Machine Learning-Based Classification on Gas Sensor Device Test-Data to Identify Presence of Toxic CO and CO_2 using SVM, RF, and KNN

Indranil Maity
Department of Electronics and Communication Engineering (ECE) Institute of Engineering and Management (IEM), University of Engineering and Management (UEM)
Kolkata, West Bengal, India
E-mail: indrasanu026@gmail.com

Tanit Ghosh
Department of Electronics and Communication Engineering (ECE) Institute of Engineering and Management (IEM), University of Engineering and Management (UEM)
Kolkata, West Bengal, India
E-mail: tanitgh24@gmail.com

Sneha Pandey
Department of Electronics and Communication Engineering (ECE) Institute of Engineering and Management (IEM), University of Engineering and Management (UEM)
Kolkata, West Bengal, India
E-mail: snehapandey835@gmail.com

Aheli Majumdar
Department of Electronics and Communication Engineering (ECE) Institute of Engineering and Management (IEM)
Kolkata, West Bengal, India
E-mail: majumdar.aheli@gmail.com

Malay Gangopadhyay
Department of Electronics and Communication Engineering (ECE) Institute of Engineering and Management (IEM), University of Engineering and Management (UEM)
Kolkata, West Bengal, India
E-mail: malay.ganguly@iem.edu.in

***Abstract*—This paper describes a machine learning (ML) driven approach for the discrimination of two poisonous gases, namely Carbon Monoxide (CO) and Carbon Dioxide (CO_2). Proper detection of them is essential as both are toxic and harmful environmental gases, which leads to several diseases. Supervised machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF) and k-Nearest Neighbors (k-NN) classifier were used here to analyse the sensor responses from Platinum (Pt)-doped Zinc Oxide (ZnO) based nanotube (NT) sensor, and distinguish between the two adsorbed gases. The dataset was divided into the ratio of 70:30. The code was written in Google Colab, which supports python 3 environment to train, verify, and test the models. The features used for model training include adsorption energy ($E_{adsorption}$), binding distance, the highest occupied molecular orbital's energy (E_{HOMO}), the lowest un-occupied molecular orbital's energy (E_{LUMO}), chemical potential (μ), ratio of electronic conductivity, sensitivity, and recovery time. The target variable indicated the gas type (CO or CO_2). Principal component analysis (PCA) helped to visualize the data by reducing its dimension. A total of 4 quality performance metrics, namely precision, along with accuracy, F1 score, and recall were calculated along with the confusion matrices. Overall, SVM achieved the best accuracy of 96.67%.**

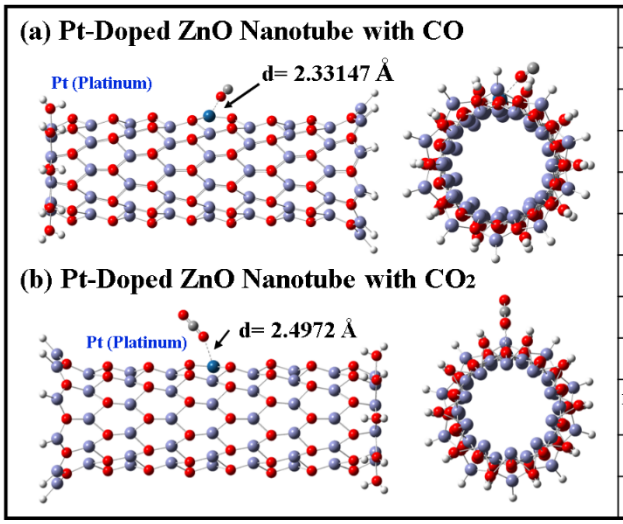
Keywords—machine learning, principal component analysis, support vector machine, k-nearest neighbor, random forest

I. INTRODUCTION

Carbon monoxide (CO) is a poisonous gas that is colorless, odorless, and tasteless [1]. It is hazardous to the environment and can be extremely harmful to human health when present at elevated concentrations [2]. Carbon dioxide (CO_2), although generally harmless at low levels, can become dangerous when its concentration increases in closed or poorly ventilated spaces. Both CO and CO_2 are significant air pollutants commonly encountered in industrial operations, household environments, and broader environmental settings. Exposure to elevated concentrations can pose serious health risks and may lead to life-threatening situations [3-4]. Therefore, dependable and efficient detection of these gases is critical for ensuring safety, preventing hazardous incidents, and reducing their potential impact on human health and the environment [4]. ZnO NTs are a prominent class of nanostructures with considerable potential in various sensing applications, owing to their unique chemical and physical characteristics [5]. Given their high surface-to-volume ratio, chemical stability, and adaptable electrical characteristics, ZnO NTs are great candidates for gas sensors [6]. Foreign atom doping alters the electronic characteristics, boosting the surface reactivity of the ZnO NT [7]. Pt is a corrosion-resistant dopant, which possesses catalytic properties and is chemically inert so it does not react with the target absorbant.

An extensive variety of machine learning and statistical methods have been investigated in previous research for forecasting CO and CO_2 emissions. Tian et al. [8] studied the application of multiple regression-based models for predicting CO_2 levels and identifying the suitable algorithms for environmental forecasting. Surbhi Kumari and Sunil Kumar Singh [9] studied the emission rate of CO_2

979-8-3315-6339-4/26/$31.00 ©2026 IEEE



Parameters	CO	CO_2
$E_{adsorption}$ (eV)	-0.7382	-0.343706
Binding Distance (Å)	2.331	2.49719
E_{HOMO} (eV)	-8.9607	-9.11989
E_{LUMO} (eV)	-5.9614	-6.1568
HOMO LUMO gap (eV)	2.9993	2.96309
E_{HOMO-1} (eV)	-9.207788108	-9.125881842
E_{LUMO+1} (eV)	-5.941333576	-5.966912277
Chemical Potential (μ)	7.461	7.638345
Electronic Conductivity Ratio	58.352	57.20250965
Sensitivity (%)	96.9	74.09796099
Recovery Time (s)	2.98	579836.7808 (ps)

Fig. 1. Structural and Electronic properties of Pt-doped ZnO Nanotubes [15] under CO, CO_2 gas interaction; used as the parameters for the ML test-train split

and predicted trends for the next 10 years, using machine learning and deep learning models. Further, Ebru Koca Akkaya and Ali Volkan Akkaya [10], repeatedly pointed out that accurate CO_2 prediction is a key factor for energy planning, directing emission reduction strategies, and global sustainability. Li et al. [11] analyzed different ML techniques including ordinary least squares regression, support vector machines, and gradient boosting, for forecasting emissions related to transportation. Moreover, Ajala et al. [12] investigated different model performances involving 14 models, including various statistical methods like ARMA, ARIMA, SARMA, and SARIMA, with daily CO_2 data from the most polluted areas worldwide as input. Moreover, Xiangqian Li and Xiaoxiao Zhang [13] narrowed down the focus to China with the goal of finding the best near-real-time forecasting model, using univariate daily emissions data from recent days. Using a refined SFS method with Light GBM, Shi et al. [14] found that just nine non-QC features were enough to build a highly accurate prediction model. This streamlined model was then used to reliably estimate CO adsorption energies, key electronic properties, and the overall stability of all layered alloys. The purpose of the paper is to execute a comparative analysis among three supervised ML algorithms on the basis of their accuracy, when applied to the dataset collected from Pt-ZnO NT gas sensor. As a comparative study, it showcases how precisely each model is able to distinguish between the target gas labels (CO and CO_2), using the provided sensor features. The dataset was formed from the readings obtained by the Pt-ZnO NT based sensor device. For a guaranteed fair and accurate evaluation, the dataset was partitioned into a 70:30 ratio (70% for training, 30% for testing). The dimensionality of the dataset was then reduced using PCA, and trained again for plotting the decision boundary. The models were developed and tested using python libraries such as NumPy, Scikit-learn, Pandas and Matplotlib.

II. Classification Model Evaluation

A. Sensor Information

The sensor utilized to build the dataset in this work was a Pt-ZnO NT structure, optimized in Gaussian 09W software. The Pt-doped ZnO had one Zn atom replaced by Pt, which formed a stable catalytic site for the interaction of CO. 11 electronic parameters were extracted from this sensor ($E_{adsorption}$, binding distance, E_{HOMO}, E_{LUMO}, μ, electronic conductivity ratio, sensitivity, and recovery time to name a few [15]) as mentioned in Fig. 1, which were later used as features to train the algorithms.

B. Support Vector Machine

SVM, a maximum margin model, is primarily utilized for regression and classification tasks, created in the 1990's by Vladimir Vapnik [16]. SVM aims to find the best hyperplane which discriminates between the classes. The working mechanism of SVM is that it selects the hyperplane with the maximum class separation, which is the distance between the support vectors [17]. Fig. 2, showcases the algorithm for implementing SVM. The cost function of the algorithm is provided in equation (1),

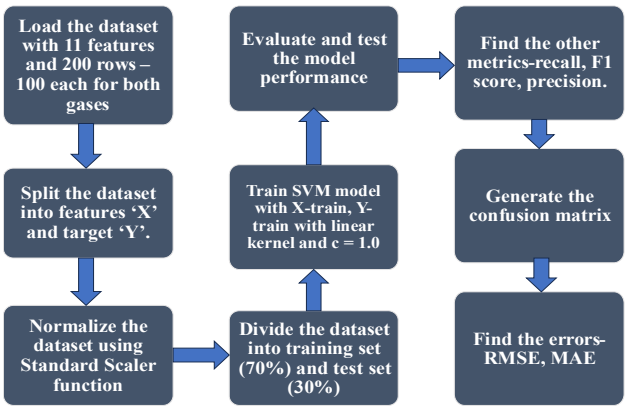


Fig. 2. Flow diagram of algorithm used in SVM model

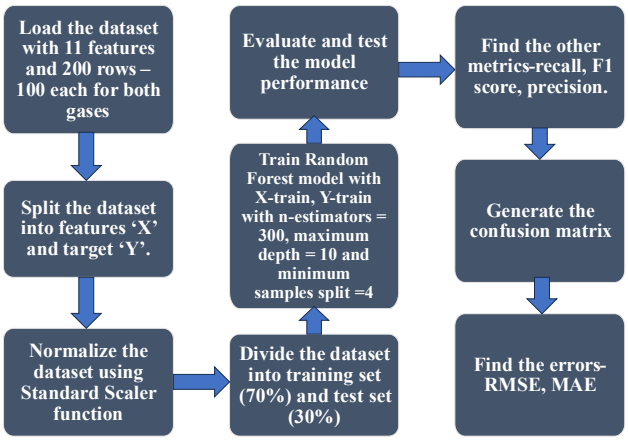


Fig. 3. Flow diagram of algorithm used in Random Forest model

where w is the weighted vector, b is the intercept, ξ is the slack variable for mis-classification ($\xi > 1$ represents misclassification), and c is the regularization parameter.

$$Cost\ Function = \min_{(w,b,\xi)} \frac{1}{.2}\|w\|^2 + c\sum_{i=1}^{N}\xi_i \qquad (1)$$

C. Random Forest

The random forest algorithm, as shown in Fig. 3, is an ensemble supervised machine learning algorithm, introduced by Leo Breiman in 2001, to improve accuracy and reduce overfitting. It uses the principle of decision trees, by using multiple of them during training and aggregating their outputs to produce a better and reliable prediction when compared with a single decision tree, which can be prone to high variance [18]. The algorithm creates each decision tree by using a technique known as bootstrap sampling. It is implemented by applying simple random sampling with replacement (SRWR) to the dataset, thus making every tree unique. Instead of using all the features at once, it picks a random subset of features to decide how to split the data. Then every tree provides their own prediction based on what it learned from the part of the data provided to it. The Gini index is given in equation (2) as:

$$Gini = 1 - \sum_{i=1}^{n}(P_i)^2 \qquad (2)$$

Each decision tree uses either Gini Index or Entropy as a metric for impurity measurement, where P_i is the probability of class i. The maximum value for gini index is 0.5 depicting impure split, whereas for entropy it is 1.

D. k-Nearest Neighbors

k-NN, a supervised ML algorithm, devised in the early years of pattern recognition research. It works on the philosophy that data points with similar characteristics appear close to each other in a feature space, and likely belong to the same category [19]. It is non-parametric and instance based in nature, i.e., it does not make a training model, but instead stores the entire dataset during prediction [20]. The Euclidean distance (d) is given in equation (3) as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

k-NN considers the k number of nearest datasets (neighbors) to the test point and the majority class found from the neighbors is the predicted output for the test point. To identify the nearest neighbors, it calculates the distance typically using the Euclidean Distance metric, as provided in equation (3). Fig. 4, shows the algorithm for the implementation of the k-nearest neighbor model for classification.

III. METHODOLOGY

The dataset was formed from the sensor response data provided by the Pt-ZnO NT based sensor device, designed for the detection of toxic CO and CO_2. The eleven features taken into consideration were, recovery time, $E_{adsorption}$, electronic conductivity ratio, binding distance, E_{HOMO}, E_{LUMO}, E_{HOMO-1}, E_{LUMO+1}, HOMO-LUMO gap, chemical potential (μ) and sensitivity, and target labels were the gases CO and CO_2. A total of 200 rows of data were used. The dataset was cleaned and verified to ensure that missing values were not present. The labels were assigned as 0 for CO and 1 for CO_2. The ML models captured the mapping between the input parameters the and output labels, and learned from it. Once trained, the models could predict the gases for new unseen sensor readings. Machine learning algorithms used were SVM, RF and k-NN. The library used to implement the models was scikit-learn. The kernel selected to classify the gases in SVM was linear after testing all other kernels, and the regularization parameter used was c = 1.0. For RF, the n-estimators was 300 and maximum depth was assigned as 10. The value for k in k-NN was 5 after hyperparameter tuning. The library used to implement the support vector classifier was scikit-learn Support Vector Classifier module. PCA was applied for dimension reduction and visualizion of the data. The regularization parameter used was c=1.0, to

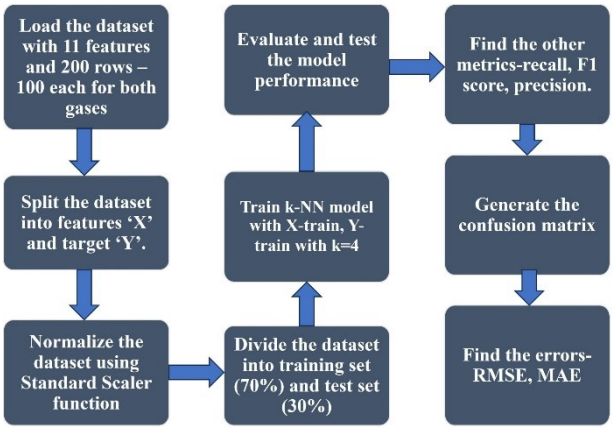


Fig. 4. Flow diagram of algorithm used in k-NN model

minimize the error. The model investigated the accuracy provided by the techniques and analysed their classification capabilities. Sensor responses for CO and CO_2 may show variations and are influenced by physiochemical factors. These algorithms were used as they are capable of learning and identifying these complicated patterns and non linear trends, and can deal with high dimensional features, which can then be visualised using PCA for better understanding.

The code was written in Google Colab supporting python 3 programming language. The dataset was divided into training and testing subsets in the ratio 70:30. To visualize the data, PCA was used [21]. Standard scalar function was used to scale the data [22], then PCA was utilized to visualize the higher dimensional data into 2-Dimensional (2D) data. These two principal components captured the maximum variance in a 2D representation. The ratio of PCA1 : PCA2 for RF, k-NN and SVM are obtained to be the same as [0.24508241 0.10788231].

IV. Results And Discussions

In Table I, it was observed that from the performance parameters, SVM attained the best results in all calculated metrics and for both gases, with k-NN performing slightly worse, followed by the RF classifier. Fig. 5, showcasing the error metrics calculated for each of the classification algorithms, further provided similar results, where SVM provided the least amount of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), both followed by RF and k-NN, which provided equal errors at the 70% testing data set. With the help of the confusion matrix, different performance analysis parameters were calculated, namely precision, which stresses on the quality of the correctly predicted values and gives information about the reliability of the model, and recall, which focuses on the coverage of the model telling how many CO or CO_2 labels were successfully identified. The F1 score is obtained by combining the precision and recall into one metric using their harmonic mean. Fig. 6(a-c) represents the confusion matrices for the 3 ML models which helps to analyze the effectiveness of the classification models. The confusion matrix is divided into four different

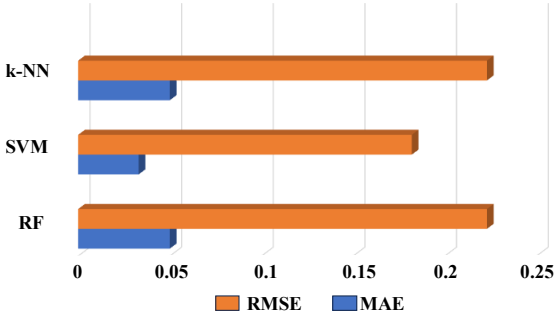


Fig. 5. Bar graph representation of the generated errors among k-NN, RF and SVM models

TABLE I. Distinction of the ML Algorithms on the Basis of Performance Metrics

Metrics	Random Forest		Support Vector Machine		k-Nearest Neighbor	
Gas	***CO***	***CO_2***	***CO***	***CO_2***	***CO***	***CO_2***
Recall	0.97	0.90	1.0	0.93	1.0	0.90
Precision	0.91	0.96	0.94	1.0	0.91	1.0
F1-score	0.94	0.93	0.97	0.97	0.95	0.95

quadrants, which are true negative (TN), false negative (FN), true positive (TP) and false positive (FP). TN and TP provide the correctly predicted CO_2 and CO cases and, FN and FP show the wrongly predicted CO_2 and CO cases

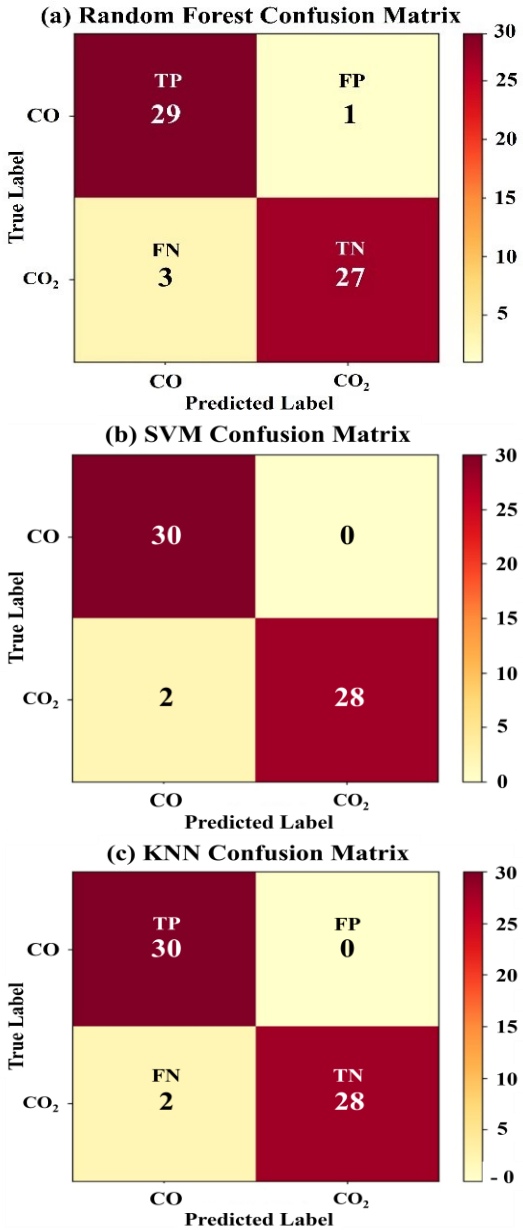


Fig. 6. Heatmap of the confusion matrix obtained by (a) the RF model, (b) the SVM model, and (c) the k-NN model

TABLE II. ACCURACY AND EXECUTION TIME FOR ML MODELS ON 80, 70, AND 60% TRAINING SET

TRAINED OVER	80%		70%		60%	
ALGORITHM	*ACCURACY (%)*	*TIME (S)*	*ACCURACY (%)*	*TIME (S)*	*ACCURACY(%)*	*TIME (S)*
RF	92.5	17.32	93.34	17.711	88.75	12.59
SVM	95.0	9.65	96.67	8.44	95.0	8.81
K-NN	95.0	13.74	95.0	11.25	90.0	9.75

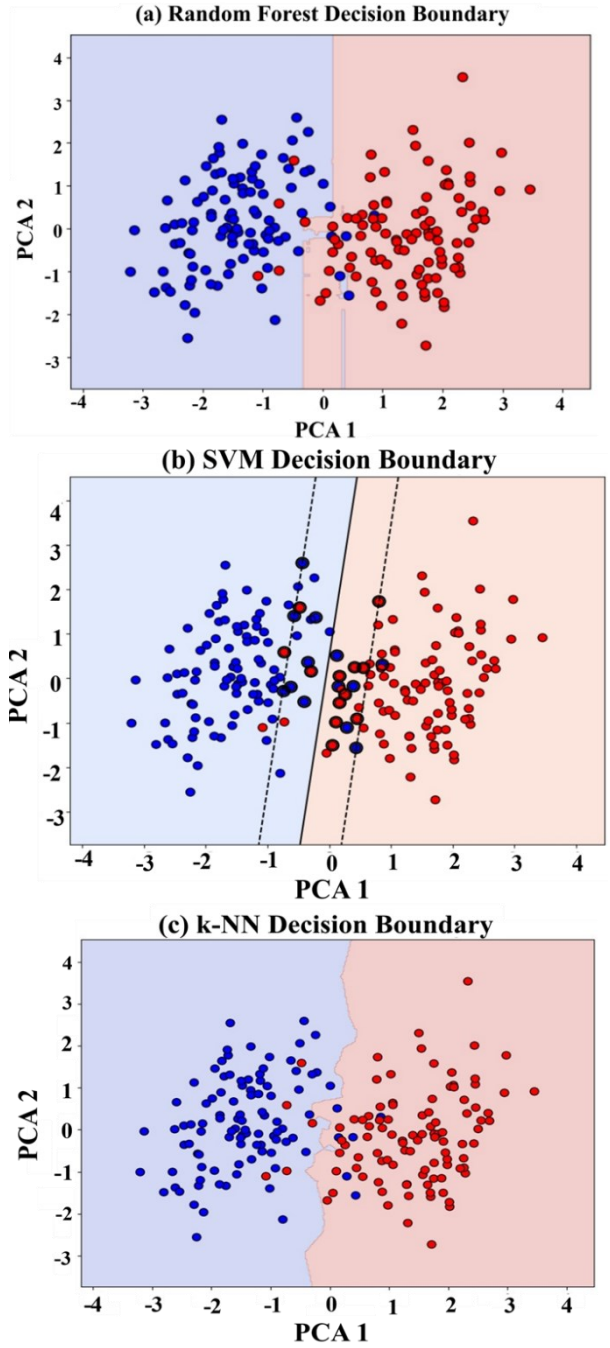


Fig. 7. Scatter plot of PCA applied testing data set of (a) Random Forest, (b) SVM, and (c) k-NN

respectively. The provided heatmap in Fig. 6(a-c)gives a visual representation for highlighting TP, TN, FP and FN [23]. In Fig. 7(a-c), the decision boundary graphs were plotted for all the three algorithms. It visualizes how accurately the classifier separates the two classes- CO and CO_2, using PCA. The random forest produces a mostly vertical, piecewise boundary showing that it forms several linear splits to separate the two classes. It illustrates a clear division between CO (blue) and CO_2 (red), with minor misclassification caused by natural randomness of ensemble. On the other hand, SVM performs the best and provides the cleanest and smoothest separation by maximizing the margin between CO and CO_2. The hyperplane is almost linear, indicating the best performance among the three algorithms. The k-NN algorithm also produced a flexible decision boundary bending according to the local patterns, though being more noise susceptible.

Table II depicts the comparison of accuracy and time taken for execution among different percentages of samples used in training the SVM, Random Forest, and k-Nearest Neighbors classifiers. It was observed that, out of the 3 training to test set ratios, the 70% training data set obtained the most accuracy. Further, for SVM, the execution time with the 70% training set was the lowest and thus the best overall. It was also noticed that SVM achieved the better accuracy and the fastest execution speed in all fields provided, thus it can be concluded that SVM is the best performing algorithm.

V. CONCLUSION

This paper conducts a comparative evaluation of the ML based classification of CO and CO_2 gases adsorbed on Pt-ZnO NT based sensor devices. The algorithms used were, Support Vector Machine, k-Nearest Neighbors and Random Forest. The dataset was divided into multiple training sets of 60, 70 and 80%, where the 70% training was found to be the best with an accuracy of 93.34%, 96.67% and 95.0% for RF, SVM and k-NN, respectively. PCA was used, where two principal components were selected for a 2D visualization, with maximum variance ratio being [0.24508241 0.10788231] for the ML models. SVM recorded the least MAE and RMSE of 0.034 and 0.1826, respectively, relative to RF and k-NN. Overall, SVM performed the best because it was robust

to high dimensional feature spaces, and handled imbalance of classes using margin maximization effectively, whereas the k-NN's performance dropped because of the factor of dimensionality. Similarly, the accuracy of RF was found to be lowered due to the overfitting, caused by the small, highly corelated sensor dataset.

ACKNOWLEDGMENT

The authors sincerely thank and express their gratitude to the Institute of Engineering & Management, Kolkata, Department of ECE for their constant help and encouragement. The authors also thank Siddhartha Bhattacharya and Ankit Biswas (IEM, Kolkata) for their valuable support.

REFERENCES

[1] L. D. Prockop and R. I. Chichkova, "Carbon monoxide intoxication: An updated review," Journal of the Neurological Sciences, vol. 262, no. 1–2, pp. 122–130, Nov. 2007.

[2] J. Beheshtian, Z. Bagheri, M. Kamfiroozi, and A. Ahmadi, "Toxic CO detection by B12N12 nanocluster," Microelectronics Journal, vol. 42, no. 12, pp. 1400–1403, 2011.

[3] N. Lalmuanchhana, B. Lalroliana, R. C. Tiwari, N. Lalhriatzuala, and R. Madaka, "Transition metal decorated ZnO monolayer for CO and NO sensing: A DFT + U study with vdW correction," Applied Surface Science, vol. 604, p. 154570, Dec. 2022.

[4] I. Maity, A. Majumdar, S. Maity, and I. Maity, "An in-depth analytical approach to unfold material-level insights in pristine and platinum-doped $MoSe_2$ nanosheet during carbon monoxide detection," Journal of Active & Passive Electronic Devices, vol. 19, no. 3, p. 217, 2025.

[5] R. Saad et al., "Fabrication of ZnO/CNTs for application in CO_2 sensor at room temperature," Nanomaterials, vol. 11, no. 11, p. 3087, Nov. 2021.

[6] Jigang et al., "Effect of platinum on the sensing performance of ZnO nanocluster to CO gas," Solid State Communications, vol. 316, p.113954, 2020.

[7] I. Maity, S. Maity, and S. Bhattacharya, "Development of foreign atom decorated ZnO nanotube based sensor devices for accurate detection of NH_3 gas leakage," Microsystem Technologies, vol. 31, pp. 3679–3694, 2025.

[8] L. Tian, Z. Zhang, Z. He, C. Yuan, Y. Xie, K. Zhang, and R. Jing, "Predicting energy-based CO_2 emissions in the United States using machine learning: A path toward mitigating climate change," *2025*, pp. 1–23, Mar. 2025.

[9] S. Kumari and S. K. Singh, "Machine learning-based time series models for effective CO_2 emission prediction in India," Environmental Science and Pollution Research, vol. 30, pp. 116601–116616, Nov. 2023.

[10] E. Koca Akkaya and A. V. Akkaya, "Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon dioxide emissions," Environmental Science and Pollution Research, vol. 30, pp. 122381–122392 2023.

[11] X. Li, A. Ren, and Q. Li, "Exploring patterns of transportation-related CO_2 emissions using machine learning methods," Sustainability, vol. 14, no. 8, p. 4588, 2022.

[12] A. A. Ajala, O. L. Adeoye, O. M. Salami, and A. Y. Jimoh, "An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models," Environmental Science and Pollution Research, vol. 32, pp. 2510–2535, 2025.

[13] X. Li and X. Zhang, "A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China," Environmental Science and Pollution Research, vol. 30, pp. 117485–117502, 2023.

[14] T.-T. Shi, G.-Y. Liu, and Z.-X. Chen, "Machine learning prediction of CO adsorption energies and properties of layered alloys using an improved feature selection algorithm," The Journal of Physical Chemistry C, vol. 127, no. 20, pp. 9573–9583, May 2023,

[15] I. Maity, S. Bhattacharya, and A. Majumdar, "Improvement in Toxic CO Detection Capabilities via Incorporating Pt Dopant into ZnO Nanotube Based Gas Sensor Devices: An Atomistic Modeling," 2024 IEEE Electron Devices Kolkata Conference (EDKCON), Kolkata, India, pp. 393–398, Nov. 2024.

[16] S. Gupta, R. Kambli, S. Wagh, and F. Kazi, "Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework," IEEE Transactions on Industrial Electronics, vol. 62, no. 4, pp. 2478–2486, Apr. 2015.

[17] I. Maity, S. Bhattacharya and S. Maity, "An Efficient Approach to Improve Diagnostic Accuracy into Findings of Breast Cancerous Cells with SVM and Logistic Regression Models," 2024 4th International Conference on Computer, Communication, Control & Information Technology (C3IT*)*, Hooghly, India, 2024.

[18] A. Parmar, R. Katariya, and V. Patel, "A review on Random Forest: An ensemble classifier," Lecture Notes on Data Engineering and Communications Technologies, pp. 758–763, Dec. 2018.

[19] I. Maity, A. Majumdar and S. Maity, "Investigation on PCA and LDA Feature Extraction Algorithms with KNN Classifier Targetting to Improve Diagnostic Accuracy into Detection of Breast Cancer Malignancy," 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 2025.

[20] M. Suyal and P. Goyal, "A review on analysis of K-nearest neighbor classification machine learning algorithms based on supervised learning," International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 43–48, 2022.

[21] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu, "Data analysis using principal component analysis," 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), Greater Noida, India, pp. 45-48, 2014.

[22] P. T. Reiss, J. Goldsmith, H. L. Shang, and R. T. Ogden, "Methods for scalar-on-function regression," International Statistical Review, vol. 85, pp. 228–249, 2017.

[23] L. F. S. Siqueira, R. F. A. Júnior, A. A. De Araújo, C. L. M. Morais, and K. M. G. Lima, "LDA vs. QDA for FT-MIR prostate cancer tissue classification", Chemometrics and Intelligent Laboratory Systems, vol. 162, pp. 123–129, 2017.