

TABLE III. CONFUSION MATRIX OF THE SIMPLE DATASET

	0	1	2	3
0	7096	120	100	11
1	95	2988	891	218
2	37	833	2566	253
3	1	246	492	640

TABLE IV. CONFUSION MATRIX OF THE AUGMENTED DATASET

	0	1	2	3
0	7073	208	45	1
1	38	3085	933	173
2	0	557	2731	401
3	0	16	378	985

TABLE V. COST BENEFIT ANALYSIS

Cost Matrix		0	1	**Confusion Matrix**	0	1
	0	0.0	1.0		7018	250
	1	5.0	0.0		21	8045
Acc.	**86.52%**					

C. Linear Regression

A quite uncomplicated LR algorithm [23-25] was employed. As a result, each response property in the supplemented dataset was converted to 3 binary characteristics (with values of 0 or 1) pertaining to each of the nominal values.

From Tables VI-VIII, the accuracy of 95.65% indicates that the model is performing extraordinarily well. According to the cost matrix, false negatives (1 misclassified as 0) have a larger cost than false positives (0 misclassified as 1). This indicates that the model has been optimized to minimize false negatives, which may be appropriate for applications where missing a positive instance is more expensive. The confusion matrix reveals that there are much more true positives (9156) than false positives (290), exhibiting that the model is performing well in properly detecting positive events. There are also much more true negatives (7256) than false negatives (28), indicating that the binary classification model is working well. The resultant linear model that will be utilized for prediction is depicted in Figure 4(a). On the supplemented data, the features with the highest coefficients are the appropriate sensor responses, displaying that these factors have a favorable influence on the decisions that can be anticipated. As mentioned above, the model has a very high overall accuracy of 95.65%. Understanding the specific domain and application context (sensor environment), on the other hand, is crucial in assessing if these levels of accuracy and misclassification costs are appropriate. The trade-off between false positives and false negatives is determined by the application's specific goals and requirements. The model works well, especially when it comes to correctly categorizing positive occurrences. The cost matrix was selected with the intention of reducing false negatives.

Three machine learning models (DT, NB, and LR) were tested on two datasets, i.e. Simple and Augmented (Table IX). For all models, the augmented dataset outperforms the simple dataset, indicating that more data or new attributes improve the prediction accuracy and reduce errors. LR outperforms the other ML models in both datasets, with NB coming second. LR

yields the highest accuracy, whereas NB yields the lowest MAE and RMSE values. It should be noted that further fine-tuning may increase performance.

TABLE VI. CONFUSION MATRIX OF THE SIMPLE DATASET

	0	1	2	3
0	8093	302	52	2
1	48	3086	895	183
2	0	568	2731	411
3	0	19	478	980

TABLE VII. CONFUSION MATRIX OF THE AUGMENTED DATASET

	0	1	2	3
0	8096	150	120	12
1	95	3988	903	260
2	36	560	2831	453
3	1	18	522	680

TABLE VIII. COST BENEFIT ANALYSIS

Cost Matrix		0	1	**Confusion Matrix**	0	1
	0	0.0	1.0		7256	290
	1	5.0	0.0		28	9156
Acc.	**95.65%**					

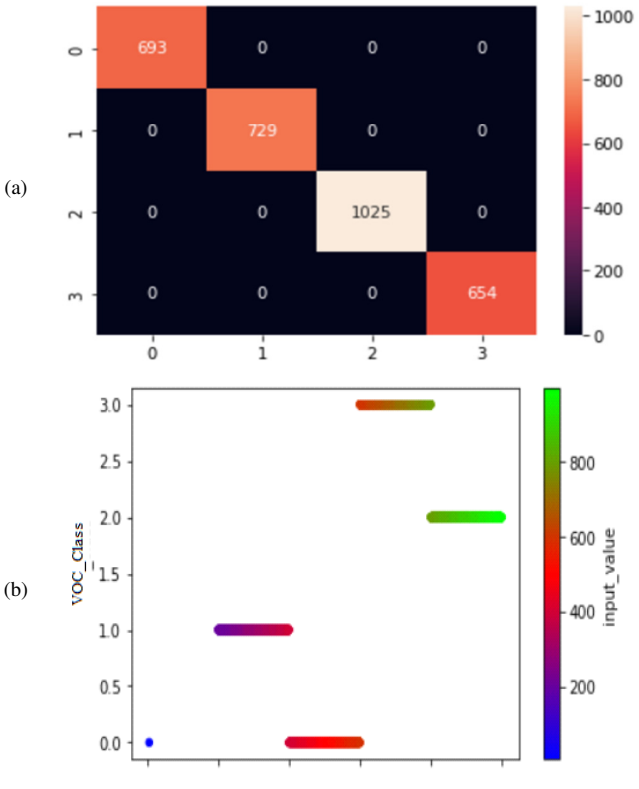

Fig. 4. (a) Classification report of the test data, (b) classified output from the modeled biosensor array.

V. CONCLUSION AND FUTURE WORK

In this study, we introduced a hierarchical system integration method for VOC classification based on sensor results. We created an augmented dataset by labelling sensor data with additional data, and then used machine learning algorithms to learn from that dataset. The examination of the obtained predicted results from both the basic and the enriched datasets proved that effective VOC classification can be forecast when using sensor data. The prediction can be improved by giving any of the three machine learning techniques new information. The application and the anticipated results have a role in the decision of which machine learning algorithm to apply. On the basis of our data, in comparison with decision trees and Naive Bayes, linear regression produce better results, however, more trials on larger datasets are needed to draw definitive conclusions.

This study findings are promising for the continuous development of the system, which will include the creation of a network of sensors in order to get more information. Also, the possibility of adding semantic technologies to the current system to enhance the data and make predictions both of which will be more accurate and more varied will be examined.

REFERENCES

[1] D. V. Dao, K. Nakamura, T. T. Bui, and S. Sugiyama, "Micro/nano-mechanical sensors and actuators based on SOI-MEMS technology," *Advances in Natural Sciences: Nanoscience and Nanotechnology*, vol. 1, no. 1, Mar. 2010, Art. no. 013001, https://doi.org/10.1088/2043-6254/1/1/013001.

[2] F. Mlawa, E. Mkoba, and N. Mduma, "A Machine Learning Model for detecting Covid-19 Misinformation in Swahili Language," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10856–10860, Jun. 2023, https://doi.org/10.48084/etasr.5636.

[3] G. Ciuti, L. Ricotti, A. Menciassi, and P. Dario, "MEMS Sensor Technologies for Human Centred Applications in Healthcare, Physical Activities, Safety and Environmental Sensing: A Review on Research Activities in Italy," *Sensors*, vol. 15, no. 3, pp. 6441–6468, Mar. 2015, https://doi.org/10.3390/s150306441.

[4] S. Gupta, K. Ramesh, S. Ahmed, and V. Kakkar, "Lab-on-Chip Technology: A Review on Design Trends and Future Scope in Biomedical Applications," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 5, pp. 311–322, Oct. 2016, https://doi.org/10.14257/ijbsbt.2016.8.5.28.

[5] Y. Bao, P. Xu, S. Cai, H. Yu, and X. Li, "Detection of volatile-organic-compounds (VOCs) in solution using cantilever-based gas sensors," *Talanta*, vol. 182, pp. 148–155, May 2018, https://doi.org/10.1016/j.talanta.2018.01.086.

[6] Y. Dong, W. Gao, Q. Zhou, Y. Zheng, and Z. You, "Characterization of the gas sensors based on polymer-coated resonant microcantilevers for the detection of volatile organic compounds," *Analytica Chimica Acta*, vol. 671, no. 1, pp. 85–91, Jun. 2010, https://doi.org/10.1016/j.aca.2010.05.007.

[7] N. Siddaiah, V. A. S. Tentu, and Z. Rehman, "Design, Simulation and Performance Analysis of Novel Cantilever Rf-Mems Switch Using Serpentine Meanders," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 4, pp. 1360–1366, Apr. 2019.

[8] M. Katta, S. Parri, M. Vamsi, K. Allu, and K. Lavanya, "Simulation Approach to Design High Sensitive Nems Based Sensor for Molecular Bio- Sensing Applications," *European Journal of Molecular & Clinical Medicine*, vol. 8, no. 3, pp. 1730–1738, Nov. 2021.

[9] N. Siddaiah, A. Pujitha, G. J. Sai, U. Gupta, and C. Chaitanya, "Sensitivity Enhancement and Optimization of Mems Piezoresistive Microcantilever Sensor for Ultra Mass Detection," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 7S, pp. 137–142, 2019.

[10] B. Najar, I. Marchioni, B. Ruffoni, A. Copetta, L. Pistelli, and L. Pistelli, "Volatilomic Analysis of Four Edible Flowers from Agastache Genus,"

Molecules, vol. 24, no. 24, Dec. 2019, Art. no. 4480, https://doi.org/10.3390/molecules24244480.

[11] M. Hodgson, H. Levin, and P. Wolkoff, "Volatile organic compounds and indoor air," *Journal of Allergy and Clinical Immunology*, vol. 94, no. 2, Part 2, pp. 296–303, Aug. 1994, https://doi.org/10.1053/ai.1994.v94.a56008.

[12] R. Bashir, "BioMEMS: state-of-the-art in detection, opportunities and prospects," *Advanced Drug Delivery Reviews*, vol. 56, no. 11, pp. 1565–1586, Sep. 2004, https://doi.org/10.1016/j.addr.2004.03.002.

[13] Y. J. Chen, "Distinct advantages and novel applications of BioMEMS," 2013.

[14] D. Doufene, S. Benharat, S. Bouazabia, and S. A. Bessedik, "Hybrid Grey Wolf and Finite Element Method (GWO-FEM) Algorithm for Enhancing High Voltage Insulator String Performance in Wet Pollution Conditions," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8765–8771, Jun. 2022, https://doi.org/10.48084/etasr.4978.

[15] G. L. Cote, R. M. Lec, and M. V. Pishko, "Emerging biomedical sensing technologies and their applications," *IEEE Sensors Journal*, vol. 3, no. 3, pp. 251–266, Jun. 2003, https://doi.org/10.1109/JSEN.2003.814656.

[16] S. M. Ho, "Fabrication of Cu4SnS4 Thin Films: A Review," *Thin Films*, vol. 10, no. 5, pp. 6161–6164, Oct. 2020, https://doi.org/10.48084/etasr.3663.

[17] C. Jayaprakash and N. Siddaiah, "Sensitivity analysis of nems cantilever to detect volatile organic compounds using finite element method: DOI: 10.48129/kjs.20501," *Kuwait Journal of Science*, vol. 50, no. 3A, Jun. 2023, https://doi.org/10.48129/kjs.20501.

[18] M. Katta and R. Sandanalakshmi, "A Technology Overview and Future Scope of Bio-Mems in Tropical Disease Detection: Review," *International Journal of Engineering & Technology*, vol. 7, no. 3.12, pp. 648–651, Jul. 2018, https://doi.org/10.14419/ijet.v7i3.12.16446.

[19] E. Frank *et al.*, "Weka-A Machine Learning Workbench for Data Mining," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Boston, MA, USA: Springer US, 2010, pp. 1269–1277.

[20] C. Zhang, C. Hu, S. Xie, and S. Cao, "Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation," *Journal of Physics: Conference Series*, vol. 1732, no. 1, Jan. 2021, Art. no. 012086, https://doi.org/10.1088/1742-6596/1732/1/012086.

[21] A. Moraru, M. Pesko, M. Porcius, C. Fortuna, and D. Mladenic, "Using Machine Learning on Sensor Data," *Journal of Computing and Information Technology*, vol. 18, no. 4, 2010, Art. no. 341, https://doi.org/10.2498/cit.1001913.

[22] X. Li *et al.*, "Integrated MEMS/NEMS Resonant Cantilevers for Ultrasensitive Biological Detection," *Journal of Sensors*, vol. 2009, Jun. 2009, Art. no. e637874, https://doi.org/10.1155/2009/637874.

[23] L. Yan, Z. Wang, Y. Liu, and Z. Ye, "Generic and Automatic Markov Random Field-Based Registration for Multimodal Remote Sensing Image Using Grayscale and Gradient Information," *Remote Sensing*, vol. 10, no. 8, Aug. 2018, Art. no. 1228, https://doi.org/10.3390/rs10081228.

[24] D. Ä. G. Ärzteblatt Redaktion Deutsches, "Linear Regression Analysis," *Deutsches Ärzteblatt*, Nov. 2010, https://doi.org/10.3238/arztebl.2010.0776.

[25] H. Basarudin *et al.*, "Evaluation of Climate Change Effects on Rain Rate Distribution in Malaysia using Hydro-Estimator for 5G and Microwave Links," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11064–11069, Aug. 2023, https://doi.org/10.48084/etasr.5552.