ISSN 2395-8618

Analysis of Human Breath with E-Nose: Diabetes Mellitus Identification via SVD-XGBoost Algorithm

Alberto Gudiño-Ochoa, J. A. García-Rodríguez, Raquel Ochoa-Ornelas, Jorge Ivan Cuevas-Chávez, Daniel Alejandro Sánchez-Arias

***Abstract*—Diabetes is a highly prevalent chronic disease worldwide. Analysis of human breath has emerged as a non-invasive method for detecting various conditions using biomarkers. Electronic noses represent a crucial tool in this breath analysis for patients with diabetes mellitus, enabling early detection and diagnosis. This study involves the evaluation of 22 healthy patients and 20 patients with diabetes mellitus using an electronic nose employing catalytic Metal-Oxide-Semiconductor (MOS) gas sensors. A computational algorithm based on Singular Value Decomposition (SVD) for feature extraction and selection was utilized, coupled with classification using the Extreme Gradient Boosting (XGBoost) algorithm. The results demonstrate that classification of singular value vectors with the XGBoost algorithm achieves an accuracy of 95.24identifying healthy and diabetic patients. This approach shows significant potential for early diagnosis of diabetes through breath analysis, highlighting the effectiveness of electronic nose technology alongside advanced computational techniques in distinguishing between patient groups.**

***Index Terms*—Diabetes mellitus, electronic nose, breath analysis, XGBoost.**

I. Introduction

Diabetes is characterized by insufficient insulin production, which leads to imbalances in blood glucose levels (BGL) and triggers cardiovascular complications, eye problems, and limb issues, including amputations [1]. Although glucometers offer high precision, their invasive method is painful and uncomfortable, especially for frequent measurements throughout the day [2]. Recent clinical research has highlighted the potential of human breath analysis in medical diagnosis [3].

Exhaled volatile organic compounds (VOCs) contain endogenous biomarkers that differentiate healthy individuals from the sick. Figure 1 illustrates the compounds present in exhaled breath. In the case of diabetes, cellular inability to absorb glucose leads to an abnormal increase in ketone bodies, including acetone, a volatile compound exhaled by the body. Consequently, diabetic patients exhibit elevated concentrations of acetone in their breath [4, 5].

Complex techniques such as gas chromatography-mass spectrometry, mass spectrometry with selected ion flow tube, and cavity ring-down spectroscopy have demonstrated precision in breath analysis but are inadequate for clinical applications due to their lack of portability, complexity, and high costs [6, 7, 8]. In contrast, electronic noses have overcome these limitations, emerging as a promising alternative by providing substantial data on acetone concentrations in breath, a key biomarker [9, 10, 11].

The development of algorithms trained to detect diabetes based on feature selection and extraction is crucial, as demonstrated in previous studies using Principal Component Analysis (PCA), achieving accuracy greater than 90% in breath sample analysis [12, 13]. Regression models have helped predict glucose levels from breath samples [14]. For qualitative detection between healthy and diabetic patients, support vector machines (SVM), k-nearest neighbors (KNN), and various neural network variations have been explored, achieving 98% accuracy using convolutional neural networks (CNN) combined with SVM [13, 15, 16]. Previous clinical studies using deep neural networks (DNN) achieved 96.29% accuracy in detecting different levels of diabetes [17]. On the other hand, decision tree-based algorithms like XGBoost achieved 99 % accuracy in artificial breath analysis [18].

This research focuses on classifying breath samples from healthy patients and those with type 1 (T1DM) and type 2 diabetes (T2DM) using an electronic nose with MOS gas catalytic sensors, employing feature extraction and selection to differentiate acetone concentrations. The effectiveness of classification with advanced algorithms like XGBoost for feature extraction with Singular Value Decomposition (SVD) is demonstrated, providing reliable results in breath differentiation between healthy and diabetic patients, thus supporting the efficacy of electronic noses for medical diagnosis.

Manuscript received on 15/11/2023, accepted for publication on 02/02/2024.

A. Gudiño-Ochoa, J.I. Cuevas-Chávez and Daniel Sánchez-Arias are with Electronics Department, Tecnoló gico Nacional de México, Instituto Tecnológico de Ciudad Guzmán, Mexico (albertogudo@hotmail.com).

J.A. García Rodriguez is with Centro Universitario del Sur, Departamento de Ciencias Computacionales e Innovación Tecnológica, Universidad de Guadalajara, Mexico.

R. Ochoa-Ornelas is with Systems and Computation Department, Tecnológico Nacional de México/Instituto Tecnológico de Ciudad Guzmán, Mexico.

ISSN 2395-8618

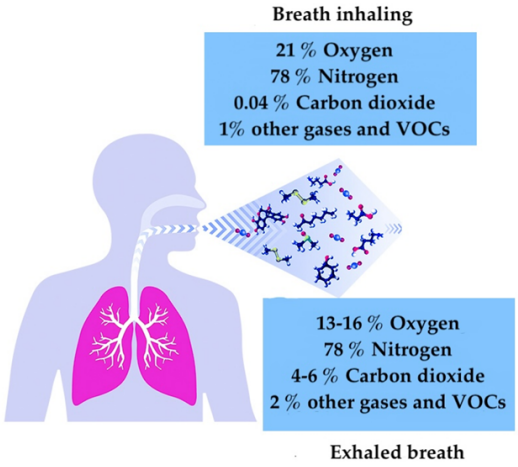

Fig. 1. Components in human breath when inhaling and exhaling

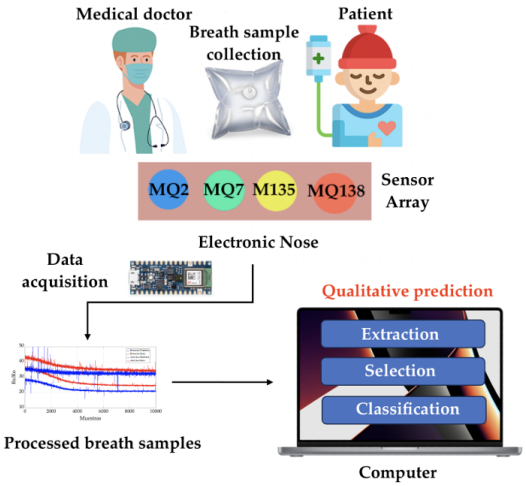

Fig. 2. Electronic nose application for the detection of diabetes mellitus

TABLE I
MATRIX OF SENSORS IN ELECTRONICNOSE

Number	Sensor	Measurement
1	MQ-2	Carbon monoxide
2	MQ-3	Alcohol
3	MQ-7	Carbon monoxide
4	MQ-135	Ketones
5	MQ-138	Acetone
6	DTH-22	Temperature and humidity
7	Mics-5521	VOCs

II. EXHALED BREATH ANALYSIS PROCEDURE

The electronic nose is composed of MQ series MOS sensors capable of identifying carbon monoxide, alcohol, acetone, ketones, VOCs, temperature, and relative humidity. The MOS sensors were previously calibrated, as they require preheating for at least 24 to 48 hours in advance. An Arduino Nano 33 BLE Sense board was incorporated due to its 12-bit ADC sampling rate with a 32-bit ARM Cortex-M4 processor. In addition to measuring the acetone biomarker in diabetes patients, gas sensors detecting other VOCs present in the breath of healthy patients were added. Table I provides a detailed matrix of sensors.

The dataset consisted of real measurements from 22 healthy patients and 20 patients with T1DM and T2DM, considering variability in ages, sampling times, and blood glucose levels. Table II shows the physical information of the patients included in this study.

In Figure 2, the procedure for collecting exhaled breath from the patients is depicted. A direct method of breath collection was employed using Tedlar medical bags to store the exhaled breath. Upon collecting the patient sample, it was manually transferred to the sample chamber in the electronic nose [13]. Figure 3 shows the breath collection of a patient with type 2 diabetes mellitus, the same procedure that was applied to the patients who participated in the present study. Measurements were initiated by connecting to the serial port and coding with Python for 90 seconds, generating 10,000 samples within that

Fig. 3. Breath sample collection from patient with T2DM

time frame. These measurements were stored along with the response information from each sensor in comma-separated values (CSV) files for subsequent analysis of the VOCs.

A. Signal preprocessing

Due to the noise present in the measurements caused by changes in the temperature and humidity of the breath [13], the Discrete Wavelet Transform (DWT) with a low-pass filter was applied to eliminate noise from the acquired signals. Subsequently, the measurements of each patient were normalized (standardized) to reduce variability in the samples and minimum values in parts per million (ppm) of the readings.

ISSN 2395-8618

TABLE II
COMPARISON BETWEEN HEALTHY PATIENTS AND PATIENTS WITH DIABETES MELLITUS.

Variable	HealthyPatients (22)	PatientswithDiabetes (20)
Age (years)	23.64±2.19	29.95±4.24
Height (m)	1.72±0.12	1.69±0.12
Weight (kg)	72.16±10.38	76.80±11.42
BMI (kg/m2)	24.47±4.02	27.33±6.20
Gender (M/f)5	13/9	9/11
Type of Diabetes (T1DM/T2DM)	-	6/14
Minimum and maximum Glucose Level (mg/dL)	80.59/94.63	199.28/303.10

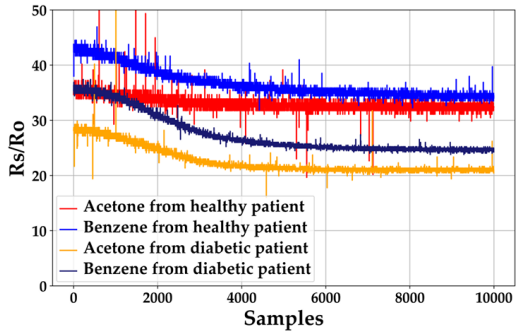

Fig. 4. Response of MOS sensor signals between a healthy patient and a patient with T2DM

This was done by establishing the Rs/Ro value in response to the MQ sensors, ensuring that no attribute is too dominant over others [17]. Figure 4 shows the relationship of the Rs/Ro values of the MQ-135 and MQ-138 sensors in response to the exhaled breath of a healthy patient and a patient with diabetes, where a lower Rs/Ro value indicates a higher concentration in ppm [16].

B. Feature extraction: SVD

After standardizing the response information of each sensor per patient, the characteristics of the data are selected and extracted using SVD. This technique in linear algebra breaks down a matrix A into three main components:

1) Matrix of left singular vectors U,
2) Diagonal matrix S containing the singular values,
3) Matrix of transposed right singular vectors V T.

The mathematical expression of the SVD decomposition for an m×n matrix A is defined as:

$$PrecisionA = U\Sigma V^T, \tag{1}$$

where:

- U is an $m \times m$ orthogonal matrix whose columns are the left singular vectors of A.
- Σ is an $m \times n$ diagonal matrix containing the singular values of A.
- V^T is the matrix transpose of an n × n orthogonal matrix whose columns are the right singular vectors of A.

The singular values of A represented by σ_i are found on the diagonal of Σ, ordered from largest to smallest,

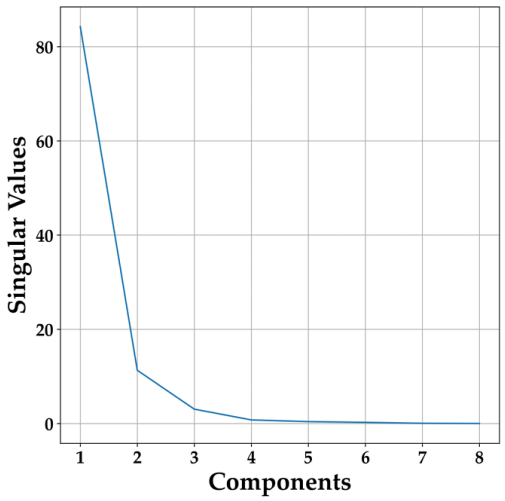

Fig. 5. Contribution of singular values according to the number of components

being indicators of the importance of each singular vector in the reconstruction of the original matrix A. The number of singular values determines the effective dimension to reduce the dimensionality of A.

Figure 5 displays the information provided by each singular value, where the first 4 components reach 99.33% of the information, allowing the implementation of a classification model. To obtain the dimensionality reduction, the reduced matrix A_k was obtained by retaining only the k most important singular values, resulting in a k-rank approximation of matrix A:

$$A_k = U_k \Sigma_k V_k^T, \tag{2}$$

where U_k, Σ_k and V_k^T are the reduced matrices with the k most important singular values. The following equations are used to calculate U, Σ and V_k^T:

- U is obtained from the eigenvectors of AA^T,
- Σ is formed by the singular values σ_i on the diagonal,
- V^T is obtained from the $A^T A$ eigenvectors.

Figure 6 plots the set of values obtained using the reduced matrix method with SVD. Distinct clusters are identified between the groups of healthy patients and those with diabetes mellitus. In addition, a second data set was generated using the left singular vector method as shown in Figure 7.

ISSN 2395-8618

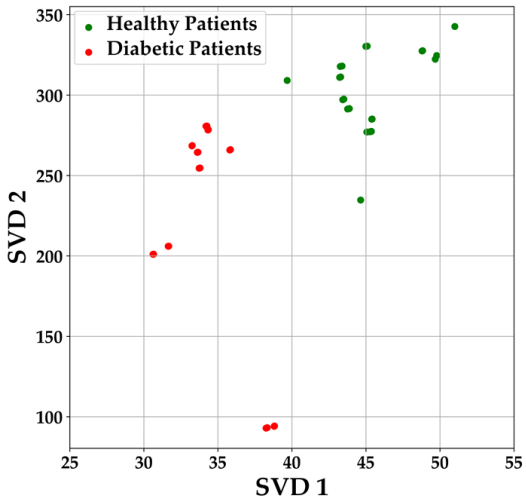

Fig. 6. Reduced matrix feature extraction using SVD

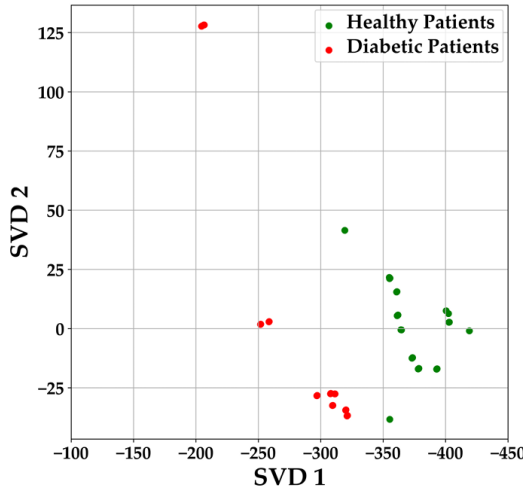

Fig. 7. Feature extraction with singular vectors on the left with SVD

These results demonstrate the feasibility of both feature extraction and feature selection techniques for a gradient-boosting-based model. This makes possible the construction of a sequence of predictive models.

C. Anomaly detection with One-Class SVM

Anomaly detection using One-Class SVM was applied to identify unusual patterns in the data that could influence classification. This technique models the common breath characteristics of healthy individuals, allowing detection of significant deviations that could indicate chemical differences associated with diabetes. Data were stored and labeled to determine the individual's health status. Figure 8 shows clusters among the normal values, suggesting that the data set is appropriate for classification.

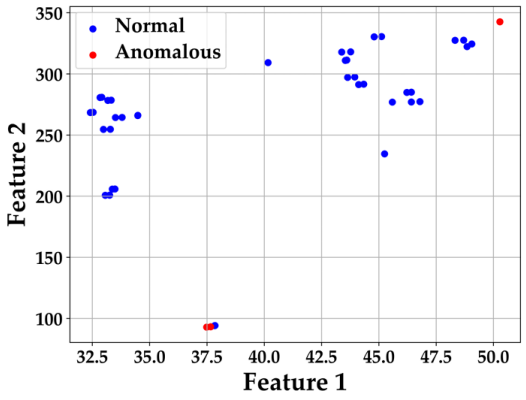

Fig. 8. Anomaly detection with One-Class SVM on data extracted with SVD

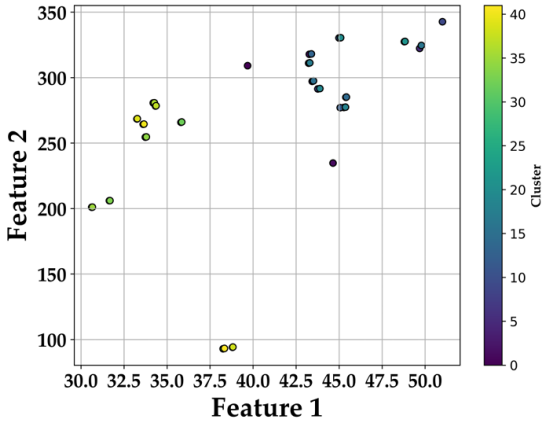

Fig. 9. DBSCAN classification of values extracted with SVD

III. Classification Results

A. Unsupervised learning: DBSCAN

In the analysis prior to the machine learning classification stage, a validation of the data was performed using an unsupervised clustering technique (DBSCAN). This allowed clusters in the data sets to be identified based on the density of the points. The parameters of the unsupervised classifier were set to an epsilon (ϵ) value of 0.2 and a minimum number of center points of 1 (MinPts), which resulted in the identification of patients with diabetes, as evidenced by the yellow-greenish hues in Figure 9, based on their elevated NGL levels.

To address the distinction between healthy patients and those with diabetes mellitus, we divided the data set into two equal parts: 50% for model training and the remaining 50% for the test set. Given the sample size limitation of healthy patients and patients with diabetes mellitus, the percentage was adequate to avoid the problem of overfitting. The XGBoost model was configured as a binary "logistic" type classification, integrating the "log-loss" evaluation metric together with lambda and alpha regularizers to avoid overfitting during the

ISSN 2395-8618

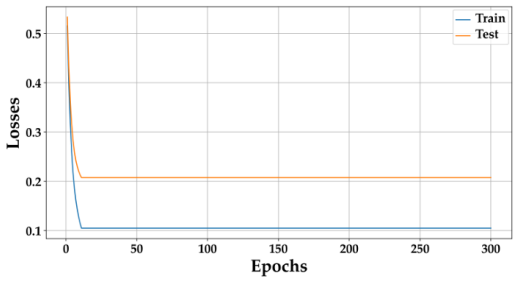

Fig. 10. DBSCAN classification of values extracted with SVD

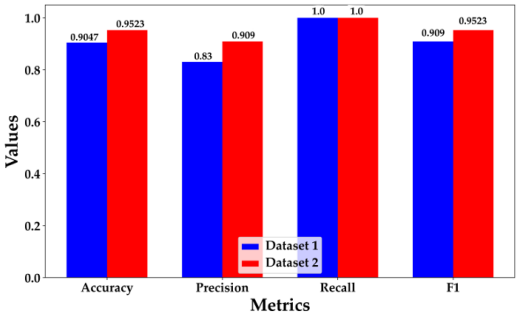

Fig. 11. Comparison of evaluation metrics between SVD sets integrated to the XGBoost model

training. We chose 300 iterations (same number of trees) for the model's training.

Figure 10 presents the evaluation of losses in both the training and test sets over time using the data obtained from the left-hand singular vector SVD set. Despite achieving a classification accuracy of 95.24%, the loss assessment indicates that a larger sample size could enhance the model's accuracy using test data from both healthy and diabetic patients. We then compared two sets: set 1, which represented features extracted by reduced matrix SVD, and set 2, which represented features extracted by left singular vector SVD. Figure 11 shows scoring and evaluation metrics that reveal a significant improvement in classification accuracy, precision, and F-1 score, improving from 90.47% to 95.24%.

B. Classification with XGBoost

The XGBoost algorithm confirms the effectiveness of the SVD extraction method using singular vectors for classification. In addition, the confusion matrix in Figure 12 reveals a single false positive in the classification between 11 healthy patients and 10 patients with diabetes. A ROC plot in Figure 13 with an area under the curve (AUC) of 0.9545 complements this finding, confirming the robustness of the model to distinguish between both patient groups.

We suggest that enhancing the classification process with XGBoost, which extracts features from human breath analysis

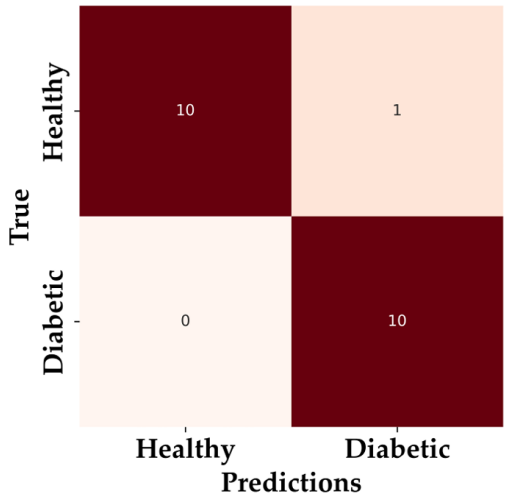

Fig. 12. Confusion matrix of the SVD-XGBoost model

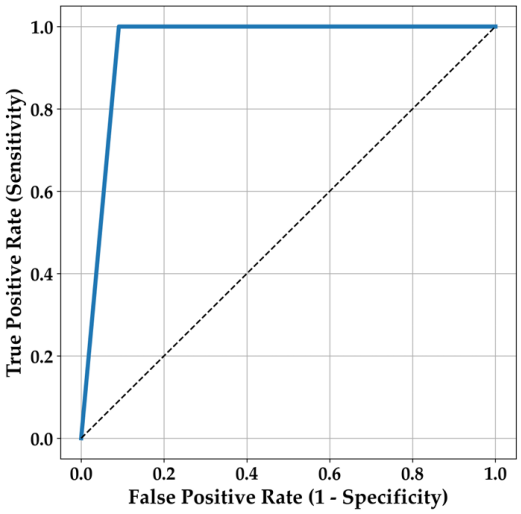

Fig. 13. Fig. 13. ROC Curve.

and VOCs to identify patients with diabetes mellitus, could boost the accuracy to 95.24% by increasing the number of patient tests and their NGL variability.

IV. Conclusion

The study has demonstrated the XGBoost algorithm's effectiveness in the classification of features extracted and selected by SVD in human breath analysis to identify diabetes mellitus. The SVD extraction technique, especially through singular vectors, has been highlighted as a reliable tool for this classification. We can detect abnormalities in the breath of diabetic patients by training the XGBoost model with data from healthy individuals and setting a cutoff for "normal". Increasing the number of tests and the variability of NGL between patients can improve this non-invasive and effective

ISSN 2395-8618

diagnosis method. These results provide a noninvasive tool for analyzing exhaled human breath and detecting diabetes mellitus. Medical treatment will depend on the type of diabetes and NGL concentration, as well as other types of clinical tests for diagnosis and habits that improve the patient's quality of life.

ACKNOWLEDGMENTS

This manuscript was supported by resources from the PROSNII-2023 program requested by Dr. Julio García-Rodríguez at the CUSUR University of Guadalajara. In addition, thanks to the Department of Computer Science and Technological Innovation of CUSUR, University of Guadalajara. All authors also thank the Instituto Tecnológico de Ciudad Guzmán (ITCG) for their support.

REFERENCES

[1] C. Petry, "Gestational diabetes: Risk factors and recent advances in its genetics and treatment," *The British journal of nutrition*, vol. 104, pp. 775–87, 2010.

[2] J. Yang, X. Yang, D. Zhao, X. Wang, W. Wei, and H. Yuan, "Association of time in range, as assessed by continuous glucose monitoring, with painful diabetic polyneuropathy," *Journal of Diabetes Investigation,* vol. 12, no. 5, pp. 828–836, 2021.

[3] E. M. Gaffney, K. Lim, and S. D. Minteer, "Breath biosensing: using electrochemical enzymatic sensors for detection of biomarkers in human breath," *Current Opinion in Electrochemistry*, vol. 23, pp. 26–30, 2020.

[4] Z. Wang and C. Wang, "Is breath acetone a biomarker of diabetes? a historical review on breath acetone measurements," *Journal of breath research,* vol. 7, p. 037109, 2013.

[5] F. Gouzi, D. Ayache, C. H´edon, N. Molinari, and A. Vicet, "Breath acetone concentration: too heterogeneous to constitute a diagnosis or prognosis biomarker in heart failure? a systematic review and meta-analysis," *Journal of Breath Research,* vol. 16, no. 1, p. 016001, 2021. DOI:10.1088/1752-7163/ac356d.

[6] T. Saidi, O. Zaim, M. Moufid, N. El Bari, R. Ionescu, and B. Bouchikhi, "Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects," *Sensors and Actuators B: Chemical,* vol. 257, pp. 178–188, 2018.

[7] M. . Storer, J. Dummer, H. Lunt, J. Scotter, F. McCartin, J. Cook, M. Swanney, D. Kendall, F. Logan, and M. Epton, "Measurement of breath acetone concentrations by selected ion flow tube mass spectrometry in type 2 diabetes," *Journal of breath research,* vol. 5, p. 046011, 2011.

[8] Z. Xue, W. Hongmei, G. Dianlong, L. Yan, X. Lei, H. Chaoqun, S. Chengyin, and C. Yannan, "On-line monitoring human breath acetone during exercise and diet by proton transfer reaction mass spectrometry," *Bioanalysis,* vol. 11, no. 1, pp. 33–40, 2019. DOI:10.4155/bio-2018-0258.

[9] P. Montuschi, N. Mores, A. Trové, C. Mondino, and P. J. Barnes, "The Electronic Nose in Respiratory Medicine," Respiration, vol. 85, no. 1, pp. 72–84, 2012. DOI:10.1159/000340044.

[10] A. D. Wilson, "Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath", *Metabolites,* vol. 5, no. 1, pp. 140–163, 2015.

[11] V. R. Nidheesh, K. M. Aswini, V. K. Unnikrishnan, K. S. Rajeev, N. Rajesh, B. K. Vasudevan, and S. Chidangil, "Breath analysis for the screening and diagnosis of diseases," *Applied Spectroscopy Reviews*, vol. 56, no. 8-10, pp. 702–732, 2021. DOI:10.1080/05704928.2020.1848857.

[12] S. Lekha and M. Suchetha, "Recent advancements and future prospects on e-nose sensors technology and machine learning approaches for non-invasive diabetes diagnosis: A review,´´ *IEEE Reviews in Biomedical Engineering,* vol. 14, pp. 127–138, 2020.

[13] A. Paleczek and A. Rydosz, "Review of the algorithms used in exhaled breath analysis for the detection of diabetes,´´ *Journal of Breath Research,* vol. 16, no. 2, p. 026003, 2022. DOI:10.1088/1752-7163/ac4916

[14] Z. Ye, J. Wang, H. Hua, X. Zhou, and Q. Li, "Precise detection and quantitative prediction of blood glucose level with an electronic nose system," *IEEE Sensors Journal,* vol. 22, no. 13, pp. 12452–12459, 2022.

[15] S. Lekha and M. Suchetha, "A novel 1-d convolution neural network with svm architecture for real-time detection applications," *IEEE Sensors Journal,* vol. 18, no. 2, pp. 724–731, 2018.

[16] A. Gudiño-Ochoa, J. A. García-Rodríguez, R. Ochoa-Ornelas, J. I. Cuevas-Chávez, and D. A. Sánchez-Arias, "Noninvasive Diabetes Detection through Human Breath Using TinyML-Powered E-Nose," *Sensors,* vol. 24, no. 4, 2024.

[17] R. Sarno, S. Izza Sabilla, D. Rahman Wijaya, and Hariyanto, "Electronic nose for detecting multilevel diabetes using optimized deep neural network," *Engineering Letters,* vol. 28, no. 1, pp. 31–42, 2020.

[18] A. Paleczek, D. Grochala, and A. Rydosz, "Artificial breath classification using xgboost algorithm for diabetes detection," *Sensors,* vol. 21, no. 12, 2021.