Journal of Electronics and Informatics (ISSN: 2582-3825)
www.irojournals.com/iroei/

Breath Based Pneumonia Detection Using Gas Sensors and Decision Tree Classifier with Risk Analysis

Bharath P M.[1], Suguna S.[2], Oviya S.[3], Sobiya M.[4], Tasneem Taj H.[5]

Department of Biomedical Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, Puducherry, India.

E-mail: [1]bharathmannappan1@gmail.com, [2]sugunasubramani54@gmail.com, [3]ooviya.s2021343@gmail.com, [4]sobim2831@gmail.com, [5]tasneem05taj@gmail.com

Abstract

Pneumonia is a serious respiratory disease that necessitates immediate diagnosis in order to minimize the chance of complications and enhance the well-being of patients. Traditional techniques for diagnosis involve chest X-ray, laboratory testing, which are both time-intensive and costly. This article focuses on a framework for detecting pneumonia using breath analysis that utilizes multi-sensor gas sensing along with machine learning for non-invasive assessment of risks associated with it. This framework makes use of MQ-3, MQ-135, and BME680 sensors to analyze Volatile Organic Compounds (VOC) and environment. The collected data is preprocessed using Min-Max scaling technique, clustered using K-Means algorithm into different risk groups and then classified by using a Decision Tree Classifier. The experimental evaluation has resulted in 95.84% accuracy, 95.12% precision, 94.87% recall and 94.99% F1-Score. The proposed system is a cost-effective and portable approach for the early pneumonia detection.

Keywords: Pneumonia Detection, Breath Analysis, Metal Oxide Gas Sensors, Volatile Organic Compounds (VOCs), Decision Tree Classifier, Non-Invasive Diagnosis.

DOI: https://doi.org/10.36548/jei.2026.3.006
Extended version of article received from **International Conference on Smart Innovations in Engineering & Technology (ICSIET 2026)**
Received: 01.06.2026, received in revised form: 08.07.2026, accepted: 25.07.2026, published: 04.08.2026

1. Introduction

Pneumonia remains as the leading cause of respiratory disease and deaths all over the world, common among young children, older patients, and people with compromised immunity. Early disease detection is crucial for efficient treatment and proper care of affected patients. The traditional means of diagnosing this condition include chest x-rays, CT scans, and laboratory tests, usually involve special equipment and professional staff as well as a lot of time. Recent developments in breath analysis show that volatile organic compounds (VOCs) found in exhaled breath can be used as biomarkers of respiratory diseases, thus allowing non-invasive testing of health status. This initiates the design and development of mobile and inexpensive devices for early detection of pneumonia.

Various methodologies for pneumonia detection have been explored in existing studies, deep learning-based image analysis, VOC analysis, gas sensor technology, and artificial intelligence-enabled respiration monitoring techniques. Image-based detection techniques resulted in good accuracy, although they depend on imaging modalities and a large annotated dataset. Similarly, considerable scope exists for breath analysis-based technology in detecting diseases using VOCs; however, many existing systems are mainly designed for the purpose of biomarker or sensor development rather than for the overall process of disease diagnosis. Moreover, issues like environmental variations, lack of clinically labeled breath data, portability, and ease of implementation on an embedded system delay the practical implementation of the available techniques. This increases the need to design a comprehensive platform where breath analysis and data processing is carried out simultaneously.

To address these challenges, a mobile breath analysis-based pneumonia detection model is introduced by using MQ-3, MQ-135 and BME680 gas sensors coupled with an Arduino embedded system for the live collection of breath features. After preprocessing the collected sensor responses, K-means clustering is used for the generation of pneumonia risk classes and Decision Tree classifier for pneumonia risk prediction. The proposed framework integrates multi-sensor VOC detection, environment parameter correction, and intelligent classification in a cohesive, cost-effective, and non-invasive testing platform. The main contribution of this research is to integrate embedded sensing and hybrid machine learning to develop an effective respiratory screening system for rapid pneumonia evaluation. The experimental results show the feasibility of the proposed framework for real-time pneumonia risk detection.

2. Literature Review

2.1 Breath-Based Disease Detection

The breath analysis approach has been developed as a promising non-invasive method of diagnosing respiratory diseases based on the presence of volatile organic compounds (VOCs) in the exhaled breath. Improvements in breath collection protocols and in analytical methodologies have increased the sensitivity and reproducibility of VOCs analysis, thereby enabling the identification of accurate biomarkers [1]. Breath analysis using high-sensitivity instruments equipped with multiple gas sensors has improved breath discrimination capacity [3]. standardized techniques for identifying VOCs have helped in reducing false identifications and increased the reliability of the test results [6]. Breath analysis kits have proved that real-time disease screening is possible through exhaled biomarkers [9]. In addition to this, breath sensors combined with advanced breath samplers have improved multi-biomarker detection while maintaining the integrity of VOCs [10], [11].

2.2 Gas Sensor-Based Respiratory Monitoring

Recent advancements in the gas sensor technology have increased the ability to detect VOCs of specific diseases for respiratory analysis. The optimized metal oxide gas sensors have proven to have greater sensitivity, selectivity, and long-term stability for the trace gas detection [5]. Also, numerical modeling has helped to develop better sensors and simplify the calibration process [8]. The high-sensitivity gas sensor array incorporated in portable breath analyzers has allowed for real-time VOC measurement [3]. Gas sensing systems based on microcontrollers have proven the viability of developing inexpensive and embedded respiratory monitoring systems [7].

2.3 Machine Learning for Pneumonia Detection

Machine learning techniques have significantly enhanced the automated pneumonia detection process. The classification of pneumonia has been successfully done by using the convolutional neural network approach through the effective learning of the lung abnormalities [2]. The artificial intelligence approach has also been used to analyze the breathing process through the use of respiratory monitoring systems with multiple sensor inputs, thereby facilitating early detection of respiratory diseases [4]. While imaging approaches perform well

for the detection of diseases, they also require special medical equipment. Breath sensing, coupled with machine learning algorithms, results as a viable alternative approach.

2.4 Research Gap

Several existing researches have demonstrated the efficacy of using breath VOCs analysis methods, gas sensor technology, and machine learning approaches in diagnosing respiratory diseases. However, the existing breath analysis devices only pay attention to biomarkers discovery, whereas pneumonia diagnosis based on machine learning utilizes chest X-ray imagery. Few works were conducted on developing a unified approach based on multi-sensor VOCs collection, environmental factors compensation, unsupervised risk labels generation, and real-time embedded pneumonia risk classification. Therefore, the proposed model involves using MQ-3, MQ-135, and BME680 gas sensors, Arduino board for data collection, K-Means clustering algorithm for generating risk labels, and Decision Tree classifier for developing inexpensive, non-invasive, and real-time pneumonia risk diagnosis system.

3. Proposed Methodology

3.1 Overall System Architecture

The proposed framework represents a system design for portable and non-invasive detection of pneumonia risk through the integration of multiple breath sensors along with the machine learning algorithms. The general structure of the proposed framework is shown in Figure 1, it includes five main modules: breath sensing, data collection from embedded systems, data pre-processing, intelligent risk prediction, and visual results display. Initially, the collection of exhaled breath samples is done through the MQ-3, MQ-135, and BME680 sensors that in aggregate collect the VOCs and environment readings associated with respiratory diseases. Then, the data obtained from the sensors is relayed to the Arduino UNO for real-time data collection and initial preprocessing before being sent to the processing unit.

Preprocessing of the collected sensor data is done by performing noise filtering, outliers removal, and feature normalization in order to increase the quality of data used by the machine learning model. In order to improve the model's ability to deal with labeled clinical datasets of breathing sounds, the K-Means algorithm is used to cluster the preprocessed data into three groups: Normal, Mild Risk, and High Risk. The labels for these clusters are then used to train

the Decision Tree classifier to learn the association between breath sensor data and risk of pneumonia. In practice, the trained classifier will predict the risk class from the breath input data, and the associated diagnostic result will be shown on the embedded 16×2 LCD screen.

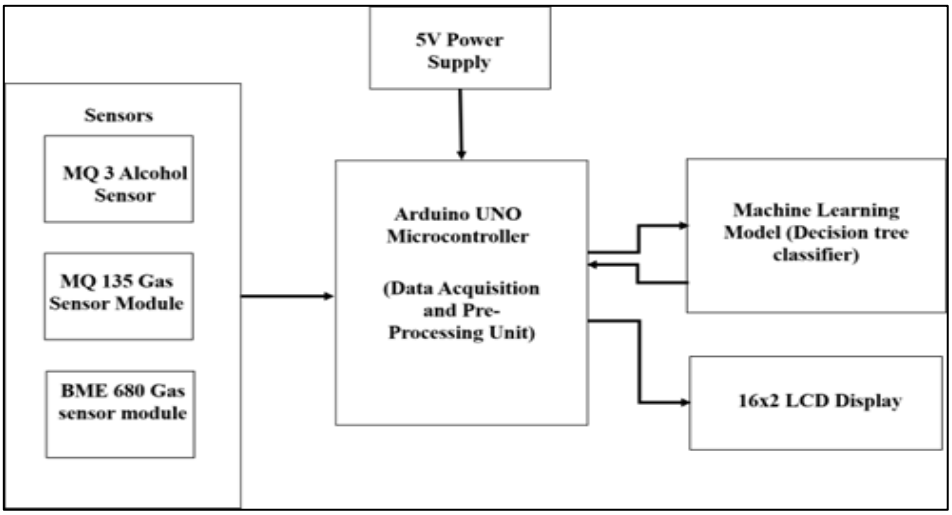

Figure 1. Overall architecture of the proposed breath-based pneumonia detection system

3.2 Hardware Implementation

The pneumonia detection model is implemented in an embedded system, which consists of Arduino UNO, MQ-3, MQ-135, BME680 sensors, and 16×2 LCD display. The hardware prototype used is depicted in Figure 2, showing that the sensing modules, microcontroller, power source, and display screen have been integrated into a portable enclosure.

The Arduino UNO serves as the central processing unit, that is responsible for receiving analog and digital signals from all sensors. This microcontroller is used for data acquisition in real time, initializing the sensors, signal conditioning, and communicating with the processing system for machine learning purposes. The output prediction is fed to the LCD display for visualization at the same time.

The MQ-135 gas sensor is used to detect air quality-related gases, like ammonia, carbon dioxide, and volatile organic compounds associated with respiratory diseases. Besides this, the MQ-3 gas sensor senses volatile compounds of alcohol and acetone in the breath of patients. The BME680 sensor is used to record environmental conditions, like temperature, humidity, pressure, and gas resistance. This helps to account for environmental factors that might affect the performance of the gas sensors.

A 16×2 LCD screen display is used to the pneumonia risk assessment level instantly, thereby making it possible to view the screening result instantly. The compact nature of the system makes it possible to implement in portable form and perform preliminary screening for respiratory disease.

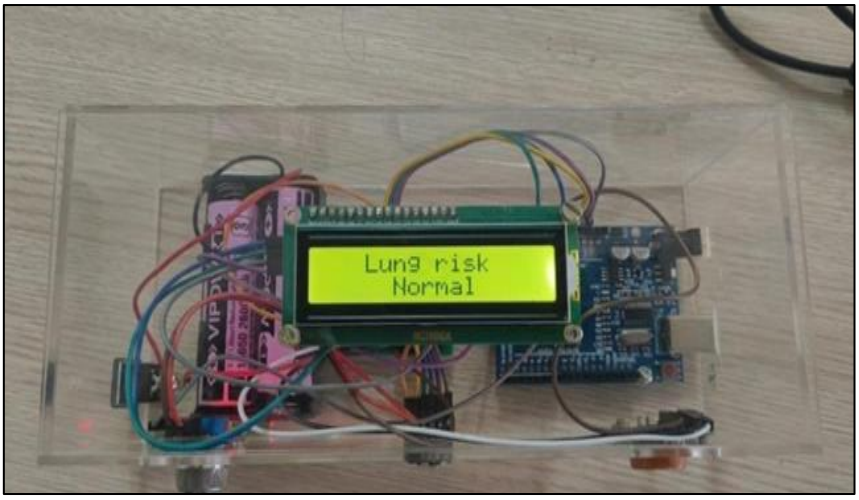

Figure 2. Real-time hardware implementation of the proposed pneumonia detection system

3.3 Dataset Description

The proposed pneumonia detection system employs a self-created breath sensor dataset which has been created with the aid of the developed embedded sensor platform. This dataset is made up of breath samples captured in real time by the sensors included in the platform including the MQ-3, MQ-135, and BME680 sensors. Breath analysis is based on VOCs, which have received extensive research interest as biomarkers for diagnosing respiratory diseases [12]. Each breath sample represents a single observation with various sensor properties related to VOCs and environmental parameters. Properties of the acquired dataset are presented in Table 1 below. Features recorded in the dataset include the MQ-3 sensor response, MQ-135 sensor response, gas resistance, temperature, humidity, and atmospheric pressure.

Table 1. Characteristics of the self-collected breath sensor dataset

Parameter	Value
Dataset Type	Self-collected breath sensor dataset
Dataset Source	Developed prototype (supported by VOC biomarkerinformation)
Number of Gas Sensors	3

Number of Recorded Features	6
Recorded Features	MQ-3 Response, MQ-135 Response, Gas Resistance, Temperature, Humidity, Pressure
Dataset Category	Unlabeled
Number of Risk Categories	3 (Normal, Mild Risk, High Risk)
Clustering Algorithm	K-Means
Classification Algorithm	Decision Tree
Data Acquisition Mode	Real-time breath sensing

Conventional supervised learning dataset, the breath data obtained are unlabeled owing to lack of clinical annotations of pneumonia categories. As such, the dataset is first processed through k-means clustering in order to create three categories of pneumonia risk: normal, mild risk, and high risk. This way, the breath-based pneumonia screening model will be able to achieve effective supervised learning while remaining practical.

3.4 Breath Data Acquisition

The breath data acquisition is performing the use of the designed embedded sensing system to measure the real-time VOCs and environmental parameters in the exhaled breath. First, the sensors such as the MQ-3, MQ-135, and BME680 are permitted to stabilize after a suitable warm-up process to ensure efficient sensing. Thereafter, participants are required to exhale directly into the sensor box to facilitate the acquisition of variations in the VOC levels together with temperature, humidity, pressure, and gas resistance. The entire breath data acquisition procedure is shown in Figure 2, the proposed hardware prototype acquires sensor readings and transmits them to the Arduino UNO.

The Arduino UNO takes analog and digital readings from the sensors and sends the collected information to the processing unit for further analysis. Information from the BME680 sensor about the environmental variables is used to consider the fluctuations in the environment so that the accuracy of the collected breath information can be enhanced. The collected data from the sensors is compiled into a dataset and sent on to the next step of preprocessing.

3.5 Data Preprocessing

The obtained breath sensors' readings are preprocessed for improving the accuracy of the data and ensuring that reliable machine learning analysis can be performed on them. First,

the readings are run through a moving average filter in order to remove any random variation and noise coming from the sensors due to the process of breathing. Next, the outliers in the data are removed to prevent any anomalous behavior of the sensor readings due to variations in the breathing process.

feature normalization is done to scale all sensor features to the same numeric range such that features with higher values do not dominate the learning process. Min-Max normalization is used for the purpose of preserving the distribution of sensors values while maintaining stability of clustering and classification processes. The normalized feature value is computed as

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x denotes the original sensor value, $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the corresponding feature, respectively, and x_{norm} is the normalized feature value. The processed dataset is then sent to the K-Means Clustering model for pneumonia risk segmentation.

3.6 K-Means-Based Risk Segmentation

The labeled breath sensor data set is unlabeled and K-Means clustering method is used to automatically cluster the preprocessed data into different clusters of pneumonia risks depending on the similar VOCs. Unsupervised learning does not depend on the clinical labeled data sets but allows meaningful grouping of the breath data according to the sensor data set features. Three clusters are identified using the clustering algorithm which corresponds to normal, mild risk, and high risk. Hence, risk segmentation has been initially done before applying supervised learning. The classification criteria used for the risk groups are shown in Table 2.

Table 2. K-means clustering classification

Cluster Label	Risk Category	Description
Cluster 1	Normal	VOC levels corresponding to healthy respiratory conditions.
Cluster 2	Mild Risk	Moderate deviations in VOC concentrations indicating possible early pneumonia symptoms.

Cluster 3	High Risk	Significant VOC variations associated with pneumonia.

Cluster labels that have been generated are considered as pseudo labels, which indicate the degree of pneumonia risk associated with each breath sample. Pseudo labels serve as the starting point for model training by using a supervised learning approach. This allows the Decision Tree algorithm to learn how different sensor values are related to different pneumonia risk categories.

3.7 Decision Tree-Based Risk Classification

The breath sensor clustering dataset is employed to train a Decision Tree classifier that predicts pneumonia risk automatically. Pseudo labels created by using K-Means clustering are attached to sensor readings. As a result, it becomes possible for the classifier to capture the correlation between patterns of VOCs and categories of pneumonia risks. Being based on rules hierarchically, the Decision Tree classifier efficiently splits the feature space into different decision regions to classify breath samples into Normal, Mild Risk, and High-Risk categories.

During the prediction stage, the trained classifier takes the preprocessed data from the sensors as input and performs the rule evaluations one by one to identify the respective risk level of pneumonia. The decision tree classifier uses a rule-based algorithm for classification, thus providing an efficient method of classifying the inputs. The predicted pneumonia risk is passed to the deployment phase for enabling visualization and user notification.

3.8 Model Development and Optimization

The Decision Tree algorithm is trained on the breath sensor dataset that has been preprocessed along with the pseudo labels obtained using the K-Means clustering algorithm. The preprocessed sensor features are fed as input into the model to create decision rules that can distinguish the three classes of pneumonia risk. In order to make the trained model generalize better and prevent the problem of overfitting, suitable values for hyperparameters such as maximum tree depth, minimum samples needed to split a node, and minimum samples required at the leaves are selected. In addition, cross-validation is used to assess the robustness of the model and feature importance analysis to find out the impact of each feature on classification.

Gini impurity criterion is used in the Decision Tree to select the best feature and the threshold at which each decision node is split. Gini impurity is defined as

$$G = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

where p_i represents the probability of samples belonging to the i^{th} class, and n denotes the total number of risk categories. A lower value of Gini impurity leads to a better separation between the classes. This gives rise to better decision boundaries and, thus, better classification of the risk of pneumonia. The tuned classifier is then used for live prediction.

3.9 Real-Time Deployment

The decision tree classifier is used to conduct the prediction of risks associated with pneumonia on a live basis by using sensor signals from breath. As such, exhaled breath is sampled using MQ-3, MQ-135, and BME680 sensors and the data transmitted by the Arduino UNO for processing. Data from the sensors undergoes similar preprocessing procedures that were undertaken during the training of the model.

The pre-processed data from the sensors are train into the Decision Tree classifier model that will predict the risk level associated with pneumonia as Normal, Mild Risk, and High Risk. This prediction is instantaneously shown on the 16×2 LCD screen. The entire deployment process allows the continual, non-invasive, and real-time assessment of the risk level of pneumonia without using traditional imaging equipment and laboratory-based diagnostic methods. The fusion of embedded sensing, intelligent data processing, and prediction capability makes it possible for the proposed system to act as a decision support system for preliminary pneumonia screening.

4. Results and Discussion

4.1 Experimental Setup

The proposed breath-based pneumonia detection system was subjected to experimental tests to determine its ability to predict real-time risks. The experiments were performed through the hardware prototype built together with the MQ-3, MQ-135, and BME680 gas sensors. The breath sensor data collected was analyzed through Python, where the K-Means algorithm was used to label the data, and then the Decision Tree classifier was used. The model's performance was evaluated based on classification metrics such as accuracy, precision, recall, and F1-score. The experimental setup used for the model testing is shown in Table 3.

Table 3. Experimental setup and evaluation configuration

Parameter	Configuration
Hardware Platform	Arduino UNO with MQ-3, MQ-135, and BME680 sensors
Programming Environment	Python 3.x
Machine Learning Library	Scikit-learn
Clustering Algorithm	K-Means
Classification Algorithm	Decision Tree
Feature Set	Six preprocessed sensor features
Data Normalization	Min-Max Normalization
Validation Strategy	Train-Test Split (80:20)
Evaluation Metrics	Accuracy, Precision, Recall, and F1-score
Output	Normal, Mild Risk, and High Risk

4.2 Multi-Sensor Response and Risk Segmentation Analysis

Efficacy of the breath-based pneumonia detection system was initially verified based on analysis of responses from the combined gas sensor array. Clustered results from sensors are depicted in Figure 3, the K-Means clustering algorithm categorizes the breath samples into three categories such as Normal, Mild Risk, and High-Risk. It is evident from the visualization that the resulting feature values for sensors vary in the VOC levels and therefore can be classified into separate clusters.

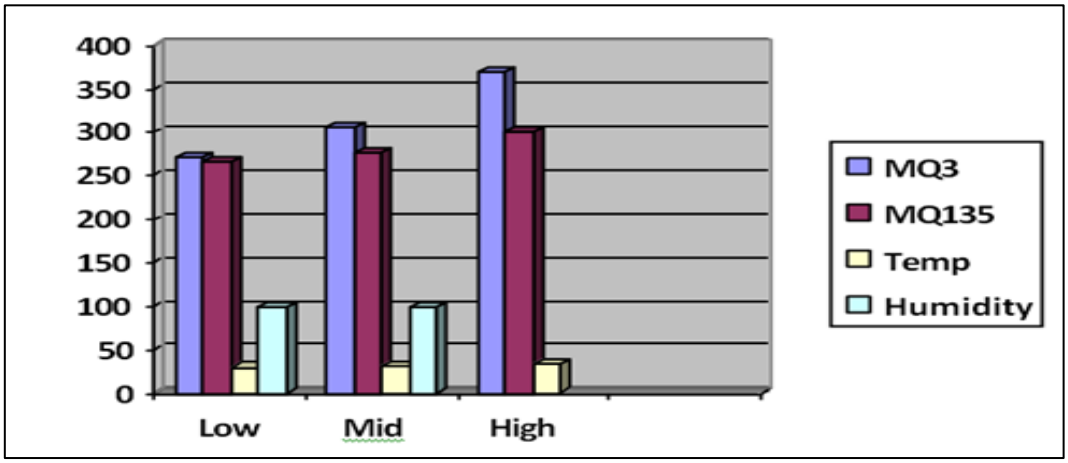

Figure 3. Multi-sensor response analysis across pneumonia risk categories

Cluster analysis shows that the samples with similar characteristics of VOCs have been clustered together and those samples having more deviation in the sensor outputs have been

classified into the risky clusters. The output of MQ-3, MQ-135, and BME680 sensors adds more dimensions to the information about alcohol concentration, air quality, gas resistance, temperature, humidity, and pressure, which means that breath characteristics are represented better. It is proved by the obtained cluster distribution, meaning that the selected features of sensors characterize the variation well in connection with pneumonia risk.

The distinct separation among the three clusters demonstrates the ability of the suggested preprocessing and K-Means clustering algorithm to classify unlabeled breath data into relevant pneumonia risk groups. The outputs from such clustered data form a solid basis for supervised classification and contribute to successful pneumonia risk prediction in real-time applications.

4.3 Decision Tree Classification Performance

The performance of the Decision Tree classifier model proposed was tested using the preprocessed breath sensor dataset along with the pseudo labels formed using K-Means clustering. The performance of the classification was measured using accuracy, precision, recall, and F1 score. The results obtained are provided in Table 4.

Table 4. Performance evaluation of the proposed decision tree classifier

Performance Metric	Value (%)
Accuracy	95.84
Precision	95.12
Recall	94.87
F1-Score	94.99

The proposed classifier was achieved an overall accuracy of 95.84%, which proved the ability of the model to classify breath samples in the three pneumonia risk categories. The high consistency in precision, recall, and F1-score reflects the consistency of the classification model, as well as the reliability of the predictions at any level of risk. This proves the effectiveness of the framework used in the pneumonia risk assessment.

5. Conclusion and Future Work

This research presented a breath-based pneumonia detection system incorporating a multi-sensor gas sensing system with machine learning that enables non-invasive and real-time

pneumonia detection. The system includes the MQ-3, MQ-135, and BME680 sensors to analyze the breath samples based on their properties related to respiratory diseases, where the K-Means clustering and Decision Tree classifiers help in conducting risk assessment and classification processes. It was found that the system provides a low-cost approach to pneumonia detection using multi-sensor breath analysis. The integration of embedded sensing and machine learning ensures fast decisions are made with low computing needs. In the future, efforts will be concentrated on increasing the breadth of the dataset using clinical validation, adding other VOC biomarkers for specific diseases, and exploring machine learning and deep learning algorithms to further improve the robustness and generalizability of the proposed system.

References

[1] Arulvasan, Wisenave, Julia Greenwood, Madeleine L. Ball, Hsuan Chou, Simon Coplowe, Owen Birch, Patrick Gordon et al. "Optimized Breath Analysis: Customized Analytical Methods and Enhanced Workflow for Broader Detection of VOCs." Metabolomics, 2025, vol. 21, no. 1: 17.

[2] Abdulsattar, Nadia Shamsulddin. "Pneumonia Classification from X-ray Images with Convolutional Neural Network." 12th International Conference on Information Technology (ICIT), IEEE 2025: 444-450.

[3] Zhang, Zhaokang, Kun Li, Guangpeng Zhang, Wenjun Liu, Xinglin Chen, and Zhaocheng He. "Design and Implementation of a High-Sensitivity Exhaled Breath Analyzer." 5th International Conference on Sensors and Information Technology, IEEE 2025: 122-126.

[4] Bharath, Mr PM, P. Kiruthika, and S. Madhusnegha. "AI-Based Respiratory Tracking and Detection System for Dyspnea." High Technology Letters, 2024, vol. 30, no. 4: 700–724.

[5] Belghiti, Mehdi Khaiaf, Mohammed El Gibari, Ahmed Rhallabi, and Amal El Fallah Serghrouchni. "Development and Optimization of Gas Sensors for Ammonia detection." 32nd Telecommunications Forum (TELFOR), IEEE 2024: 1-3.

[6] Arulvasan, Wisenave, Hsuan Chou, Julia Greenwood, Madeleine L. Ball, Owen Birch, Simon Coplowe, Patrick Gordon et al. "High-Quality Identification of Volatile Organic Compounds (VOCs) Originating from Breath." Metabolomics, 2024, vol. 20, no. 5: 102.

[7] Bekh, Igor, Serge Mamilov, Dmytro Velyhotskyi, Illya Porozinski, and Igor Yashan. "Prototype Device for Measuring Carbon Monoxide Levels in Human Exhalation." IEEE 42nd International Conference on Electronics and Nanotechnology (ELNANO), IEEE 2024: 364-368.

[8] Gupta, Rahul, Anil Kumar, Mukesh Kumar, Pradeep Kumar, and Dinesh Kumar. "Numerical Modelling of Cu2O-Based Gas Sensor for Detection of Volatile Organic Compounds in Human Breath." In International Conference on Emergent Converging Technologies and Biomedical Systems, Springer Nature Singapore 2022: 297-304.

[9] Sayo, Kevin, and Winncy Y. Du. "Portable Diabetic Ketoacidosis Detector via the Human Exhaled Breath Analysis." IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), IEEE 2023: 1212-1217.

[10] Massey, Roslyn S., Bruno Gamero, and Ravi Prakash. "A System-on-Board Integrated Multi-Analyte PoC biosensor for Combined Analysis of Saliva and Exhaled Breath." 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE 2022: 904-909.

[11] Xiong, Hangming, Jiaying Sun, Yingying Xue, Weijei Yu, Yong Zhou, Hao Wan, and Ping Wang. "Design of Breath Sampling Device and Procedure for Volatile Organic Compounds and Exhaled Breath Condensate." IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), IEEE 2022: 1-3.

[12] Human Breathomics Database (HBDB). A Comprehensive Database of Volatile Organic Compounds (VOCs) in Human Exhaled Breath. National Yang Ming Chiao Tung University. Available: https://hbdb.cmdm.tw/