Submitted to *INFORMS Journal on Computing*
manuscript (Please, provide the manuscript number!)

Paving the Way for Cancer Detection: A Sensor-Based E-Nose Approach in Clinical Research

Premal Doshi
School of Information Systems, College of Business, University of South Florida, Tampa, FL 33620, USA, premaldoshi@usf.edu,

Arash Takshi
Department of Electrical Engineering, College of Engineering, University of South Florida, Tampa, FL 33620, USA., atakshi@usf.edu,

Ehsan Sheybani
School of Information Systems, College of Business, University of South Florida, Tampa, FL 33620, USA, sheybani@usf.edu,

In this study, we propose an innovative sensor-based approach for the early detection of diseases, focusing on the potential future application in cancer detection. Utilizing an electronic nose (e-nose) system, this method offers a low-cost, highly accurate, and rapid solution that could significantly enhance early diagnosis capabilities in clinical settings. While our current application does not include cancer detection, the e-nose technology, with its fast response time and precision, holds promise for future integration into cancer diagnostics, potentially saving lives and reducing costs for hospitals and insurance companies. We explore the capabilities of machine learning models for classification and estimation tasks, which are crucial in various industries, necessitating accurate and reliable methodologies. We introduce an innovative approach using the Gradient Boosting Machine (GBM) algorithm, leveraging sensor data for enhanced classification and estimation. To mitigate overfitting and enhance model generalization, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized to generate synthetic samples. A rigorous feature selection process was employed, resulting in a highly accurate model achieving 99.22% accuracy. Additionally, we propose a method to estimate the proportion and concentration of components in mixtures based on sensor data fusion. Our study also investigates the concept of optimized sensor combinations, achieving comparable accuracy with fewer sensors, thus emphasizing the importance of feature selection in enhancing cost efficiency without compromising performance. Overall, our research underscores the significance of advanced machine learning techniques in classification and estimation tasks, demonstrating their utility in diverse real-world scenarios.

Key words: Gas classification; Concentration estimation; Synthetic Minority Over-sampling Technique (SMOTE); affordable solutions; Machine Learning; Gradient Boosting Machines (GBM); Hyperparameter Tuning; E-Nose

1. Introduction.

Early detection of diseases is crucial for effective treatment and improved patient outcomes, particularly in the case of cancer. Interest in diagnosing cancer using volatile organic compounds (VOCs) found in exhaled breath has been growing[Hanna et al. (2019),Farraia et al. (2019)]. Human exhaled breath encompasses a diverse array of over 3000 volatile organic compounds (VOCs) in the gaseous state, which can be identified through advanced laboratory techniques such as gas chromatography and mass spectrometry. Among these exhaled VOCs are molecules including alkanes, benzene derivatives, acetone, dimethyl sulfide, phenol, and various aromatic compounds[Dragonieri et al. (2017)].The composition of these VOCs has been observed to undergo alterations in a growing spectrum of medical conditions, including cancers[Gordon et al. (1985),Behera et al. (2019)].Recent investigations have revealed that specific volatile organic compounds (VOCs), such as isopropanol, acetone, pentane, and benzene, can function as biomarkers for lung cancer [Machado et al. (2005),Phillips et al. (2008)]. To date, an array of electronic nose (eNose) models has been employed to distinguish the unique breathprints of lung cancer patients from those of healthy individuals [Machado et al. (2005),D'Amico et al. (2010),Dragonieri et al. (2009),Mazzone et al. (2007),Hakim et al. (2011)]. These promising results were corroborated by robust and reproducible data across multiple research groups, demonstrating the ability of eNoses to differentiate lung cancer from other respiratory diseases [Machado et al. (2005),Dragonieri et al. (2009),Natale et al. (2003),Tran et al. (2010),McWilliams et al. (2015)]. Consequently, these findings suggest that eNoses hold potential as clinical screening tools for early detection and diagnosis.

Traditional diagnostic methods, while effective, are often expensive, time-consuming, and require highly specialized equipment and personnel. In response to these challenges, we explore the potential of an electronic nose (e-nose) system as a novel, cost-effective, and accurate approach for early disease detection. Traditional methods for VOC analysis often rely on expensive and cumbersome equipment, limiting their practicality for widespread deployment.The most precise method for identifying VOCs combines gas chromatography with mass spectrometry (GC-MS), offering high accuracy and selective detection of individual VOCs[Langford et al. (2014)].However, GC-MS is time-consuming, costly, and requires skilled professionals to operate[Wilson (2015),Krilaviciute et al. (2015),Sun et al. (2016)].

In recent years, advancements in sensor technology and machine learning techniques have paved the way for more efficient and cost-effective approaches to gas sensing.

The e-nose technology leverages sensor data to identify volatile organic compounds (VOCs) associated with various diseases, providing a fast and reliable diagnostic tool.An electronic nose (eNose), as delineated by prior research, is a sophisticated instrument comprising an array of electronic chemical sensors coupled with an advanced pattern-recognition system, adept at discerning both simple and intricate odors[Gardner and Bartlett (1994)]. eNoses can identify complex olfactory patterns by juxtaposing the incoming scent with previously acquired patterns[Persaud and Dodd (1982)], thereby generating distinctive breathprints. The detection mechanism operates when volatile organic compounds (VOCs) interact with the sensor surfaces of the eNose, inducing alterations in the sensors' conductivity[van de Goor et al. (2018)]. These changes are subsequently detected by transducers and converted into electrical signals, which form unique VOC signatures[Gardner and Bartlett (1994)].

Over the past few decades, numerous pattern recognition algorithms have emerged for gas classification. In addition to chemical methods and advancements in interdisciplinary technologies, various Artificial Intelligence (AI) techniques have been reported in the literature for gas detection. Numerous machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machines (SVM), have been proposed for this purpose[Yin et al. (2015)]. In prior investigations, efforts have been made to discern the existence of either of two gases, with the capability to discern a blend of the two. Moreover, concentration estimation has been conducted for a volatile organic compound (VOC) composed of methanol and acetone, spanning from 40 to 400 ppm and 22 to 220 ppm, respectively[Khalaf et al. (2008)]. In order to attain a high recognition rate, a combination of sensors with diverse selectivity patterns is employed, necessitating the integration of pattern recognition techniques with the sensor array[Pardo and Sberveglieri (2005)]. Notably, for the electronic nose applications, the K-Nearest Neighbor (KNN) method was introduced[Gutierrez-Osuna et al. (2003),Yang et al. (2016)]. While simple and effective, this method necessitates the storage of training data, rendering it unsuitable for memory-constrained scenarios. In response to this limitation, a Gaussian Mixture Model (GMM) approach was introduced for odor classification [Belhouari et al. (2005)]), offering a potential solution.[Liu et al. (2015)] proposed two network architectures, Deep

Belief Networks and Stacked Autoencoders, to extract abstract gas features from E-nose data, followed by Softmax classifiers constructed from these features. These methodologies employ sequential methods based directly on gas sensor data. However, relying solely on gas sensor-based detection and identification presents several challenges. Primarily, the proportion of gas in the air can be very low, making it difficult for standard gas sensors to identify specific gases, leading to false negatives or false positives and thus compromising detection accuracy. Additionally, low-cost sensors tend to be less sensitive, potentially resulting in inaccurate measurements. Another method observed for gas detection is thermal imaging. Gas leakage often causes a temperature increase in the surrounding area, which can be detected and analyzed using thermal imaging cameras. This method has been applied in systems designed for Methane and Ethane gas leak detection [Hamilton and Charalambous (2020),Avila (2005),Marathe (2019)]. Jadin and Ghazali [Jadin and Ghazali (2014)] described a gas leak detection system using infrared image analysis, incorporating stages like data acquisition, image preprocessing, feature extraction, and classification. In a research they [Bilgera et al. (2018)] demonstrated a fusion of various AI models for Gas Source Localization using six different gas sensors to pinpoint leakage locations. [Pan et al. (2019)] introduced a deep learning approach utilizing a hybrid framework of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to extract sequential information from transient response curves. A rapid Gas Recognition algorithm based on a hybrid CNN and Recurrent Neural Network (RNN) is discussed in [Brahim-Belhouari et al. (2005)], showing that the fusion model surpasses Support Vector Machine (SVM), Random Forest, and k-nearest neighbors.

Although our current application does not extend to cancer detection, the e-nose system shows great promise for future use in this area. By integrating this technology into clinical practice, hospitals could achieve significant cost savings, streamline diagnostic processes, and ultimately improve patient care while reducing the financial burden on insurance companies. Our study focuses on recognizing various VOCs both individually and in mixtures, assessing each gas's signature or range, and detecting new occurrences of previously identified gases. Specifically, we investigated the tasks of VOC classification and concentration estimation using a Gradient Boosting Machine (GBM) model. Additionally, we used SMOTE to increase our training dataset by generating synthetic samples, while the test set consisted only of the actual experimental data. Through iterative feature selection, we

identified a subset of nine key features that significantly enhanced model accuracy while reducing complexity. The selected features—Nitric Oxide (NO), Total Volatile Organic Compounds (TVOC) (ppb), Ammonia (NH3)-Echem, Nitrogen Dioxide (NO2), Alcohol-Acetone, Alcohol, Ammonia (NH3)-Ethanol, Humidity (%), and Oxygen (O2) (%)—were instrumental in achieving an impressive accuracy of 99.22%.

2. Experimental Section

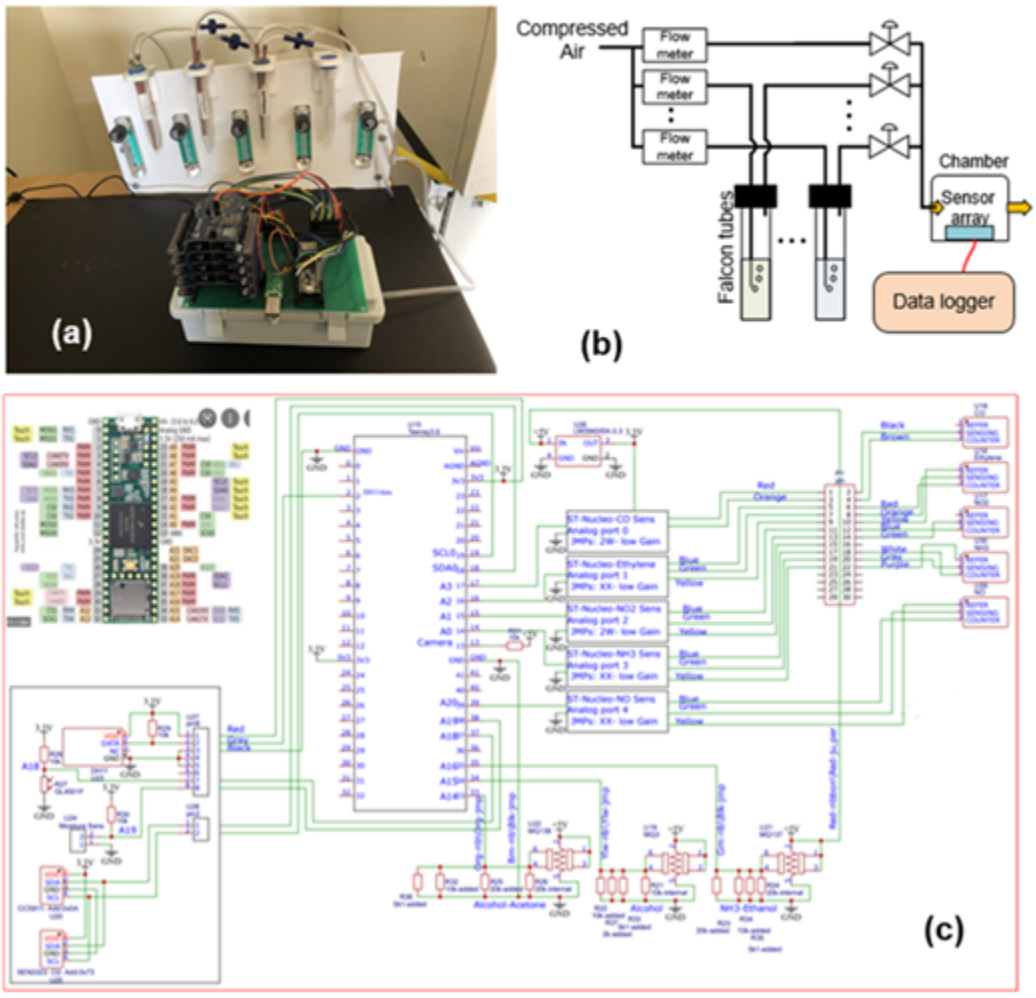

Figure 1 **(a) a picture and (b) a schematic of the testing setup. (c) Circuit diagram of the microcontroller system with the sensors and the signal conditioning sub-systems.**

Figure 1 shows a setup that was made for conducting the experiments. An IP65 ABS plastic sealed electric junction box (15x10x7 cm) was used as the chamber. 11 sensors were installed inside the box. Six of the sensors were electrochemical gas sensors: SGX-4NH3-1000 (ammonia sensor from SGX sensortech), SGX4NO-250 (NO sensor from SGX

sensortech), SGX-4NO2-2E (NO2 sensor from SGX sensortech), TGS5141 (CO sensor from Figaro), ME3-C2H4 (ethylene sensor from Winsen), and Gravity oxygen sensor (from Gravity). Also, three metal oxide sensors (MQ3 sensitive to alcohol, MQ137 sensitive to NH3 and ethanol, MQ138 sensitive to alcohol and acetone) were installed inside the box. Additionally, a humidity sensor equipped with a temperature sensor (DH11) and an air quality sensor (CCS811) were purchased from Amazon and placed in the box. To read the signals from the electrochemical sensors each electrochemical sensor (except the oxygen sensor) was connected to a P-NUCLEO-IKA02A1 (STMicroelectronics) board. The Gravity oxygen sensor and CCS811 were equipped with an I2C port. The metal oxide sensors were connected to an external resistor to form a voltage divider for reading the signals.Humidity and temperature fluctuations exert a substantial influence on the performance of a wide range of sensors[Khalaf et al. (2008)]. This observation, corroborated by previous studies, was also substantiated in our own investigation. A TeenSy 3.6 microcontroller was used for data logging on an SD card. The outputs of the signal conditioning boards were connected to different analog inputs or the I2C port of the microcontroller. The schematic of the circuit is shown in Figure 1. (c). The microcontroller was programmed to read the sensor signals every 5 seconds and save the data on the SD card.

A custom-designed setup was devised to subject the sensors to diverse volatile organic compounds (VOCs), illustrated in Figure 1. (a) and (b). This setup featured falcon tubes housing the testing VOCs, namely Isopropanol, ethanol, and Methanol. Within each falcon tube, two plastic tubes were inserted through the cap—one submerged in the VOC solution and the other positioned above the liquid surface.

Pressurized air, regulated by a flowmeter valve, was directed into the VOC solution, generating bubbles to facilitate VOC release. Simultaneously, a second tube above the liquid surface transported the gas laden with VOCs out of the falcon tube and into the chamber box. By adjusting the flowmeter, the rate of VOC insertion at varying levels could be controlled. For experimental purposes and simplicity in training, each experiment was conducted with only one VOC at a consistent rate. The actual concentration of VOCs was determined by measuring the consumption of liquids over a fixed period at a steady flow rate.

2.1. E-Nose

Constructed with an array of sensors, the device translates volatile organic compound (VOC) data into electronic signals. As gas samples permeate the sensor array, the odor molecules prompt reversible physicochemical alterations in the sensor materials, leading to shifts in electrical properties such as resistance and electrical potential.

3. Procedure for Obtaining Data Sets

Baseline Procedure:

1. Initiate the sensor device setup and allow it to operate for a duration of 5 minutes without the passage of compressed air or any other substances through the sensors. Subsequently, extract the collected data.

Procedure for Ethanol, Methanol, and Isopropanol:

Gas_Type	0.2L/min	0.5L/min	0.8L/min
Ethanol	200ppm	123ppm	161ppm
Methanol	292ppm	137ppm	164ppm
Isopropanol	217ppm	117ppm	143ppm

Table 1 Ethanol, Methanol, and Isopropanol at Different Flow Rates

1. Pour approximately 6mL of ethanol, methanol, or isopropanol into a test tube.
2. Affix the test tube to the setup.
3. Activate the air valve.
4. Adjust the flowmeter to a rate of 0.2L/min, 0.5L/min, or 0.8L/min, depending on the desired chemical concentration intended to pass through the sensors.
5. Initiate the operation of the sensor device setup and allow it to operate for a duration of 5 minutes.
6. Upon completion of the designated time, deactivate the flowmeters and air valve, detach the device, and ventilate the sensor box to facilitate air circulation.
7. Retrieve the data using a micro-SD card and transfer it to an Excel file for further analysis.

Procedure for Gas Mixtures:

1. Prepare mixtures of ethanol, methanol, and isopropanol in suitable proportions.
2. Follow steps 2 through 6 to pass the mixture of gases through the sensors.

3. Repeat the aforementioned process for various combinations and concentrations of gases as required.

4. Retrieve the data using a micro-SD card and transfer it to an Excel file for subsequent analysis.

This additional step facilitates the evaluation of the sensor device's response to gas mixtures, contributing to a comprehensive assessment of its performance.

4. Proposed Approach and Implementation

To address the task of gas classification and concentration estimation, we employ a multi-step approach combining data preprocessing, feature engineering, and machine learning techniques. Below, we outline the key components of our approach and demonstrate their implementation using Python programming language and relevant libraries.

4.1. Data Extraction and Analysis

Figure 2 **Test tubes and flowmeters used to produce the gases**

The primary dataset originates from the experimental setup depicted in the accompanying image. In this setup, an experiment is conducted without any base liquid present in the apparatus to establish baseline data, as illustrated in Figure 2. The experiment involves the utilization of three distinct quantities of ethanol, methanol, and isopropanol, detailed in Table 1. During the experiment, the gas is circulated through sensors for five minutes, and the resulting data is captured in an Excel sheet, as depicted in Figure 3. This

ms	Ammonia (NH3)-Echem	Nitrogen Dioxide (NO2)	Ethylene	Carbon Monoxide (CO)	Alcohol-Acetone	Alcohol	Ammonia (NH3)-Ethanol	Blank	LDR (light sensor)	Moisture	Nitric Oxide (NO)	Carbon Dioxide (CO2) (ppm)	Total Volatile Organic Compounds (TVOC) (ppb)	Oxygen (O2) (%)	Humidity (%)	Temp (deg C)	hic (look into)	Gas_Type
3779	8374	8292	8306	8277	11802	21402	40662	53384	64868	65128	49092	0	0	19.33	67	19.7	19.47	Ethanol 200
9196	8356	8296	8308	8342	11911	22432	40989	53125	65020	65080	49138	400	0	19.34	67	19.4	19.14	Ethanol 200
14596	8352	8303	8318	8279	11936	22725	40314	53092	65139	65061	49131	400	0	19.35	67	19.5	19.25	Ethanol 200
19995	8351	8260	8287	8304	11871	22676	39124	53187	64894	65118	49134	400	0	19.36	67	19.4	19.14	Ethanol 200
25400	8349	8270	8298	8287	12021	22710	38280	53134	65176	65027	49076	400	0	19.37	67	19.7	19.47	Ethanol 200
30798	8349	8286	8311	8300	12303	22880	37736	53171	65227	64982	49106	400	0	19.37	67	19.6	19.36	Ethanol 200
36196	8354	8286	8301	8294	12196	22736	37090	53068	64979	64995	49086	402	0	19.36	67	19.6	19.36	Ethanol 200
41594	8336	8310	8295	8282	12250	22692	36680	53017	65058	64967	49100	400	0	19.36	67	19.5	19.25	Ethanol 200
47003	8346	8314	8301	8277	12356	22833	36470	53144	65191	64968	49162	411	1	19.36	67	19.4	19.14	Ethanol 200
52401	8336	8288	8298	8282	12019	22400	36127	53051	65146	65042	49134	400	0	19.36	67	19.4	19.14	Ethanol 200
57801	8340	8303	8310	8278	11880	22315	35796	53161	65253	64891	49140	411	1	19.36	67	19.5	19.25	Ethanol 200
63199	8350	8308	8323	8301	11757	22156	35643	53082	65141	65113	49082	410	1	19.36	67	19.4	19.14	Ethanol 200
68601	8339	8251	8322	8282	11818	22223	35612	53174	65014	65202	49061	416	2	19.35	67	19.5	19.25	Ethanol 200
74001	8350	8263	8288	8284	11866	22168	35637	53110	65140	65159	49135	434	5	19.35	67	19.4	19.14	Ethanol 200
79399	8313	8298	8305	8292	11940	22131	35766	53053	65149	64991	49152	434	5	19.35	67	19.5	19.25	Ethanol 200

Figure 3 **The snippet of the extracted Excel file**

dataset encompasses several fields, including ammonia Echem, nitrogen dioxide, ethylene monoxide, alcohol acetone, alcohol, ammonia ethanol, LDR, moisture, nitric oxide, carbon dioxide, total volatile organic compounds, oxygen, humidity, temperature, HIC, and gas type. Notably, the target variable is gas type, while the remaining variables serve as inputs. The dataset is partitioned into training and testing sets using an 80:20 ratio, facilitating further data analysis and model evaluation.

4.2. Data Preprocessing and Compilation

In the original dataset, gas measurements were recorded across multiple Excel sheets as shown in Table 2, each representing a specific gas type and concentration level. These sheets varied in size, with dimensions indicated as rows and columns, denoting the number of observations and features, respectively. Through data preprocessing, these individual sheets were merged into a unified dataset, facilitating comprehensive analysis and modeling for gas classification and concentration estimation.Utilizing the Pandas library, we read Excel sheets containing sensor readings for different gas types, such as Ethanol, Methanol, and Isopropanol, among others.

4.3. Handling Class Imbalance

Numerous studies have demonstrated that employing a combination of over-sampling techniques to generate synthetic samples for the minority (abnormal) class, in conjunction with under-sampling of the majority (normal) class, results in enhanced classifier performance when compared to utilizing solely under-sampling of the majority class[Chawla

Table 2 Sheet Information

Sheet	**Shape**
Ethanol 200ppm	(114, 18)
Ethanol 123ppm	(114, 18)
Ethanol 161ppm	(114, 18)
Methanol 292ppm	(114, 18)
Methanol 137ppm	(114, 18)
Methanol 164ppm	(114, 18)
Isopropanol 217ppm	(114, 18)
Isopropanol 117ppm	(114, 18)
Isopropanol 143ppm	(114, 18)
E200 & M292	(171, 18)
E123 & M137	(171, 18)
E161 & M164	(171, 18)
E200 & I217	(171, 18)
E123 & I117	(171, 18)
E161 & I143	(171, 18)
M292 & I217	(171, 18)
M137 & I117	(171, 18)
M164 & I143	(171, 18)

et al. (2002)].Utilizing an under-sampling strategy and a boosting approach, the model trains multiple predictive classifiers and reidentifies challenging examples, where the end goal is to get balanced-class dataset[Zhao et al. (2023)]. A recent study conducted a novel theoretical analysis of the Synthetic Minority Over-sampling Technique (SMOTE) method by deriving the probability distribution of the SMOTE-generated samples[Elreedy et al. (2023)]. For each minority class sample, SMOTE randomly selects one of its k nearest neighbors. It then creates a synthetic sample along the line segment connecting the original sample and the selected neighbor. The synthetic sample is generated at a randomly chosen point along this line segment. SMOTE generates synthetic samples for the minority class to address the class imbalance. The algorithm works as follows:

1. For each sample $\mathbf{x_i}$ in the minority class, identify its k-nearest neighbors.

2. Generate a synthetic sample $\mathbf{x}_{\text{new}}$ by randomly selecting one of its k-nearest neighbors $\mathbf{x_{ij}}$ and interpolating between them:

$$\mathbf{x}_{\text{new}} = \mathbf{x_i} + \delta \cdot (\mathbf{x_{ij}} - \mathbf{x_i}) \tag{1}$$

Here, δ is a random number between 0 and 1.

3. Repeat the above steps for the desired number of synthetic samples.

In our study, we applied SMOTE to the training set with a specified sampling strategy to balance the classes. The target number of samples for each class was set to be three times the majority class. The equations used for this approach are:

$$\text{target\_samples}[c] = \max(\text{class\_counts}) \times 3 \tag{2}$$

where target_samples is the desired number of samples for each class c, and max(class_counts) is the maximum number of samples in the majority class. To generate additional synthetic samples and mitigate the risk of overfitting, Synthetic Minority Oversampling Technique (SMOTE) was employed. By augmenting the training dataset with synthetic samples, SMOTE effectively addresses class imbalance while preventing overfitting. We utilize the imblearn library to implement SMOTE and achieve a balanced training dataset.

5. Modeling Approach: Utilizing Gradient Boosting Machine (GBM)

GBM stands for Gradient Boosting Machine, a popular machine learning algorithm for classification and regression tasks. Gradient Boosting builds an ensemble of decision trees sequentially, with each tree learning from the errors of the previous ones. Multinomial classification is ubiquitous in various domains, including healthcare, finance, and marketing. GBM algorithms have gained popularity for their ability to handle complex datasets and deliver high predictive accuracy[Chen and Guestrin (2016)]. In a study, Gradient-boosted trees were employed in a large-scale mixed-integer nonlinear nonconvex optimization problem, for optimizing pretrained regression tree models for decision-making in chemical process catalyst selection, which signifies the versatility of GBM[Mistry et al. (2020)].

Boosting stands as a formidable learning paradigm introduced in recent years, aimed at synthesizing the predictions of numerous "weak" classifiers into a robust model. This approach essentially transforms a collective assembly of "weak" learners into a potent

"strong" learner, showcasing the efficacy of ensemble techniques [Hastie et al. (2009), Schapire (1999)].Gradient Boosting Machines (GBMs) fuse additive models with gradient descent, facilitating the optimization of a loss function. However, unlike traditional numerical optimization methods, GBMs employ boosting functions to iteratively enhance performance by advancing along the gradient direction [Friedman (2001)]. The parameters (β_m, α_m) are computed as:

$$(\beta_m, \alpha_m) = \arg\min_{\beta,\alpha} \sum_{i=1}^{n} L\left(y^{(i)}, F_{m-1}(x^{(i)}) + \beta h(x^{(i)}; \alpha)\right) \tag{1}$$

Using vectored notation, the updated ensemble function $F_m(X)$ is expressed as:

$$F_m(X) = F_{m-1}(X) + \eta \Delta_m(X) \tag{2}$$

Equation (1) illustrates the calculation of parameters where the objective is to minimize the loss function over the training dataset. This process entails adjusting the boosting functions to move in the direction of the gradient, thereby refining the model's predictive capabilities. Using vectored notation as depicted in Equation (2), the updated ensemble function incorporates the incremental changes to the previous ensemble function, which is scaled by a learning rate. This iterative approach enables GBMs to gradually improve the model's performance by iteratively fitting to the training data while avoiding overfitting.

5.1. Model Parameters and Variable Importance

Hyperparameter tuning is a critical step in optimizing machine learning models. The goal is to find the set of hyperparameters θ that minimizes the loss function L on the validation set. This process can be expressed as:

$$\theta^* = \arg\min_{\theta} L(\theta; X_{\text{val}}, y_{\text{val}}) \tag{3}$$

Here, θ represents the hyperparameters, L is the loss function (accuracy in this case), and $(X_{\text{val}}, y_{\text{val}})$ are the validation data and labels.

Grid search cross-validation evaluates each combination of hyperparameters from the defined grid:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{k} \sum_{i=1}^{k} L(\theta; X_{\text{train}_i}, y_{\text{train}_i}, X_{\text{val}_i}, y_{\text{val}_i}) \tag{4}$$

Where Θ is the set of all possible hyperparameter combinations, k is the number of folds in cross-validation, and $(X_{\text{train}_i}, y_{\text{train}_i}, X_{\text{val}_i}, y_{\text{val}_i})$ are the training and validation data and labels for the $i^{th} fold$.

The hyperparameter tuning process was conducted to optimize the performance of the Gradient Boosting Machine (GBM) model. Using the scikit-learn library in Python, a grid search approach was employed to explore combinations of hyperparameters, including learning rate, subsample, and max features. The parameter grid consisted of specific values for each hyperparameter: learning rate [0.05, 0.1, 0.2], subsample [0.8, 0.9, 1.0], and max features ['sqrt', 'log2', None]. The grid search was performed with 3-fold cross-validation to evaluate the model's performance. The best hyperparameters were determined based on the highest cross-validation accuracy. The finalized GBM model, initialized with the optimal hyperparameters, was trained on the resampled training dataset and evaluated on an independent test set using accuracy as the evaluation metric. We further optimize our model by tuning hyperparameters to maximize its performance. Based on our experiments, the best hyperparameters for the Gradient Boosting Machine (GBM) algorithm are {'learning_rate': 0.2, 'max_features': None, 'subsample': 0.9}. We achieve an accuracy of 98.63% when using all features of the dataset.

5.2. Feature Selection for Gas Classification and Concentration Estimation

In this section, we outline the feature selection process utilized to identify the most informative features for gas classification and concentration estimation. The process involved training a Gradient Boosting Machine (GBM) model and iteratively eliminating the lowest performing features.

Initially, the GBM model was trained using all available features as shown in 4a, resulting in an accuracy of 98.63%. Surprisingly, during this stage, the ethylene sensor (ME3-C2H4) was identified as one of the lowest performing features. Despite its known importance, it exhibited poor performance in this context.

In subsequent iterations, although ethylene sensor (ME3-C2H4) consistently ranked as one of the lowest performing features, it was retained due to its significance in gas classification and concentration estimation. By eliminating other low performing features while keeping ethylene sensor (ME3-C2H4) in the feature set as shown in Figure 4b, the model accuracy improved to 99.03%. Further refinement was achieved by removing all the least-performing features, including ethylene sensor (ME3-C2H4) as shown in Figure 4c. This

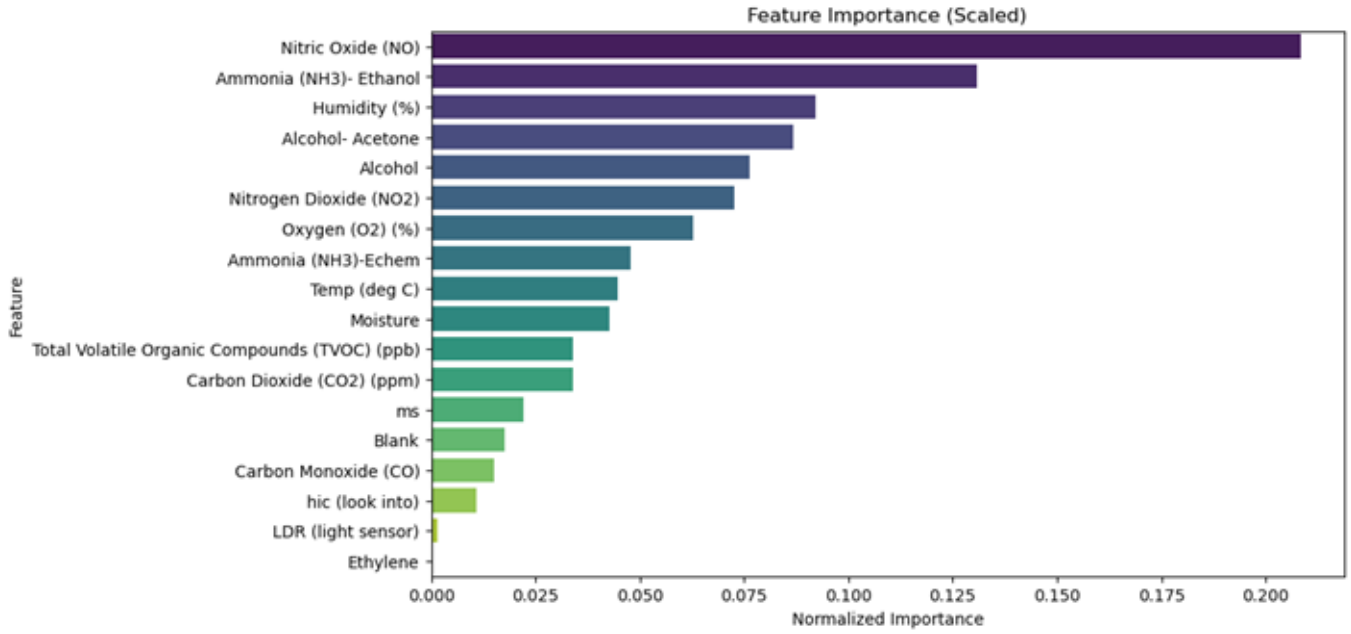

Figure 4a "Normalized Feature Importance of All Features in the GBM Model

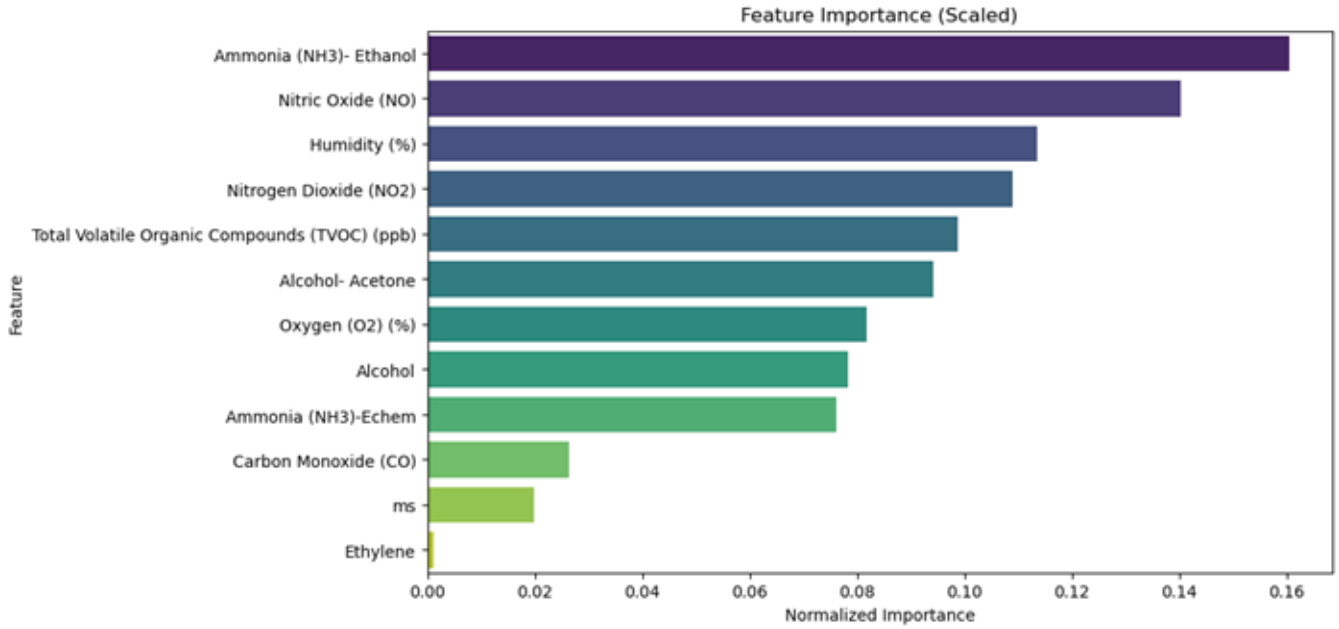

Figure 4b Feature Selection Process: Retaining Ethylene Sensor (ME3-C2H4) Despite Low Performance for Gas Classification and Concentration Estimation

resulted in an optimized model with an accuracy of 99.22%. Notably, the number of features used was reduced to just 9 from the initial 18 features present in the dataset. These results underscore the critical role of feature selection in optimizing the performance of gas classification and concentration estimation models, ensuring both accuracy and efficiency.

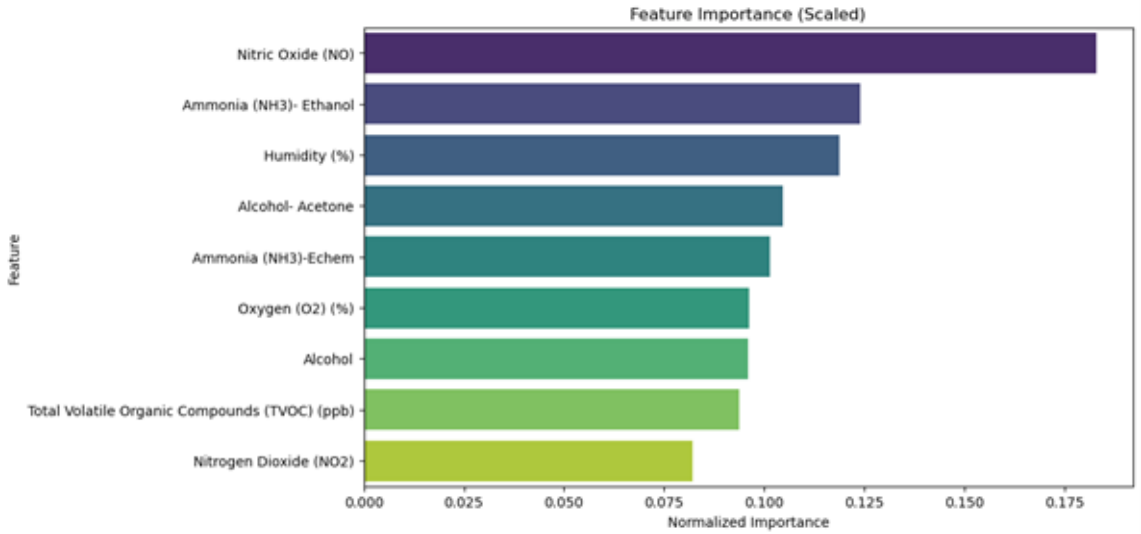

Figure 4c "Feature Refinement: Removal of Least-Performing Features, Including Ethylene Sensor (ME3-C2H4)

6. Results

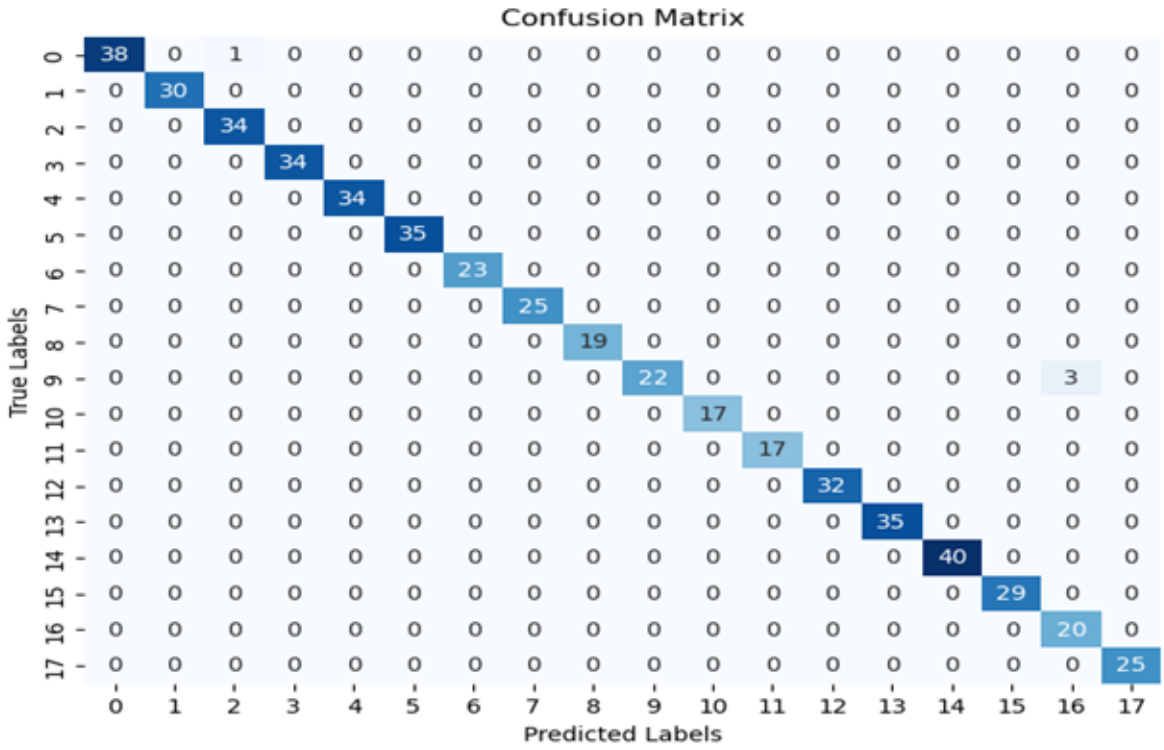

Figure 5 Confusion Matrix for Gas Classification

Figure 5 presents the confusion matrix generated during the evaluation of our Gradient Boosting Machine (GBM) model for gas classification and concentration estimation. Each row and column of the confusion matrix corresponds to a specific gas class, represented by

Table 3 Class Labels

Label	Class Description
0	E123 & I117
1	E123 & M137
2	E161 & I143
3	E161 & M164
4	E200 & I217
5	E200 & M292
6	Ethanol 123ppm
7	Ethanol 161ppm
8	Ethanol 200ppm
9	Isopropanol 117ppm
10	Isopropanol 143ppm
11	Isopropanol 217ppm
12	M137 & I117
13	M164 & I143
14	M292 & I217
15	Methanol 137ppm
16	Methanol 164ppm
17	Methanol 292ppm

the labels as shown in Table 3. Based on the confusion matrix as shown in Figure 5, Label 9, which represents Isopropanol 117ppm, was misclassified 3 times out of 25, resulting in an accuracy of 88% for this class. Evidently, this can be observed from the graph depicting the Accuracy and Counts of Actual and Correct Predictions by Class as shown in Figure 6. The function is designed to interpret the output of a gas mixture prediction model. It takes the predicted gas mixtures generated by the model as input, which are represented as a list of strings containing the names and concentrations of gases (e.g., ['E161 & I143']). The function then extracts the gas names and concentrations from each predicted gas mixture and computes the total concentration of all gases in the mixture. For instance, if the predicted gas mixture is ['E161 & I143'], the function computes the concentration of E

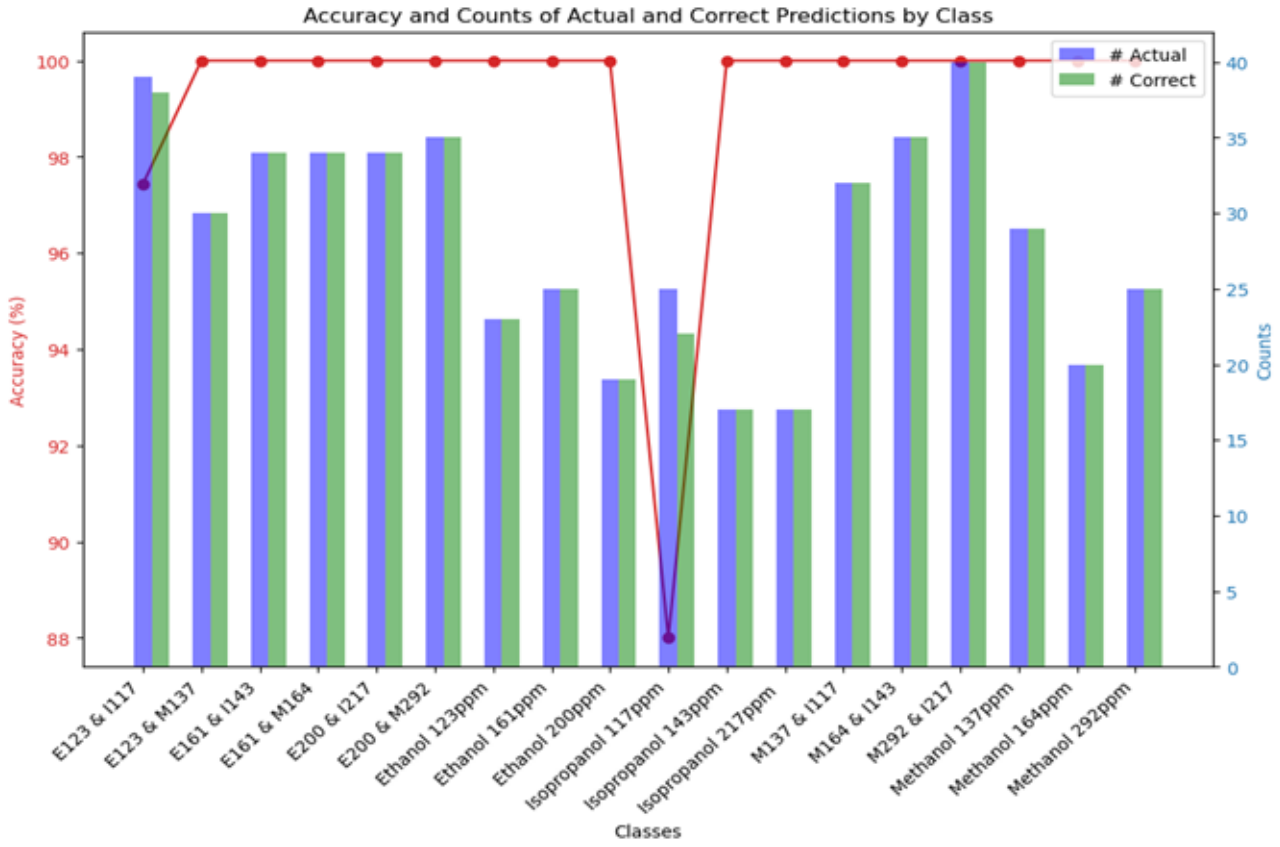

Figure 6 **Accuracy and Counts of Actual and Correct Predictions by Class**

as 161 ppm and the concentration of I as 143 ppm. Moreover, it calculates that E accounts for 52.96% of the mixture, while I constitutes 47.04%

This function aids in the analysis and interpretation of the model's predictions, providing insights into the composition and proportions of gases in the predicted mixtures, which is essential for various applications It facilitates a detailed examination of the model's performance and the implications of its predictions in real-world scenarios.

7. Conclusion and Future Scope

This study includes recognizing various gas types individually, assessing each gas's signature or range, and finding fresh inputs of the previously investigated gases. In this study, we investigated the task of gas classification and concentration estimation using a Gradient Boosting Machine (GBM) model. Through iterative feature selection, we identified a subset of nine key features that significantly contributed to model accuracy while reducing complexity. Notably, despite its initial low performance, the ethylene sensor (ME3-C2H4) was retained in the feature set due to its importance in gas characterization. The selected

features, including Nitric Oxide (NO), Total Volatile Organic Compounds (TVOC) (ppb), Ammonia (NH3)-Echem, Nitrogen Dioxide (NO2), Alcohol- Acetone, Alcohol, Ammonia (NH3)- Ethanol, Humidity (%), and Oxygen (O2) (%), proved pivotal in achieving an impressive accuracy of 99.22%. Furthermore, the model demonstrated robustness when tested on actual data without the use of synthetic generative data, showcasing its real-world applicability. By employing Synthetic Minority Over-sampling Technique (SMOTE) during training, we addressed class imbalance concerns and prevented overfitting, thereby enhancing the generalization capability of the model.

The significant advantages of E-noses, such as their high sensitivity, real-time analysis capabilities, ease of operation, and portability, could make them a promising platform for diagnosis of Cancer using blood samples and passing breath which has VOCs. This potential application serves as the basis for our future study, which will focus on the modification of the Gradient Boosting Machine (GBM) model developed herein to enable multi-classification in the context of cancer diagnosis. The implementation of electronic nose (e-nose) technology in clinical settings represents a groundbreaking shift towards more efficient and cost-effective disease detection methods. Our study demonstrates that the e-nose system achieves high accuracy in identifying VOCs associated with various diseases, offering a rapid and affordable alternative to traditional diagnostic techniques. While it is not yet used for cancer detection, the potential for future application in this field is significant. This advancement could revolutionize early disease detection, ensuring timely treatment and improved survival rates. Moreover, the reduced diagnostic costs and faster turnaround times would benefit both hospitals and insurance companies, reinforcing the e-nose's value as a vital tool in modern healthcare. Continued research and development in this area will further enhance the capabilities and applications of e-nose technology, solidifying its role in future cancer diagnostics and beyond.

Acknowledgments

We gratefully acknowledge the funding support from the USF SM campus IRG and Pioneer grants, which made this research possible.

Appendix. GitHub Source Code

https://github.com/premal11/Gas-Classification-and-Concentration-Estimation-using-Gradient-Boosting.git

References

Avila L (2005) Leak detection with thermal imaging. URL https://patents.google.com/patent/US6866089B2.

Behera B, Joshi R, Anil Vishnu G, Bhalerao S, Pandya HJ (2019) Electronic nose: A non-invasive technology for breath analysis of diabetes and lung cancer patients. *Journal of Breath Research* 13(2):024001, URL http://dx.doi.org/10.1088/1752-7163/aafc77.

Belhouari S, Bermak A, Shi M, Chan P (2005) Fast and robust gas identification system using an integrated gas sensor technology and gaussian mixture models. *IEEE Sens. J.* 5:1433–1444.

Bilgera C, Yamamoto A, Sawano M, Matsukura H, Ishida H (2018) Application of convolutional long short-term memory neural networks to signals collected from a sensor network for autonomous gas source localization in outdoor environments. *Sensors* 18(4484), URL http://dx.doi.org/10.3390/s18124484.

Brahim-Belhouari S, Bermak A, Shi M, Chan P (2005) Fast and robust gas identification system using an integrated gas sensor technology and gaussian mixture models. *IEEE Sens. J.* 5:1433–1444, URL http://dx.doi.org/10.1109/JSEN.2005.858971.

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)* 16:321–357, URL http://dx.doi.org/10.1613/jair.953.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dragonieri S, Annema J, Schot R, van der Schee M, Spanevello A, Carratú P, Resta O, Rabe K, Sterk P (2009) An electronic nose in the discrimination of patients with non-small cell lung cancer and copd. *Lung Cancer* 64:166–170, URL http://dx.doi.org/10.1016/j.lungcan.2008.08.008.

Dragonieri S, Pennazza G, Carratu P, Resta O (2017) Electronic nose technology in respiratory diseases. *Lung* 195:157–165, URL http://dx.doi.org/10.1007/s00408-017-9987-3.

D'Amico A, Pennazza G, Santonico M, Martinelli E, Roscioni C, Galluccio G, Paolesse R, Natale CD (2010) An investigation on electronic nose diagnosis of lung cancer. *Lung Cancer* 68:170–176, URL http://dx.doi.org/10.1016/j.lungcan.2009.11.003.

Elreedy D, Atiya A, Kamalov F (2023) A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Mach Learn* URL http://dx.doi.org/10.1007/s10994-022-06296-4.

Farraia M, Cavaleiro Rufo J, Paciência I, Mendes F, Delgado L, Moreira A (2019) The electronic nose technology in clinical diagnosis: a systematic review. *Porto Biomed J.* 4(4):e42, URL http://dx.doi.org/10.1097/j.pbj.0000000000000042.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5):1189–1232, URL http://dx.doi.org/10.1214/aos/1013203451.

Gardner J, Bartlett P (1994) A brief history of electronic noses. *Sens. Actuators B Chem.* 18:210–211, URL http://dx.doi.org/10.1016/0925-4005(94)87085-3.

Gordon S, Szidon J, Krotoszynski B, Gibbons R, O'Neill H (1985) Volatile organic compounds in exhaled air from patients with lung cancer. *Clinical Chemistry* 31:1278–1282, URL http://dx.doi.org/10.1093/clinchem/31.8.1278.

Gutierrez-Osuna R, Gutierrez-Galvez A, Powar N (2003) Transient response analysis for temperature-modulated chemoresistors. *Sens. Actuators B Chem.* 93:57–66.

Hakim M, Billan S, Tisch U, Peng G, Dvrokind I, Marom O, Abdah-Bortnyak R, Kuten A, Haick H (2011) Diagnosis of head-and-neck cancer from exhaled breath. *Br. J. Cancer* 104:1649–1655, URL http://dx.doi.org/10.1038/bjc.2011.128.

Hamilton S, Charalambous B (2020) *Leak Detection: Technology and Implementation* (London, UK: IWA Publishing).

Hanna G, Boshier P, Markar S, Romano A (2019) Accuracy and methodologic challenges of volatile organic compound-based exhaled breath tests for cancer diagnosis: a systematic review and meta-analysis. *JAMA Oncol.* 5(1):e182815, URL http://dx.doi.org/10.1001/jamaoncol.2018.2815.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning* (Springer New York, NY), 2nd edition.

Jadin M, Ghazali K (2014) Gas leakage detection using thermal imaging technique. *Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 302–306 (Cambridge, UK), URL http://dx.doi.org/10.1109/UKSim.2014.80.

Khalaf W, Pace C, Gaudioso M (2008) Gas detection via machine learning. *MDPI* .

Krilaviciute A, Heiss J, Leja M, Kupcinskas J, Haick H, Brenner H (2015) Detection of cancer through exhaled breath: a systematic review. *Oncotarget* 6(36):38643–38657, URL http://dx.doi.org/10.18632/oncotarget.5938.

Langford V, Graves I, McEwan M (2014) Rapid monitoring of volatile organic compounds: a comparison between gas chromatography/mass spectrometry and selected ion flow tube mass spectrometry. *Rapid Communications in Mass Spectrometry* 28(1):10–18, URL http://dx.doi.org/10.1002/rcm.6747.

Liu Q, Hu X, Ye M, Cheng X, Li F (2015) Gas recognition under sensor drift by using deep learning. *Int. J. Intell. Syst.* 30:907–922, URL http://dx.doi.org/10.1002/int.21746.

Machado R, Laskowski D, Deffenderfer O, Burch T, Zheng S, Mazzone P, Mekhail T, Jennings C, Stoller J, et al JP (2005) Detection of lung cancer by sensor array analyses of exhaled breath. *Am. J. Respir. Crit. Care Med.* 171:1286–1291, URL http://dx.doi.org/10.1164/rccm.200409-1184OC.

Marathe S (2019) Leveraging drone based imaging technology for pipeline and rou monitoring survey. *SPE Symposium: Asia Pacific Health, Safety, Security, Environment and Social Responsibility* (Kuala Lumpur, Malaysia: Society of Petroleum Engineers), URL https://www.onepetro.org/conference-paper/SPE-195543-MS.

Mazzone P, Hammel J, Dweik R, Na J, Czich C, Laskowski D, Mekhail T (2007) Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array. *Thorax* 62:565–568, URL http://dx.doi.org/10.1136/thx.2006.072892.

McWilliams A, Beigi P, Srinidhi A, Lam S, MacAulay C (2015) Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose. *IEEE Trans. Biomed. Eng.* 62:2044–2054, URL http://dx.doi.org/10.1109/TBME.2015.2409092.

Mistry M, Letsios D, Krennrich G, Lee RM, Misener R (2020) Mixed-integer convex nonlinear optimization with gradient-boosted trees embedded. *INFORMS Journal on Computing* 33(3):1103–1119, URL http://dx.doi.org/10.1287/ijoc.2020.0993.

Natale CD, Macagnano A, Martinelli E, Paolesse R, D'Arcangelo G, Roscioni C, Finazzi-Agrò A, D'Amico A (2003) Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosens. Bioelectron.* 18:1209–1218, URL http://dx.doi.org/10.1016/S0956-5663(03)00086-1.

Pan X, Zhang H, Ye W, Bermak A, Zhao X (2019) A fast and robust gas recognition algorithm based on hybrid convolutional and recurrent neural network. *IEEE Access* 7:100954–100963, URL http://dx.doi.org/10.1109/ACCESS.2019.2931050.

Pardo M, Sberveglieri G (2005) Classification of electronic nose data with support vector machines. *Sensors and Actuators B* 107:730–737.

Persaud K, Dodd G (1982) Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* 299:352–355, URL http://dx.doi.org/10.1038/299352a0.

Phillips M, Altorki N, Austin J, Cameron R, Cataneo R, Kloss R, Maxfield R, Munawar M, Pass H, et al AR (2008) Detection of lung cancer using weighted digital analysis of breath biomarkers. *Clin. Chim. Acta* 393:76–84, URL http://dx.doi.org/10.1016/j.cca.2008.02.021.

Schapire RE (1999) A brief introduction to boosting. *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2, ser. IJCAI'99*, 1401–1406 (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.).

Sun X, Shao K, Wang T (2016) Detection of volatile organic compounds (vocs) from exhaled breath as noninvasive methods for cancer diagnosis. *Analytical and Bioanalytical Chemistry* 408(11):2759–2780, URL http://dx.doi.org/10.1007/s00216-015-9200-6.

Tran V, Chan H, Thurston M, Jackson P, Lewis C, Yates D, Bell G, Thomas P (2010) Breath analysis of lung cancer patients using an electronic nose detection system. *IEEE Sens. J.* 10:1514–1518, URL http://dx.doi.org/10.1109/JSEN.2009.2038356.

van de Goor R, van Hooren M, Dingemans A, Kremer B, Kross K (2018) Training and validating a portable electronic nose for lung cancer screening. *J. Thorac. Oncol.* 13:676–681, URL http://dx.doi.org/10.1016/j.jtho.2018.01.024.

Wilson A (2015) Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath. *Metabolites* 5(1):140–163, URL `http://dx.doi.org/10.3390/metabo5010140`.

Yang J, Sun Z, Chen Y (2016) Fault detection using the clustering-knn rule for gas sensor arrays. *Sensors* 28:2069.

Yin X, Zhang L, Tian F, Zhang D (2015) Temperature modulated gas sensing e-nose system for low-cost and fast detection. *IEEE Sens. J.* 16:464–474.

Zhao H, Zhao C, Zhang X, Liu N, Zhu H, Liu Q, Xiong H (2023) An ensemble learning approach with gradient resampling for class-imbalance problems. *INFORMS Journal on Computing* 35(4):747–763, URL `http://dx.doi.org/10.1287/ijoc.2023.1274`.