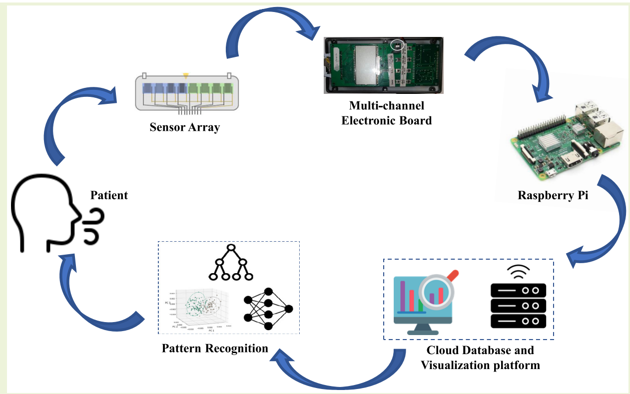
# A Wireless Sensor Array System Coupled With AI-Driven Data Analysis Towards Remote Monitoring of Human Breaths

Dong Dinh, Guojun Shang, Shan Yan, Jin Luo, Aimin Huang, Lefu Yang, Susan Lu, and Chuan-Jian Zhong




***Abstract*—The development of low-cost point-of-care sensor systems is essential for the screening and diagnostics of different diseases. However, this type of application requires effective integration of different sensor hardware and electronics in a portable, wireless, and reliable platform. We report herein the development of such a platform by integrating a nanostructured chemiresistive sensing array (CSA) with a low-current multichannel electronics board (MEB) and a Raspberry Pi board (RPB). The system allows the collection of data from the sensor array responses to volatile organic compounds (VOCs) and human breaths (HBs), then transfers the data through a serial connection from MEB to RPB. After processing and restructuring the data, RPB will wirelessly upload it to MongoDB Atlas (MDBA) cloud database. A workstation periodically retrieves the data from the MDBA cloud database and trains them with customized machine learning models. The best result feeds back to the MDBA cloud server, providing a pretrained model for a future prediction or disease identification. At the same time, the real-time sensor response data are displayed on the Thing speak portal. Once programed, the system runs in an independent mode without a PC connection with various functions, including remote monitoring services and ad hoc applications that are typically not accessible from traditional stationary monitoring systems housed in hospitals and laboratories. Some of these functions are demonstrated by testing the performance in sensing HB samples with and without simulated lung cancer-specific VOCs, showing promises for potential applications in remote breath monitoring and screening of lung cancer.**

***Index Terms*— Chemiresistive sensing array (CSA), human breath (HB) monitoring, lung cancer screening, multichannel electronics board (MEB), remote monitoring, volatile organic compounds (VOCs), wireless sensor array.**

Manuscript received 6 March 2023; accepted 23 March 2023. Date of publication 7 April 2023; date of current version 14 July 2023. This work was supported in part by the National Science Foundation under Grant CHE 2102482 and Grant IIP 1640669, in part by the Binghamton University Harpur College Faculty Research Seed Grant and Watson—United Health Services (UHS) Collaboration Seed Grant, and in part by Integrated Electronics Engineering Center (IEEC). The associate editor coordinating the review of this article and approving it for publication was Prof. Huang C. Lee. *(Corresponding authors: Susan Lu; Chuan-Jian Zhong.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Binghamton University IRB.

Dong Dinh and Susan Lu are with the Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902 USA (e-mail: ddinh2@binghamton.edu; slu@binghamton.edu).

Guojun Shang, Shan Yan, Jin Luo, and Chuan-Jian Zhong are with the Department of Chemistry, State University of New York at Binghamton, Binghamton, NY 13902 USA (e-mail: gshang1@binghamton.edu; syan4@binghamton.edu; jluo@binghamton.edu; cjzhong@binghamton.edu).

Aimin Huang and Lefu Yang are with Xiamen Smart NanoSensing Technologies, Xiamen, Fujian 361000, China (e-mail: aimin.huang@snsingtech.com; lfyang@xmu.edu.cn).

Digital Object Identifier 10.1109/JSEN.2023.3264205

## I. Introduction

According to the World Health Organization, lung cancer is the most common cancer and one of the leading causes of death around the world for both men and women [1], [2]. Early diagnosis (Stage I) could lead to a higher five-year survival rate (67%) than later stages [3]. Symptoms alone cannot be used to diagnose lung cancer since they typically do not show until the damage has been done. New methods for detecting cancer at the early stage are expected to enhance survival dramatically. As a candidate for the early diagnosis of lung cancer, breath analysis of volatile organic compounds (VOCs) attracted a considerable amount of scientific and clinical interest in recent years in medical applications due to its non-invasiveness, real-time capability, and wide applicability [4]. The investigation of breath and samples from cancer cell culture headspace can be performed using a variety of analytical methods. Gas chromatography coupled with mass spectrometry (GC-MS) is one of the most popular methods for analyzing breath samples [5]. A study conducted on breath samples from 65 patients with different stages of lung cancer and undergoing different treatment regimens using GC-MS successfully identified eight potential marker compounds, which showed the difference in concentration between healthy volunteers versus cancer patients [6]. Another study used GC-MS coupled with statistical methods to identify VOCs found in biological samples such as exhaled air from 108 patients with lung cancer, 121 healthy volunteers, and 24 people with other lung diseases. They were able to get an 80% correct classification rate for validation data [7]. However, GC-MS instruments are not portable and expensive. Furthermore, sample preparation, data analysis, and analyte identification are all time-consuming [8], [9]. Chemiresistor arrays made with molecularly capped gold nanoparticles as sensing materials have been used for the detection of VOCs in breath samples, which can potentially be used as a portable and cost-effective diagnostic tool for the early detection of lung cancer [10], [11], [12], [13], [14], [15], [34]. The use of pattern recognition algorithms in combination with these arrays has also been shown to improve their accuracy in detecting VOCs [16], [17], [18], [19], [25]. Recent studies on nanoparticle-based chemiresistor arrays have also shown the capability to detect tumors, diabetes, and other disorders [20], [21], [22], [23], [24], [27].

In contrast to most conventional methods that are invasive, expensive, time-consuming, and need specialized equipment and biomarker preconcentration [20], [22], [26], the integration of chemical sensors in point-of-care testing (POCT) provides noninvasive, real-time, and cost-effective detection of VOCs in breath samples. Such sensors can aid in the early diagnosis of diseases such as lung cancer [10], [11], [12], [13], [14], [15]. However, POCTs currently face challenges such as lower accuracy compared to traditional laboratory testing, high-cost manufacturing, and less user-friendly operation. To address some of the challenges, the use of point-of-care diagnosis (POCD) [32], [33] aided by machine learning techniques, cloud computing, and the Internet of Things (IoT) applications can improve the accuracy and user-friendliness of these devices and ultimately lead to applications in early detection and diagnosis of diseases. The use of color and thermal cameras for the development of monitoring systems to monitor and classify breathing patterns has shown some promise to enable automated health assessments [36]. While there are many reports on different sensor systems for detecting cancer-specific VOCs, there have been no reports demonstrating the integration of a nanoparticle-structured chemiresistor sensor array with a low excitation current detector in a wireless and portable device for HB detection and monitoring. In this article, we describe the design of a wireless portable sensor array monitoring system, which integrates a low excitation current chemiresistor detector developed in our recent work [25] with a wireless function, an IoT interface, and a database. The system can be deployed anywhere wirelessly in under 5 min, and it automatically uploads simultaneously collected sensor array breath data to the Google Cloud server for flexible data collection and analysis. This allows for not only low-temperature drift and low power consumption, but also real-time and remote monitoring capabilities, making it an ideal solution for a wide range of applications. In this report, results will be shown to demonstrate some of these functions by testing the performance of the integrated system in monitoring HB samples with and without simulated lung cancer-specific VOCs. Implications for potential applications in remote breath monitoring and screening of lung cancer will also be discussed.

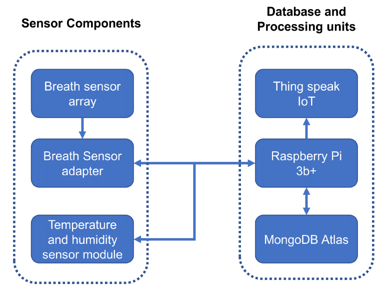



Fig. 1. Block diagram showing the wireless sensor array system in terms of the electronic communications between the sensor components and the database/processing units.

## II. System Architecture, Integration, and Implementation

### A. System Architecture

Fig. 1 shows the block diagram for the wireless sensor system, which includes sensor array electronics and Raspberry Pi board (RPB). The sensor array system first collects data from the chemiresistive responses to VOCs and sends the data to the RPB through a serial connection. RPB will then process and reorganize sensor array response data into a BSON format data file. The processed Binary JSON (BSON) data file will then be uploaded through a wireless connection to MongoDB Atlas (MDBA), which is the global cloud database service for modern applications. A workstation will pull down data from the MDBA cloud service after every certain period to train with ensemble learning models and deep learning models. The updated best output result will be uploaded back to the MDBA cloud server and stored in a separate BSON file as the pretrained model for future prediction.

The RPB can also use the pretrained model on the MDBA cloud server and output the classification result of the collected data. MongoDB is a document-oriented NoSQL database that is used for large-scale data storage. MongoDB employs collections and documents rather than tables and rows as in traditional relational databases. Documents are made up of key-value pairs, which are the fundamental unit of data in MongoDB. Collections are comparable to relational database tables in that they include collections of documents and functions. MongoDB stores data in JSON-like documents with variable structure, resulting in a dynamic, customizable schema. MongoDB was also created for high availability and scalability, with integrated replication and auto-sharing ability. The MDBA service is a global cloud database service developed and operated by the MongoDB team. While MongoDB is under the databases category, MDBA falls under MongoDB's Hosting.

### B. Data Acquisition and Processing

RPB is used as the main processing unit for the wireless sensor array system. It is not only able to process and transfer sensor data to the cloud database but also generates necessary commands to synchronize different components' operations. Serial protocols were heavily used to connect the peripherals (sensors) to RPB. These serial protocols are serial peripheral interface (SPI), integrated circuit interface or inter-integrated circuit (I2C), and standard universal asynchronous receiver transmitter (UART).

The humidity and temperature sensor module are connected to the RPB's general-purpose input–output (GPIO) pins. The sensor array electronic board is connected to RPB's USB port. The main concept of the wiring is that digital sensors are connected directly to the RPB's GPIO, while analog sensors are connected to an analog-to-digital converter (ADC), which in turn is directly connected to the RPB. RPB comes with both wired and wireless networks; it is extremely useful to get Wi-Fi ready for an IoT applications.

Python programming language is heavily used to create an autonomous process for collecting, processing, and transferring sensor response data. Necessary Python packages include Pandas, NumPy, scikit-learn, scipy, RPi.GPIO, pyserial, and urllib3. For remoting communication with Ras-Pi, MobaXterm provides not only one terminal but also the ability to open multiple terminals within a single window in a tabbed manner. Each terminal may be for a different purpose. A terminal may be connected to a remote shell via Secure Socket Shell while another is a File Transfer Protocol to some other host.

### C. Integration of Sensor Array and Electronic Board

Microfabrication is used for creating interdigitated microelectrodes (IME) [5]. Using the Nordiko 2000 vacuum sputtering deposition equipment, thin film electrodes were created on a glass substrate. Photolithography was used to pattern the IME device's construction. IME devices were utilized for signal transduction. The device features an array of 200 gold microelectrodes, each with a length of 200 $\mu$m, a width of 10 $\mu$m, and a 5 $\mu$m gap, on a 1-mm thick glass substrate. The microelectrodes were 100 nm thick and were fabricated at the Cornell NanoScale Science and Technology Facility. The chemiresistor sensor array is constructed by a thin film assembly of gold nanoparticles on the IME device. Gold nanoparticles with core diameters of 2 and 5 nm and decanethiolate (DT) monolayer shells were used [9], [10], [11], [12]. 11-mercaptoundecanoic acid (MUA), 1,9-nonanedithiol (NDT), 1,4-butanedithiol (BDT), 1,5-pentanedithiol (PDT), and 1,6-hexanedithiol (HDT) were used as linker molecules to assemble the gold nanoparticles. Hexane was used to create vapors (Hx, Fisher). There were two types of nanoparticle thin films used in this study: 1) NDT-linked nanoparticles (NDT-$Au_{nm}$) and 2) MUA-linked nanoparticles (MUA-$Au_{nm}$). The thin films were made by following the "exchanging-cross-linking-precipitation" method [13]. An exchange of linker molecules (NDT, MUA) with gold-bond alkanethiolates was followed by crosslinking and precipitation via Au-S bonding at both ends of NDT or hydrogen bonding at the carboxylic acid terminals of MUA. The eight-sensor array consists of the following sensing films: MUA-$Au_{2\,nm}$, MUA-$Au_{5\,nm}$, BDT-$Au_{2\,nm}$, BDT-$Au_{5\,nm}$, PDT-$Au_{2\,nm}$, HDT-$Au_{2\,nm}$, NDT-$Au_{2\,nm}$, and ND-$Au_{5\,nm}$.

Fig. 2 shows a block diagram for the design and integration of the multichannel electronics board (MEB) [25] with the nanostructured chemiresistor sensor array. The MEB was designed with the ability to automatically search for the channel based on the sensor's initial resistance value (Ri) and lock it in the most appropriate resistance range. This also minimizes the difference and time delay in the measured value caused by resistance switching at the measuring range's limit. The resistance is measured using the handheld chemiresistor detector (HCD) by measuring the voltage drop while a continuous current is sent through the sensor. Each channel's constant current source (6.0 nA–1.2 mA) includes 16 different current values, each corresponding to a different resistance range. For the specified sensors array interface, the input channels are configured as a modular array (Fig. 2). The output voltage signal for the sensor's resistance measured by each channel is transferred to an ADC, from which a microcontroller unit (MCU) calculates the resistance value. The electronic is designed in such a way that each channel of the array sensors can be monitored quickly. Through the microprocessor interface, the measured resistance values are sent to the readout. The adsorption of VOCs in the nanoparticle thin films on the interdigitated microelectrode device causes changes in interparticle spacing and dielectric medium properties. Such changes are transduced by the IME into changes in electrical resistance, which is detected by the low-current MEB. A change in electrical resistance depends on the chemical properties and concentration of the VOC and the interaction with the nanoparticle thin film. The array provides the response signals for data processing, leading to recognition, speciation, and quantitation.

### D. IoT Platform and Database

Two IoT platforms were used as servers for collecting and storing the data and building the database.

*1) Thinkspeak Server:* Thingspeak is a web-based open API IoT platform [4], [5], [6] that stores sensor data from various sensor notes and can visualize data output in graphical form at the web level. Thingspeak communicates with the help of an internet connection, which acts as a data packet carrier between the wireless connected sensor notes and the Thingspeak cloud to retrieve, save/store, analyze, observe, and work on the sensed data from the connected sensors to RPB. Moreover, Thingspeak also has the ability to construct a public-based channel for data analysis and estimation. To engage the wireless sensor note in sensing the data and transmitting it across the Internet, the data requires to be uploaded to Cloud. The Cloud uses graphic visualization operations and is available in the form of a virtual server to tell us about our real-time analog data. The IoT helps to not only bring all information together and permits us to communicate with our sensor notes but also different sensor notes to interact with each other.

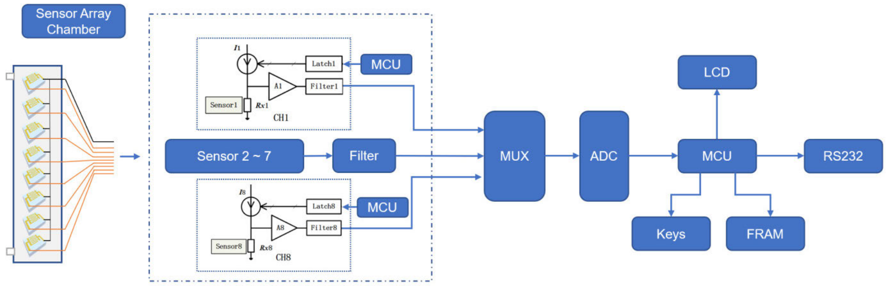



Fig. 2. Block diagram for the design of the multichannel electronic board for interfacing with the individual sensors in the nanostructured chemiresistor array.


Fig. 3. Operational flows for the sensor data collection and analytics with the IoT core architecture.

*2) Google Cloud Server:* Fig. 3 shows a typical sensor data collection and analytics with Google Cloud IoT Core Architecture. The sensor data from the Raspberry-pi are transmitted as JSON over HTTP to serverless Google Cloud Function HTTPS endpoints. Then, Google's Protocol Buffers are used to transmit binary data over HTTP. It features a reduction in the message size contained in the request payload from JSON to Protobuf, which reduces system latency and cost. Data received by Cloud Functions over HTTP will be published asynchronously to Google Cloud Pub/Sub. A second Cloud Function will respond to all published events and push the messages to MDBA on GCP. Once in Atlas, the data will be aggregated, transformed, analyzed, and built by machine learning models using tools such as MongoDB Compass, Jupyter Notebooks, and Google's Artificial Intelligent (AI) Platform Notebooks.

## *E. Hardware and Software Implementation*

As shown in Fig. 4, the BME280 sensor is connected to RPB through a GPIO connection using an SPI interface. The chemiresistor sensor board is connected to the sensor array and R-Pi through the serial protocol. Both chemiresistor sensor board and R-Pi are powered by external portable battery packs. The system is also equipped with temperature and humidity sensors. BME280 sensor is an environmental sensor with temperature, barometric pressure, and humidity. This sensor is great for indoor environmental sensing and can even be used in both I2C and SPI. The sensor can measure humidity with $\pm 3\%$ accuracy, barometric pressure with $\pm 1$ h Pa absolute accuracy, and temperature with $\pm 1.0$ °C accuracy. For simple easy wiring, we can go with I2C. If we want to connect a series of sensors without worrying about I2C address collisions, SPI can be a preferred option.

RPB uses PiOS (previously called Raspbian) as an operating system. Raspberry Pi Imager is a quick and easy way to install Raspberry Pi OS and other operating systems to a microSD card, ready to use with RPB. To program on the RPB and use it to control sensors, we used Virtual Network Computing (VNC) or Secure Shell Protocol (SSH) client on the main computer. The headless RPB is activated by writing an empty text file named "ssh" (no file extension) to the root of the directory of the micro SD card. When it sees the "ssh" on its first boot-up, Raspberry Pi OS will automatically enable ssh, allowing remote access to the Pi command line.

# III. Data Collection and Analysis Methodologies

To determine the performance of the sensor system in recognizing different mixtures of VOCs in real HBs, breath samples were collected with healthy subjects. All subjects were asked to take 3–5 deep breaths before sample collection. Breath samples were collected into 1L Tedlar sampling bags. To reduce the humidity level to an acceptable range for effective data analysis, we used a plastic mouthpiece to blow through 1-foot-long silicone tubing and allowed the sampling bag to stabilize for an additional 5 h. This helped to bring the relative humidity to a level between 30% and 35%. The healthy volunteers were asked to be fasting for 12 h prior to their breath collection. All samples were analyzed between 5 and 24 h after collection, and multiple repeated tests were conducted.

In this study, simulated breath samples were created by adding specific lung cancer-related volatile VOCs to healthy breath samples collected from human subjects. The VOCs chosen for this study were toluene, 2-butanone, and 2,3,4-trimethyl-pentane, which were generated by a gas flow and mixing method. The gas flow was precisely controlled by an array of calibrated Aalborg mass flow controllers (AFC-2600), as previously described in the literature [16], [35]. The concentration of the targeted gases in breath samples is 9 ppm each. This method allows for the creation of simulated breath samples with representative lung cancer-specific VOCs, providing a useful tool for evaluating the performance of the sensor system.

The sensor array's response data from HBs and simulated cancer breaths are analyzed by principal component analysis (PCA), $k$-nearest neighbor (KNN'), support vector machine

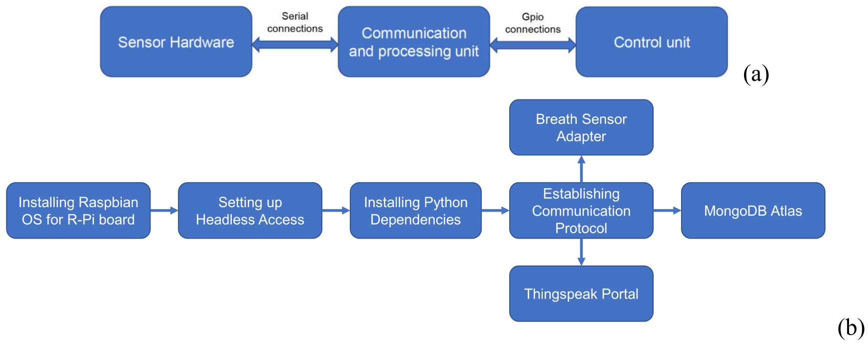



Fig. 4. Block diagrams showing the communications in the sensor and wireless hardware. (a) Hardware connections for the sensor array (eliminate the real photographs and add figure flow), temperature/humidity sensors, electronic board, and R-Pi. (b) Setting up Raspberry Pi.

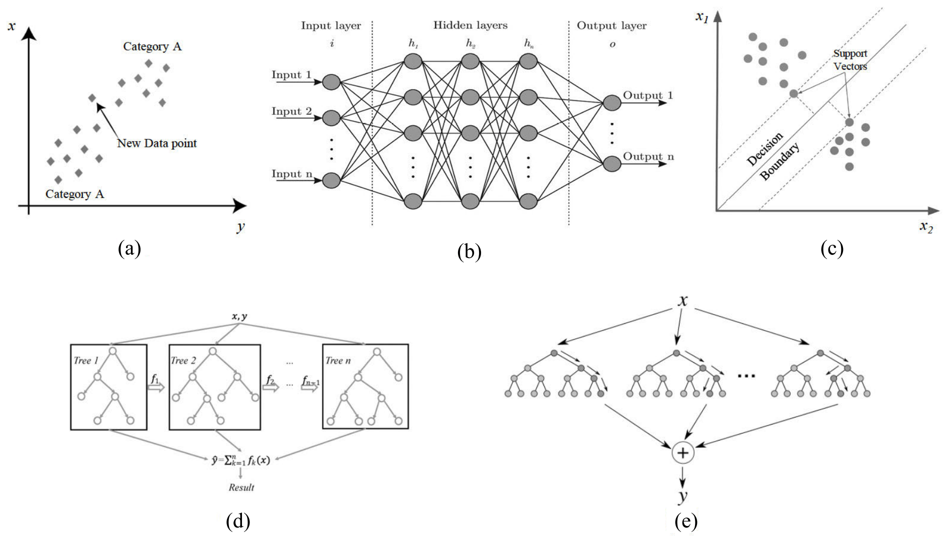



Fig. 5. Schematic of several different statistical methods used for data analysis in this work. (a) KNN, (b) ANN, (c) SVM, (d) XGBoost, and (e) RF classifier.

(SVM), random forest (RF), XGBoost, and artificial neural network (ANN), which are summarized as follows.

PCA is used to explore and reduce the dimensionality of the data. For our analysis, PCA was applied to sensor response to create new, uncorrelated variables and reduce the dimensions of a set of variables while preserving the majority amount of original variance among variables. In addition to PCA, five different statistical methods (Fig. 5) were used for the data analysis of the collected sensor array data.

The KNN classifier algorithm is based on the idea of feature similarity to estimate the value of new response variables. As it is shown in Fig. 5(a), for an unknown set of predictor variables, the value of classification is calculated by averaging the classification values of $k$-nearest observations. Euclidean and Mahalanobis Distance can be used to determine the distance between new unknown observations and training observations.

SVM is mainly used for classification problems by developing a maximum margin hyperplane between two classes. In Fig. 5(c), SVM essentially finds the Decision Boundary. If we had 1-D data, we would separate the data using a single threshold value. If we had 3-D data, SVM's output is a plane separating the two classes. Finally, if the data is more than three dimensions, the decision boundary is a hyperplane which is nothing but a plane in higher dimensions.

The RF is an ensemble learning algorithm proposed by Breiman [28] to predict discrete-dependent variables by bagging multiple decision trees into the model and averaging all predictions for results. RF algorithm can obtain better prediction accuracy with good tolerance for extreme value and noise. The procedure of the RF classifier is summarized in Fig. 5(e). In the first part of the algorithm, multiple subsamples are extracted from the original samples using the bootstrap resampling method, and a decision tree is applied for each sample. In the second step, the final prediction from the RF classifier is computed by an average of the predictions produced by the trees in the forest.

XGBoost [29] is a tree-based ensemble machine learning algorithm that is built on a gradient-boosting framework. As it is visualized in Fig. 6(d), the algorithm consists of optimized gradient boosting that has been enhanced through parallel processing, tree-pruning, handling missing values, and regularization to avoid overfitting/bias. Since its introduction, the advantages of XGBoost have been validated by several machine learning and data mining challenges [30].

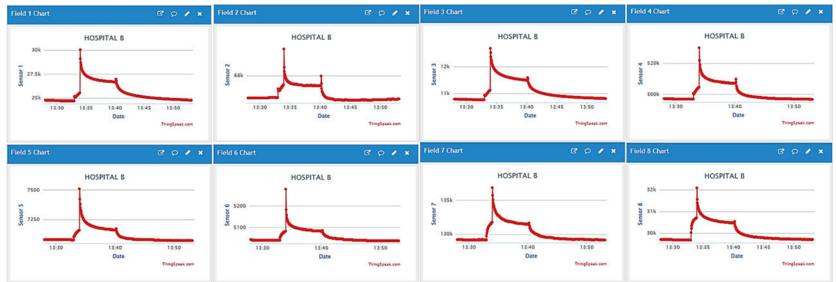



Fig. 6. Screenshot of the real-time sensor array response (Ω) profiles recorded during the data collection and processing, as displayed on the Thingspeak Portal.

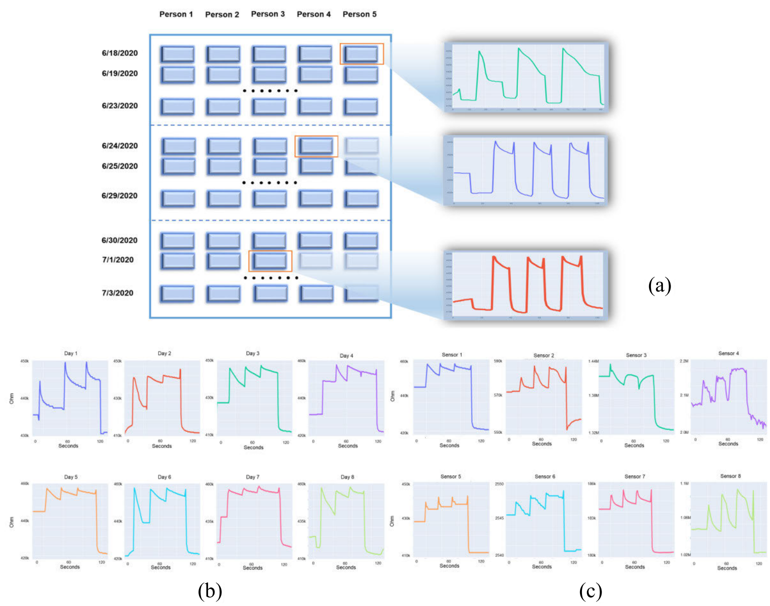



Fig. 7. Screenshots of the historical data displays during remote monitoring of HBs on the database. (a) Temporal remote monitoring of breath sample data structuring and response profiles from remote tracking of breath samples, (b) id1 with sensor 1 for eight days of data collection under different data collection protocols, and (c) record of id1 with sensor 1.

ANN is a methodology motivated by a biological network of neurons. ANN can be used as a powerful tool to capture complex or nonlinear relationships between predictor and response variables [31]. As seen in Fig. 6(b), a basic neural network model includes an input layer, one or more hidden layers in the middle, and an output layer at the end. Each layer in the ANN model consists of a collection of nodes. The output of each node is calculated by multiplication between inputs and activation function. Each connection among nodes across different layers has a certain weight which regulates the signal between the neurons. Weights between nodes are continuously updated based on feedback from output errors [31].

Despite the fact that some machine learning models, such as ANN, require large amounts of data for training, using the original data collected from the limited number of healthy volunteers at different time points with added simulated noise is adequate to evaluate the performance of the sensor system in terms of shallow machine learning models like KNN, SVM, RF, and XGBoost.

## IV. Results and Discussion

### A. Remote Monitoring and Simulation of HBs

To evaluate the performance of the integrated system for remote monitoring of HBs, two different sensor array stations A and B were set up in two different physical locations. There were five volunteers and sensor testing lasted more than 22 days. Sensor responses were recorded through the Thingspeak IoT portal from any place with an internet connection and web browser connectivity. Data from sensor responses were saved in both the Thingspeak IoT portal and an MDBA Atlas Database on Google Cloud Service.

Once the error-free coding is established, then the program is executed, and the sensor data output is shown from a local platform such as LCD 16 × 2 display. The local data is transferred to the Thingspeak Cloud via the Internet and can be visualized on a global platform. Fig. 6 shows the graphical output at the Thingspeak cloud observed upon logging in to the Thingspeak website with a secure username and password.

Fig. 7 shows the data layout on the display screen at one of the monitoring stations (a) and examples of the sensor response profiles from one of the sensors (Sensor 1) in the sensor array for a subject (id2) on eight different days of the continuous data collection (b). The system displayed excellent stability and continuity, demonstrating the ability of the system to capture different patterns from a specific person on different days and the viability of wirelessly transferring and storing data in a cloud database. From the recorded data for id2 with Sensors 1–8, we can analyze the different response

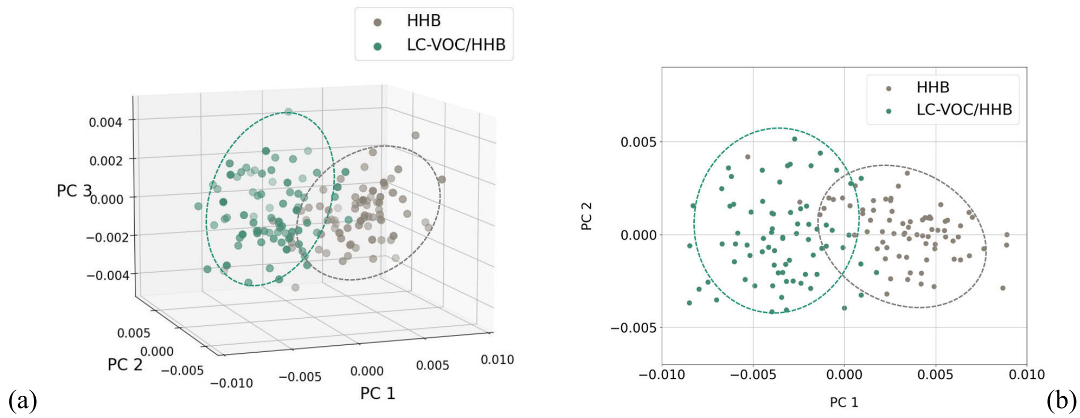



Fig. 8. (a) Three-dimensional and (b) 2-D PCA plots of the results from the sensor array system in monitoring breath samples with and without spiked lung cancer-specific VOCs. Samples: HB and HB with *LC*-specific VOCs.

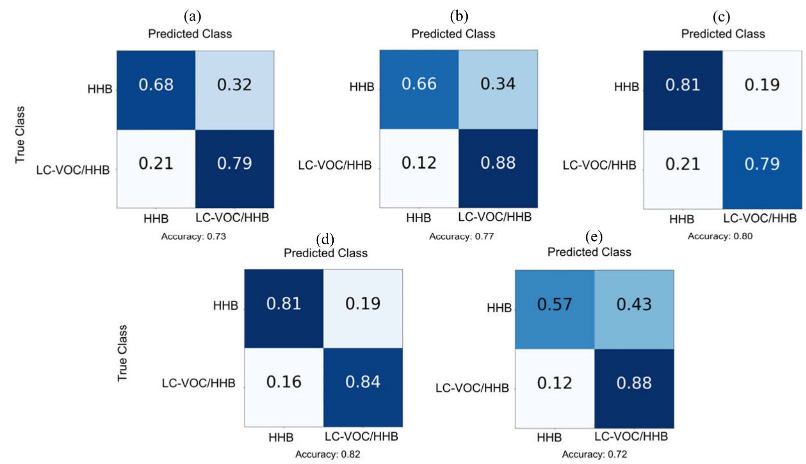



Fig. 9. Confusion matrix result of (a) KNN, (b) SVM, (c) RF, (d) XGBoost, and (e) ANN classifier. Samples: HB and HB with *LC*-specific VOCs.

patterns, thereby achieving differential sensing information for assessing multianalytes. The information is processed for further machine learning analysis. Moreover, the use of sensor arrays with multiple sensors enables the effective management of the impact of defective sensors on system performance.

## *B. Pattern Recognition of the Sensor Responses*

To illustrate how the data collected are used for pattern recognition, we show here one specific example of remote screening of HB samples with and without spiked lung cancer-specific VOCs. In this study, a dataset of 150 data points, including 75 healthy breath samples and 75 simulated breath samples from lung cancer patients was used. These data were divided into 70% for training and 30% for testing. The combination of the abundance ratio range [19] and Monte Carlo simulations was used to test the performance of the machine learning model in noisy conditions. The calibrated responses of all vapors to random and systematic errors are assumed to have a Gaussian distribution. Through multiple repeated sampling of these error-enhanced responses and assigning each sample an identity based on the abundance ratio range determined during calibration, the expected frequency and nature of recognition errors under standard operating conditions can be evaluated. Data for the sample testing were collected on three consecutive days, in which 50% of data were collected 5 h after the sample collection and the other 50% were collected 24 h after the sample collection. By doing that, the performance of the sensor array system could be evaluated in terms of sample shelf time variations. Fig. 8 shows the PCA analysis results from the collected data. The result displays two separate clusters with a small overlapping area, demonstrating a good separation among the HB samples with and without spiked lung cancer-specific VOCs.

We also examined the performance evaluation of different analytic methods in terms of sensitivity and selectivity. Fig. 9 shows a representative set of the confusion matrixes of classification results from various machine learning techniques, where the diagonal elements represent the total correct values predicted per class. The darker the color, the greater the number. There are two types of color coding, darker and lighter colors on the diagonal elements, which illustrate how well the model performs in terms of sensitivity and selectivity in recognition of healthy and *LC*-VOC-spiked HBs.

Five different statistical methods were used to assess the sensitivity and selectivity of the sensor array data. Among the five different methods, the XGBoost classifier performs the best with an 82% overall correct classification rate and successfully identified 84% of the breath samples with lung cancer-specific VOCs, i.e., 84% sensitivity, while the selectivity is 81%. The next best appears to be the RF Classifier, showing 79% sensitivity and 81% selectivity. Even though XGBoost shows a better overall classification rate and high selectivity (81%) compared to ANN, ANN appears to correctly classify more breath samples with lung cancer-specific VOCs, i.e., higher sensitivity (88%) but with worse selectivity (57%). The KNN classifier performs worst among all methods, with a very low overall correct classification rate but also low sensitivity and selectivity.

## V. Summary

In summary, we have demonstrated an effective strategy for the integration of a nanostructured chemiresistive sensing array (CSA) with a low-current MEB and an RPB as a portable and wireless sensor platform. The coupling of the sensor array and electronics board with RPB allows effective data collection and database development. The sensor array system collects data from HBs with and without certain lung cancer-specific VOCs. The data are first transferred to the RBP through a serial connection and then RPB enables the processing of the sensor array response data, which is followed by wirelessly uploading to a cloud database, MDBA. A workstation can pull down data from the MDBA cloud database after a certain period of time to train with ensemble learning and deep learning models. The updated best output result is then uploaded back to the MDBA cloud server and stored in a separate BSON file as the pretrained model for future prediction. Real-time sensor data are visualized in the form of charts in Thing speak. The integrated system can work in standalone mode without the requirement of a PC once programed. This system is viable for potential point-of-care and remote monitoring of HBs for screening for lung cancer. This viability is demonstrated by the performance evaluation of the system which is coupled with different analytic methods to determine the sensitivity and selectivity in monitoring HB samples with and without lung cancer-specific VOCs, demonstrating a high recognition rate. The repeatability and reliability of the system are also evaluated by results from continuous monitoring of the HB samples, including remote monitoring services and ad hoc applications. This capability is important for applications in areas that are typically not accessible from traditional stationary monitoring systems housed in hospitals and laboratories [37]. The results from the performance evaluation of the system in monitoring HB samples with and without simulated lung cancer-specific VOCs have shown promise for potential applications in remote breath monitoring and screening of lung cancer. This work has demonstrated the integration of our laboratory-developed low-current multichannel detection electronics board and the nanostructured chemiresistor sensor array with the commercially available wireless function and cloud data techniques for breath monitoring, which represents a new development in the system hardware in our work. The AI model further amplifies the system's capabilities, providing a comprehensive understanding of the specific database setup, the protocol for wireless data transfer, the type of breath samples being monitored, and the specific machine learning algorithm used toward a comprehensive understanding of the optimization of the wireless system's capabilities for potential applications.

## References


[1] T. N. Zamay et al., "Current and prospective protein biomarkers of lung cancer," *Cancers*, vol. 9, no. 11, p. 155, 2017.

[2] Y. Adiguzel and H. Kulah, "Breath sensors for lung cancer diagnosis," *Biosensors Bioelectron.*, vol. 65, pp. 121–138, Mar. 2015.

[3] C. F. Mountain, "Revisions in the international system for staging lung cancer," *Chest*, vol. 111, no. 6, pp. 1710–1717, Jun. 1997.

[4] H. P. Chan, C. Lewis, and P. S. Thomas, "Exhaled breath analysis: Novel approach for early detection of lung cancer," *Lung Cancer*, vol. 63, no. 2, pp. 164–168, Feb. 2009.

[5] J. K. Schubert, W. Miekisch, K. Geiger, and G. F. Nöldge-Schomburg, "Breath analysis in critically ill patients: Potential and limitations," *Expert Rev. Mol. Diagnostics*, vol. 4, no. 5, pp. 619–629, Sep. 2004.

[6] M. Ligor et al., "Determination of volatile organic compounds in exhaled breath of patients with lung cancer using solid phase microextraction and gas chromatography mass spectrometry," *Clin. Chem. Lab. Med.*, vol. 47, no. 5, pp. 550–560, Jan. 2009.

[7] J. Rudnicka, M. Walczak, T. Kowalkowski, T. Jezierski, and B. Buszewski, "Determination of volatile organic compounds as potential markers of lung cancer by gas chromatography–mass spectrometry versus trained dogs," *Sens. Actuators B, Chem.*, vol. 202, pp. 615–621, Oct. 2014.

[8] N. J. W. Rattray, Z. Hamrang, D. K. Trivedi, R. Goodacre, and S. J. Fowler, "Taking your breath away: Metabolomics breathes life in to personalized medicine," *Trends Biotechnol.*, vol. 32, no. 10, pp. 538–548, Oct. 2014.

[9] J. Rudnicka, T. Kowalkowski, T. Ligor, and B. Buszewski, "Determination of volatile organic compounds as biomarkers of lung cancer by SPME–GC–TOF/MS and chemometrics," *J. Chromatography B*, vol. 879, no. 30, pp. 3360–3366, Nov. 2011.

[10] A. W. Snow, F. K. Perkins, M. G. Ancona, J. T. Robinson, E. S. Snow, and E. E. Foos, "Disordered nanomaterials for chemielectric vapor sensing: A review," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1301–1320, Mar. 2015.

[11] N. Olichwer, E. W. Leib, A. H. Halfar, A. Petrov, and T. Vossmeyer, "Cross-linked gold nanoparticles on polyethylene: Resistive responses to tensile strain and vapors," *ACS Appl. Mater. Interfaces*, vol. 4, pp. 6151–6161, Nov. 2012.

[12] F. J. Ibañez and F. P. Zamborini, "Chemiresistive sensing with chemically modified metal and alloy nanoparticles," *Small*, vol. 8, no. 2, pp. 174–202, Jan. 2011.

[13] M. C. Leopold, R. L. Donkers, D. Georganopoulou, M. Fisher, F. P. Zamborini, and R. W. Murray, "Growth, conductivity, and vapor response properties of metal ion-carboxylate linked nanoparticle films," *Faraday Discuss.*, vol. 125, p. 63, Jul. 2004.

[14] C.-L. Chang, S.-W. Chiu, and K.-T. Tang, "An ADC-free adaptive interface circuit of resistive sensor for electronic nose system," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 2012–2015.

[15] L. Han et al., "Nanoparticle-structured sensing array materials and pattern recognition for VOC detection," *Sens. Actuators B, Chem.*, vol. 106, no. 1, pp. 431–441, Apr. 2005.

[16] L. Wang et al., "Array of molecularly mediated thin film assemblies of nanoparticles: Correlation of vapor sensing with interparticle spatial properties," *J. Amer. Chem. Soc.*, vol. 129, no. 7, pp. 2161–2170, 2007.

[17] L. Wang, J. Luo, M. J. Schadt, and C.-J. Zhong, "Thin film assemblies of molecularly-linked metal nanoparticles and multifunctional properties," *Langmuir*, vol. 26, no. 2, pp. 618–632, Jan. 2010.

[18] L. Wang et al., "Sensing arrays constructed from nanoparticle thin films and interdigitated microelectrodes," *Sensors*, vol. 6, no. 6, pp. 667–679, Jun. 2006.

[19] G. Peng, U. Tisch, and H. Haick, "Detection of nonpolar molecules by means of carrier scattering in random networks of carbon nanotubes: Toward diagnosis of diseases via breath samples," *Nano Lett.*, vol. 9, no. 4, pp. 1362–1368, Apr. 2009.

[20] M. Hakim et al., "Volatile organic compounds of lung cancer and possible biochemical pathways," *Chem. Rev.*, vol. 112, no. 11, pp. 5949–5966, Nov. 2012.

[21] M. Phillips et al., "Volatile organic compounds in breath as markers of lung cancer: A cross-sectional study," *Lancet*, vol. 353, no. 9168, pp. 1930–1933, Jun. 1999.

[22] M. K. Nakhleh et al., "Sensor arrays based on nanoparticles for early detection of kidney injury by breath samples," *Nanomedicine, Nanotechnol., Biol. Med.*, vol. 10, no. 8, pp. 1767–1776, Nov. 2014.
[23] S. W. Chiu and K. T. Tang, "Towards a chemiresistive sensor-integrated electronic nose: A review," *Sensors*, vol. 13, no. 10, pp. 14214–14247, Oct. 2013.
[24] R. C. Hughes et al., "Integrated chemiresistor array for small sensor platforms," in *Proc. SPIE*, 2000, pp. 1–12.
[25] Z. Wang et al., "A low-current and multi-channel chemiresistor array sensor device," *Sensors*, vol. 22, no. 7, p. 2781, Apr. 2022.
[26] W. Zhao et al., "Detection of mixed volatile organic compounds and lung cancer breaths using chemiresistor arrays with crosslinked nanoparticle thin films," *Sens. Actuators B, Chem.*, vol. 232, pp. 292–299, Sep. 2016.
[27] S. Yan et al., "Nano-filamented textile sensor platform with high structure sensitivity," *ACS Appl. Mater. Interfaces*, vol. 14, no. 13, pp. 15391–15400, Apr. 2022.
[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[29] T. Chen and C. Guestrin, "XGBoost," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[30] A. Kadiyala and A. Kumar, "Applications of Python to evaluate the performance of decision tree-based boosting algorithms," *Environ. Prog. Sustain. Energy*, vol. 37, no. 2, pp. 618–623, Mar. 2018.
[31] A. Sharma, P. K. Sahoo, R. K. Tripathi, and L. C. Meher, "Artificial neural network-based prediction of performance and emission characteristics of CI engine using polanga as a biodiesel," *Int. J. Ambient Energy*, vol. 37, no. 6, pp. 559–570, Nov. 2016.
[32] S. K. Vashist, "Point-of-care diagnostics: Recent advances and trends," *Biosensors*, vol. 7, no. 4, p. 62, Dec. 2017.
[33] S. Gavrilaş, C. Ş. Ursachi, S. Perţa-Crişan, and F.-D. Munteanu, "Recent trends in biosensors for environmental quality monitoring," *Sensors*, vol. 22, no. 4, p. 1513, Feb. 2022.
[34] H.-W. Cheng et al., "Assessing plasmonic nanoprobes in electromagnetic field enhancement for SERS detection of biomarkers," *Sensors*, vol. 21, no. 24, p. 8345, Dec. 2021.
[35] M. J. Schadt, W. Cheung, J. Luo, and C.-J. Zhong, "Molecularly tuned size selectivity in thermal processing of gold nanoparticles," *Chem. Mater.*, vol. 18, no. 22, pp. 5147–5149, Oct. 2006.
[36] J. Kunczik, K. Hubbermann, L. Mösch, A. Follmann, M. Czaplik, and C. Barbosa Pereira, "Breathing pattern monitoring by using remote sensors," *Sensors*, vol. 22, no. 22, p. 8854, Nov. 2022.
[37] M. Ali, A. Elsayed, A. Mendez, Y. Savaria, M. Sawan, and M. Sawan, "Contact and remote breathing rate monitoring techniques: A review," *IEEE Sensors J.*, vol. 21, no. 13, pp. 14569–14586, Jul. 2021.

**Dong Dinh** is pursuing the Ph.D. degree with the Department of Systems Science and Industrial Engineering, State University of New York (SUNY) at Binghamton, Binghamton, NY, USA.

His current research interests include sensor/biosensor optimization, system development, pattern recognition, and the Internet of Things (IoT).

**Guojun Shang** is pursuing the Ph.D. degree with the Department of Materials Science and Engineering and Chemistry, State University of New York (SUNY) at Binghamton, Binghamton, NY, USA.

Her current research interests include chemical sensors/biosensors, functional nanomaterials, nanotechnology, nanofabrication, and wearable devices and applications.

**Shan Yan** received the Ph.D. degree in chemistry from the State University of New York (SUNY) at Binghamton, Binghamton, NY, USA, in 2019.

She is currently a Senior Scientist with the Stony Brook and Brookhaven National Laboratory, SUNY. Her research interests include chemical sensors/biosensors, batteries, energy storage materials, nanofabrication, and nanotechnology.

**Jin Luo** received the Ph.D. degree in physical chemistry from Xiamen University, Xiamen, China, and The University of Tokyo, Tokyo, Japan, in 1990.

He is a Senior Scientist with the Department of Chemistry, State University of New York (SUNY) at Binghamton, Binghamton, NY, USA. His research interests include sensors, catalysts, nanomaterials, microelectronics, and functional devices.

**Aimin Huang** received the Ph.D. degree in chemistry from the University of Connecticut, Stamford, CT, USA, in 2001.

She is a Senior Scientist at Xiamen Smart NanoSensing, Xiamen, China. Her research interests include catalysts, sensors, nanomaterials, and functional devices.

**Lefu Yang** received the Ph.D. degree in chemistry from Xiamen University, Xiamen, China, in 1998.

He is an Associate Professor at the Department of Chemistry, Xiamen University, and a Senior Scientist at Xiamen Smart NanoSensing, Xiamen. His research interests include catalysts, sensors, nanomaterials, and nanotechnology.

**Susan Lu** received the Ph.D. degree from the Department of Industrial Engineering, Texas Tech University, Lubbock, TX, USA, in 1999.

She is a Professor with the Department of Systems Science and Industrial Engineering, State University of New York (SUNY) at Binghamton, Binghamton, NY, USA. Her current research interests include pattern recognition, reliability, and process optimization.

**Chuan-Jian Zhong** received the Ph.D. degree in chemistry from Xiamen University, Xiamen, China, in 1989.

He was a Postdoctoral at the Fritz-Haber Institute and the University of Minnesota, Minneapolis, MN, USA, and an Associate Scientist at DOE-Ames Laboratory, Ames, IA, USA. He is a Professor in chemistry at the State University of New York (SUNY) at Binghamton, Binghamton, NY, USA. His research interests include chemical sensors/biosensors, heterogeneous catalysts, interfacial chemistry, and functional nanomaterials and devices.