FIELD OF SCIENCE ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE AUTOMATION, ELECTRONICS, ELECTRICAL ENGINEERING AND SPACE TECHNOLOGIES

DOCTORAL DISSERTATION

Machine Learning in Electronic Nose Systems: Estimation of Metabolic Parameters and Disease Detection

Author: mgr inż. Anna Magdalena Paleczek

Supervisor: prof. dr hab. inż. Artur Rydosz

Completed at: AGH University of Krakow,
Faculty of Computer Science, Electronics and Telecommunications

Kraków, 2025

AGH

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

DZIEDZINA NAUK INŻYNIERYJNO-TECHNICZNYCH

DYSCYPLINA AUTOMATYKA, ELEKTRONIKA, ELEKTROTECHNIKA i TECHNOLOGIE KOSMICZNE

ROZPRAWA DOKTORSKA

Uczenie maszynowe w systemach elektronicznego nosa: estymacja parametrów metabolicznych i detekcja chorób

Autor: mgr inż. Anna Magdalena Paleczek

Promotor rozprawy: prof. dr hab. inż. Artur Rydosz

Praca wykonana: Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Informatyki, Elektroniki i Telekomunikacji

Kraków, 2025

I would like to sincerely thank my Supervisor prof. dr hab. inż. Artur Rydosz for guiding my scientific development and for his expertise and support throughout this dissertation.

I am grateful to all the members of the Biomarkers Analysis LAB AGH research group for their collaboration and inspiring discussions. Special thanks go to dr inż. Dominik Grochala for his valuable guidance and support during the experimental work.

I also thank the Department of Prosthodontics and Orthodontics at the Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland for providing the opportunity to conduct the medical experiment, and lek. dent. Justyna Grochala for leading the research on the medical side.

Finally, I am grateful to my family and friends for their support and encouragement throughout this work, with special thanks to my Mum for always believing in me.

Acknowledgements

This research work has been co-financially supported by the AGH - IDUB Excellence Initiative – Research University Grant, “Analysis of exhaled air as an innovative tool for medical diagnosis” ID: IDUB 4122 and National Science Centre (NCN) under SONATA grant, “Investigation of the influence of GLAD technique for 3S properties (sensitivity, selectivity, stability) of gas sensors with the enhanced response for diabetes biomarkers in exhaled human breath”, ID: 2017/26/D/ST7/00355.

The research was conducted in cooperation with the Department of Prosthodontics and Orthodontics at the Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland under the medical experiment entitled „Research on selected risk factors of metabolic syndrome and their impact on the condition of the prosthetic base”. Bioethics Committee 1072.6120.40.2023 (granted 14/06/2023).

Abstract

In recent years, there has been an increase in the number of individuals affected by metabolic disorders, such as abdominal obesity, insulin resistance, hypertension, and dyslipidaemia, which significantly elevate the risk of cardiovascular diseases and type II diabetes. Early monitoring of metabolic parameters enables the rapid detection of abnormalities and the implementation of preventive measures, including lifestyle changes, which can reduce the risk of complications.

Analysis of exhaled breath represents a promising diagnostic tool that has garnered considerable attention in scientific research over the past few years. Apart from the major compounds of exhaled air, such as nitrogen, carbon dioxide and oxygen, exhaled breath contains thousands of volatile organic compounds that can serve as indicators of physiological processes occurring in the body. Breath can be analysed using precise laboratory instruments, such as gas chromatographs coupled with mass spectrometry, or with electronic noses composed of a matrix of gas sensors. Due to the multidimensional nature of the data and the complexity of breath composition, machine learning algorithms are usually employed to analyse signals from breath-analysis devices, with the aim of disease detection and prediction of health parameters.

The literature review on diseases studied for detection through exhaled breath analysis has also been included in the dissertation to provide an overview of current challenges from both medical and engineering perspectives. In this thesis, the Author also discusses the current applications of data pre-processing methods and various machine learning and Artificial Intelligence algorithms for analysing signals obtained from systems based on electronic noses.

The research activities conducted during the doctoral process include laboratory studies in which an electronic nose system and machine learning algorithms were developed to detect diabetes in simulated exhaled breath, as well as the clinical evaluation during the medical experiment. The developed e-nose system achieved accuracy of 99% for a diabetes detection in simulated exhaled breath. The prediction of acetone concentration (a diabetes biomarker) in gas mixtures achieved a mean absolute error of 0.248 ppm. However, in the presence of high ethanol concentrations, which serve as an interfering factor, the mean error increased to 0.568 ppm. Combined with classification algorithms, the electronic nose was able to distinguish three metabolic states based on synthetic gas mixtures with accuracies of 95%, 79%, and 88% for samples simulating healthy individuals, prediabetic individuals, and diabetic patients, respectively.

In a medical experiment conducted with the approval of the Jagiellonian University Bioethical Committee (KBET: 1072.6120.40.2023) on a group of 151 participants, it was demonstrated that the developed electronic nose system, supported by machine learning algorithms, can predict total cholesterol, glucose, and uric acid levels. The mean absolute errors were 31.33 mg/dL for total cholesterol, 19.32 mg/dL for glucose, and 1.43 mg/dL for uric acid, respectively.

The results presented in this doctoral dissertation confirm the potential for developing a portable, non-invasive device for the early detection of metabolic disorders, thereby enabling faster treatment and the prevention of complications, such as cardiovascular diseases and diabetes.

Streszczenie

W ostatnich latach obserwuje się wzrost liczby osób dotkniętych zaburzeniami metabolicznymi, takimi jak otyłość brzuszna, insulinooporność, nadciśnienie tętnicze czy dyslipidemia, co znacząco zwiększa ryzyko wystąpienia chorób sercowo-naczyniowych oraz cukrzycy typu II. Wczesne monitorowanie parametrów metabolicznych pozwala na szybką detekcję nieprawidłowości i wdrożenie działań profilaktycznych, w tym zmiany stylu życia, co może ograniczyć ryzyko wystąpienia powikłań.

Analiza wydychanego powietrza stanowi obecnie obiecujące narzędzie diagnostyczne, które w ostatnich latach zyskało dużą popularność w badaniach naukowych. Wydychane powietrze zawiera tysiące lotnych związków organicznych, które mogą służyć jako wskaźniki zachodzących w organizmie procesów. Oddech może być analizowany za pomocą laboratoryjnych, precyzyjnych urządzeń jak chromatografy gazowe sprzężone ze spektrometrią masową lub z wykorzystaniem elektronicznego nosa, składającego się z matrycy sensorów gazowych. Ze względu na wielowymiarowość danych oraz złożoność składu wydychanego powietrza do analizy sygnałów z urządzeń analizujących oddech stosowane są algorytmy uczenia maszynowego, których zadaniem jest detekcja chorób i predykcja parametrów zdrowotnych.

W niniejszej rozprawie zamieszczono również przegląd literatury dotyczący chorób, których wykrywanie możliwe jest poprzez analizę wydychanego powietrza, aby przedstawić przegląd obecnych wyzwań z perspektywy medycznej i inżynieryjnej. Omówiono również dotychczasowe zastosowania metod wstępnej obróbki danych oraz różnych algorytmów uczenia maszynowego i sztucznej inteligencji do analizy sygnałów uzyskiwanych z systemów bazujących na elektronicznych nosach.

Prace badawcze prowadzone w ramach pracy doktorskiej obejmują badania laboratoryjne, w ramach których opracowano system elektronicznego nosa i algorytmy uczenia maszynowego umożliwiające wykrywanie cukrzycy w symulowanym wydechu, a także ocenę kliniczną w trakcie eksperymentu medycznego. System osiągnął dokładność detekcji cukrzycy, w symulowanym oddechu, na poziomie 99%. Predykcja stężenia acetonu (biomarkera cukrzycy), w mieszankach gazowych osiągnęła średni błąd bezwzględny 0,248 ppm, natomiast w obecności w mieszankach wysokich stężeń etanolu - jako czynnika interferencyjnego - średni błąd wyniósł 0,568 ppm. W połączeniu z algorytmami klasyfikacji elektroniczny nos umożliwił rozróżnienie trzech stanów metabolicznych na podstawie

syntetycznych mieszanek gazowych z precyzją odpowiednio 95%, 79% i 88% dla próbek symulujących osoby zdrowe, w stanie przedcukrzycowym oraz chore na cukrzycę.

W ramach eksperymentu medycznego, przeprowadzonego za zgodą Komisji Bioetycznej UJ (KBET: 1072.6120.40.2023) na grupie 151 osób, wykazano, że system elektronicznego wspierany algorytmami uczenia maszynowego umożliwia przewidywanie poziomów cholesterolu całkowitego, glukozy oraz kwasu moczowego. Średni błąd bezwzględny wyniósł odpowiednio 31,33 mg/dl dla całkowitego cholesterolu, 19,32 mg/dl dla glukozy oraz 1,43 mg/dl dla kwasu moczowego.

Wyniki badań zaprezentowane w tej rozprawie doktorskiej potwierdzają możliwość opracowania przenośnego, nieinwazyjnego urządzenia do wczesnego wykrywania zaburzeń metabolicznych, co pozwala na szybsze leczenie i zapobieganie powikłaniom, takim jak cukrzyca czy choroby układu krążenia.

Table of Contents

List of papers

This thesis is based on and incorporates the following papers:

Chapter 2:

[AP1] A. Paleczek and A. Rydosz, ‘Review of the algorithms used in exhaled breath analysis for the detection of diabetes’, *J Breath Res*, vol. 16, no. 2, p. 026003, Jan. 2022, doi: 10.1088/1752-7163/AC4916.

[AP2] A. Paleczek, ‘Recent achievements of exhaled breath analysis at the research stage—Artificial intelligence and machine learning algorithms’, *Exhaled Breath Analysis*, pp. 325–355, Jan. 2025, doi: 10.1016/B978-0-443-33796-3.00005-2.

Chapter 3:

[AP3] A. Paleczek, D. Grochala, and A. Rydosz, ‘Artificial breath classification using XGBoost algorithm for diabetes detection’, *Sensors*, vol. 21, no. 12, 2021, doi: 10.3390/s21124187.

[AP4] A. Paleczek and A. Rydosz, ‘The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies’, *Sens Actuators B Chem*, vol. 408, p. 135550, Jun. 2024, doi: 10.1016/J.SNB.2024.135550.

[AP5] A. Paleczek, D. Grochala, and A. Rydosz, ‘Diabetes classification and acetone concentrations prediction in gas mixtures with high ethanol content’, 2024.

Chapter 4:

[AP6] A. Paleczek *et al.*, ‘Noninvasive Total Cholesterol Level Measurement Using an E-Nose System and Machine Learning on Exhaled Breath Samples’, *ACS Sens*, Nov. 2024, doi: 10.1021/ACSSENSORS.4C02198.

[AP7] A. Paleczek et al. ‘Revolutionizing Health Monitoring: A Three-Gas Sensor System Powered by Machine Learning for Predicting Cholesterol, Glucose, and Uric Acid Levels from Exhaled Breath’, 2025

List of Abbreviations

AI	– Artificial Intelligence
ATS	– American Thoracic Society
AUC	– Area Under Curve
BMI	– Body mass index
CAS	– Chemical Abstracts Service
CKD	– Chronic kidney disease
COPD	– Chronic obstructive pulmonary disease
CRC	– Colorectal cancer
e-nose	– Electronic nose
EBA	– Exhaled breath analysis
FeNO	– Fractional nitric oxide
FTIR	– Fourier Transform Infrared Spectroscopy
GC-MS	– Gas chromatography coupled with mass spectrometry
HDL	– High-Density Lipoprotein
IR	– Infrared Spectroscopy
kNN	– k-Nearest Neighbours
LDA	– Linear Discriminant Analysis
LDL	– Low-Density Lipoprotein
LightGBM	– Light Gradient Boosting Machine
MAE	– Mean absolute error
MAPE	– Mean absolute percentage error
ML	– Machine learning
MOS/MOX	– Metal-Oxide Semiconductor
PCA	– Principal Component Analysis

PTR-MS	– Proton Transfer Reaction Mass Spectrometry
RH	– Relative humidity
ROC	– Receiver operating characteristic curve
SD	– Standard deviation
SIBO	– Small intestinal bacterial overgrowth
SIFT-MS	– Selected Ion Flow Tube Mass Spectrometry
SVM	– Support Vector Machines
UBT	– Urea breath test
VOCs	– Volatile organic compounds
WHO	– World Health Organization
XGBoost	– eXtremeGradientBoosting

List of Figures

1. Introduction

Exhaled breath analysis (EBA) is currently one of the most promising areas of development in non-invasive medical diagnostic tools. Human breath contains hundreds and even thousands of volatile organic compounds (VOCs), which can be products of metabolic processes occurring within the body (endogenous biomarkers) or result from environmental influences (exogenous biomarkers) [1]. Their chemical profile can provide valuable information about a patient's health and indicate the presence of specific disorders. The advantages of this method include its complete non-invasiveness, safety, and potential use in screening and disease monitoring.

In recent years, there has been a growing interest in the use of electronic nose (e-nose) systems for breath analysis [2], [3], [4]. These systems comprise a set of selected gas sensors and data processing modules, often supported by machine learning (ML) methods, that enable the identification of complex patterns that occur in sensor signals. This enables not only the detection of individual biomarkers but also the classification of entire respiratory profiles characteristic of specific diseases. Different diseases produce different patterns in breath; sometimes it is a change in the level of one or more biomarkers [5], [6], [7], [8], [9], and sometimes it is a change in the entire VOC profile [10], [11]. Exhaled breath analysis is an approved diagnostic method for detecting conditions such as asthma [12], [13] or lactose intolerance [14]. Research into metabolic diseases, such as diabetes and hypercholesterolemia, is critical, as early detection and monitoring are crucial for maintaining population health [15], [16], [17], [18].

Despite the great potential of this method, several research challenges remain. The VOCs profile is complex and can be influenced by numerous environmental and physiological factors. Classical analysis methods are insufficient for interpreting complex and nonlinear signals generated by sensors in the e-nose system. Therefore, machine learning algorithms play a crucial role in pattern recognition in multidimensional data, predicting biochemical parameter values, and classifying samples corresponding to various health states.

This doctoral dissertation was written as a series of research papers and focuses on the application of machine learning algorithms to the analysis of e-nose data in the context of detecting metabolic diseases. This work aimed to develop and validate an approach combining laboratory studies on model gas mixtures with the analysis of exhaled air samples from patients.

The research hypothesis: the e-nose system, supported by dedicated machine learning algorithms, can effectively predict selected metabolic parameters and classify samples according to patient health status.

The dissertation is structured in **four** main parts.

Chapter 2 presents a literature review on exhaled breath analysis in disease diagnosis, with particular emphasis on the role of e-nose systems and machine learning algorithms. This part contains two research papers: a book chapter discussing the application of e-nose and ML in medicine [AP1], and a review paper on machine learning algorithms used in diabetes detection [AP2].

Chapter 3 presents the results of laboratory studies conducted on artificially prepared gas mixtures, including an analysis of the feasibility of detecting acetone and classifying samples. This chapter contains three research papers concerning [AP3, AP4, AP5]: the detection of acetone as a biomarker of diabetes, the prediction of its concentration in the presence of ethanol, and the classification of samples corresponding to various diabetes states.

Chapter 4 presents the results of clinical studies conducted on a group of 151 individuals, including the prediction of cholesterol, glucose, and uric acid concentrations based on exhaled air analysis. This chapter contains two research papers: a full paper on cholesterol prediction [AP6] and a conference paper [AP7] on the prediction of glucose and uric acid levels.

The final chapter, **Chapter 5**, summarises the results and indicates directions for further research and potential applications of the developed system in clinical practice.

In summary, this work presents a consistent series of studies on the use of the electronic nose and machine learning algorithms in the diagnosis of metabolic diseases, focusing on both the laboratory and clinical phases. The obtained results confirm the potential of this approach as a non-invasive, safe, and user-friendly method that could find applications in future medical practices.

2. Machine learning algorithms for disease detection in exhaled breath using e-nose systems

Exhaled breath analysis is a promising diagnostic tool that has attracted significant attention in clinical research over the past few years. Breath contains up to thousands of volatile organic compounds, which act as biomarkers for processes occurring within the human body. These biomarkers can be classified into endogenous, resulting from metabolic and biochemical activities, and exogenous, originating from external sources such as the environment, diet, or smoking [1]. Variations in the concentrations or ratios of these compounds can indicate the presence of specific diseases, sometimes before known clinical symptoms appear. This technique is entirely non-invasive, safe, quick, and well-accepted by patients, contributing to its increasing popularity and application in screening and health monitoring. An electronic nose for EBA is a device equipped with a set of chemical sensors that respond to the presence of various volatile compounds in exhaled air. Scheme of breath analysis measurement techniques is shown in Figure 2.1.

Typical sensors used in the e-nose include [19], [20]:

- Metal-Oxide Semiconductor (MOS/MOX) sensors - respond to oxidising or reducing gases by a change in conductivity. MOS sensors are highly sensitive, but affected by temperature and humidity and have limited selectivity [21].
- Electrochemical sensors - generate an electrical signal as a result of chemical reactions between gases at electrodes. Electrochemical sensors are highly selective and used to detect individual gases at low concentrations [22].
- Optical sensors - use absorption, fluorescence, or changes in light intensity in response to the presence of specific gases. Optical sensors offer high selectivity and sensitivity; however, they are more expensive, consume more power, and are larger than MOS or electrochemical sensors. They are also sensitive to dirt and moisture, detecting only selected radiation-absorbing gases (e.g., CO_2, CH_4, CO), rather than the entire spectrum of compounds present in breath [23], [24], [25].

Exhaled air can also be analysed using precise laboratory techniques, such as:

- Gas chromatography coupled with mass spectrometry (GC-MS) - enables the identification and quantification of individual volatile compounds [4], [26].
- Infrared spectroscopy (IR, FTIR) - allows the detection of characteristic absorption bands of gases [27], [28].
- Electron ion spectroscopy (PTR-MS, SIFT-MS) - enables the rapid and direct detection of many VOCs in real time [29], [30].

Unlike these precise laboratory methods, the e-nose commonly focuses on recognising characteristic sensor response patterns, known as breathprints [31]. These patterns reflect the unique composition of VOCs in the sample, including both endogenous and exogenous biomarkers. Sometimes e-nose systems are designed to detect the concentration of selected biomarkers. The e-nose offers several advantages, including fast analysis, eliminating the need for complicated sample preparation, and the possibility of making the device small enough to be portable [32]. This makes it suitable for point-of-care diagnostic and screening systems.

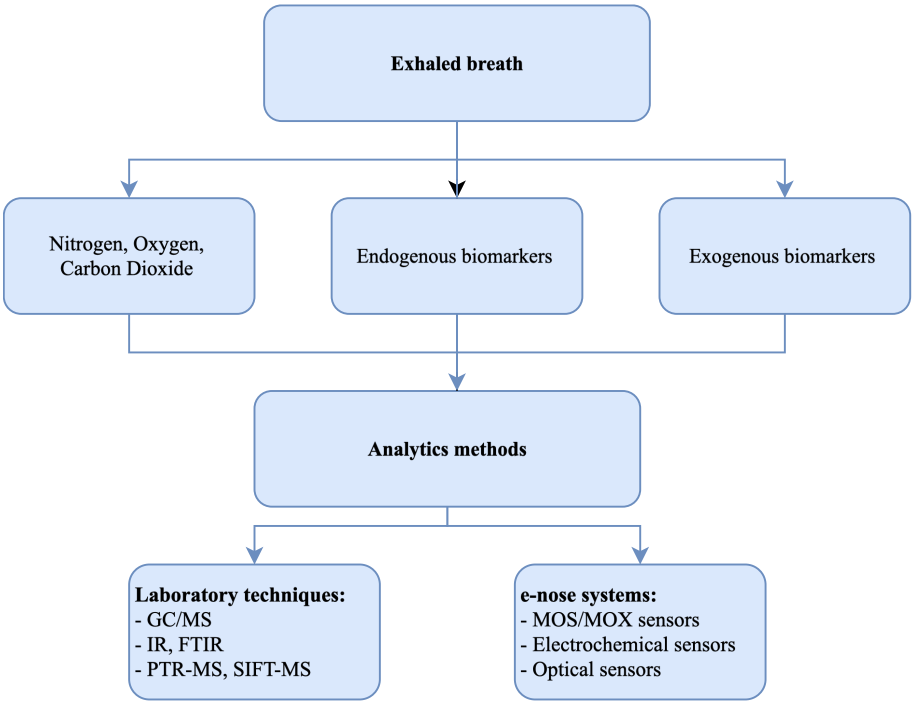


Fig. 2.1. Breath analysis scheme.

In general, exhaled breath analysis can be done in two ways: online, where the patient blows directly into the device [25], [33], [34], and offline [35], [36], where the breath is collected in special bags [37], [38], [39] or specially designed breath samplers like ReCIVA® [40] and analysed later.

Despite the increasing use of electronic noses in medical diagnostics, these systems face significant limitations arising from both sensor properties and the nature of exhaled air.

First, the signals produced by e-nose sensors are characterised by high noise levels, caused by random fluctuations and environmental interference. Many gas sensor types can also exhibit time drift, which is a gradual change in response unrelated to variations in gas concentration. This drift may result from material ageing, adsorption of compounds, or changes in operating conditions [41], [42], [43].

Second, measurement results are influenced by environmental factors such as temperature, humidity, and the presence of other VOCs in the breath sample. In exhaled breath analysis, high breath humidity is the main factor affecting sensor response. Additionally, signals can vary depending on individual patient factors, such as diet, lifestyle and metabolic rhythm, introducing notable inter-individual variability [44], [45].

Another limitation involves the high complexity and nonlinearity of the signals, stemming from interactions among multiple compounds within a sample and the sensor array's characteristics. Furthermore, some sensors are cross-sensitive, responding to multiple compounds, which complicates interpretation and highlights the need to use an e-nose system supported by signal analysis and machine learning algorithms [18], [46], [47].

The challenges above show why we need advanced approaches to process and analyse data. Signal processing can involve removing noise, adjusting signals to a common scale, correcting sensor drift, and reducing the amount of data while keeping the important information. These methods help make signals more stable and easier to compare, even when there are short-term changes in the environment or small differences between sensors.

Machine learning algorithms facilitate the identification of subtle patterns within the multidimensional, nonlinear data generated by the sensor array. Most popular techniques include, for example, Support Vector Machines (SVM) [48], [49], Decision Trees [50], eXtremeGradientBoosting (XGBoost) [49], [51], and even more sophisticated approaches such as Deep Learning [52], [53], ensemble methods and gradient boosted algorithms.

The combination of an appropriately designed e-nose system (including hardware and measurement technology, data pre-processing techniques, and machine learning algorithms) can classify breath samples to identify specific diseases or metabolic disorders.

It can also estimate health parameters that are typically measured from blood samples. The use of machine learning algorithms can also reduce the effects of environmental and inter-individual variability. A key step in the analysis is ensuring the interpretability of the algorithms' results and decisions, which helps identify the most crucial signal features, supports interpretation and system optimisation, and builds trust among physicians and patients.

In practice, combining the e-nose with machine learning creates screening tools that are highly accurate and sensitive, enabling the early detection of diseases with minimal patient discomfort. This approach supports both clinical research and the creation of real-time health monitoring and screening systems that can be used directly at the point of care. The e-nose development roadmap is shown in Figure 2.2.

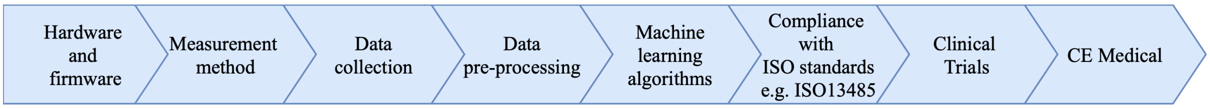


Fig. 2.2. E-nose development-to-market roadmap.

This chapter first provides an overview of diseases detectable via exhaled breath analysis (Section 2.1). Subsequently, it discusses the role of Artificial Intelligence (AI) in e-nose data processing (Section 2.2) and reviews algorithms and gas sensors specifically applied to diabetes detection (Section 2.3).

2.1. Overview of Diseases Detectable Through Exhaled Breath Analysis

Medical literature increasingly emphasises the fact that exhaled air can reflect the health of many organs and body systems. Characteristic exhaled breath profiles can signal the presence of, for example, metabolic [54], [55], cancer [56], [57], [58], [59], [60], respiratory [2], [61], [62] diseases. In clinical practice, breath analysis enables the early detection of certain conditions, often before visible symptoms appear.

The complexity and variability of these signals require the use of appropriate data analysis methods. Several machine learning and signal processing algorithms have been described in the literature to support the classification of breath samples, the identification of subtle differences between patient groups, and the potential correlations with health status. These techniques include supervised, unsupervised, and hybrid methods, which allow for the extraction of meaningful information from large, multidimensional, and nonlinear datasets.

2.1.1. Cancers

Cancer is a group of diseases in which abnormal cells begin to grow uncontrollably and can spread to other parts of the body. According to the World Health Organization (WHO), it is the second leading cause of death worldwide – in 2018, an estimated 9.6 million people died from cancer, accounting for 1 in 6 deaths. The most common cancers in men include lung, prostate, colon, stomach, and liver cancers, while in women, the most common are breast, colon, lung, cervical, and thyroid cancers [63].

The global cancer problem is growing, having a huge impact on the health, lives, and finances of people and healthcare systems. The WHO estimates that 30% to 50% of cancer deaths could be avoided. This can be achieved by making healthier lifestyle choices, such as not smoking, eating a balanced diet, staying active, and limiting alcohol consumption, and by utilising proven prevention strategies, including vaccinations and regular health check-ups. By taking these steps, many cancers could be prevented before they even start. Early detection and effective treatment significantly increase the chances of survival and reduce side effects and treatment costs [63].

There is growing hope for breath analysis, which may pave the way for rapid, non-invasive, and early diagnosis, providing the opportunity for more effective treatment even before symptoms appear. Scientists are currently analysing human breath to detect non-invasively: breast [57], [64], [65], [66], lung [4], [20], [67], colorectal [64], [68], [69], [70], prostate [64], [71], [72], [73] and thyroid [3], [74] cancers.

In the case of colorectal cancer (CRC), an increase in the level of alcohols, ketones, aldehydes [60] and commonly a difference in the exhaled level of dinitrogen oxide, nitrous acid, acetic acid, xylene, 1,3-butadiene [56], ammonia, ethanol, propanol [75], ethylbenzene, methylbenzene, and tetradecane has been found. For example, Haick *et al.* patented 1,3,5-cycloheptatriene as a CRC biomarker, which is not observed in breath samples from patients with other cancers [59], [76]. Zonta *et al.* used a fabricated 5-sensor array, principal component analysis (PCA) and SVM to detect cancer in collected breath samples. They achieved sensitivity and specificity at 95% [69]. Malagu *et al.* proposed a sensor array composed of twelve metal-oxide semiconducting films. They tested its responses for gas mixtures containing benzene, methane, and nitrogen oxide as a potential biomarker of CRC [68]. The Aeonose (The eNose Company), which includes three metal-oxide sensors, was used to analyse 511 exhaled breath samples. Sensor data were compressed using a Tucker3-like solution, and an artificial neural network was trained, resulting in an area under curve (AUC) of 0.84. The cited studies demonstrate that a specific set of VOCs is characteristic of CRC, and the use of an array of sensors combined with machine learning algorithms enables the detection of CRC in exhaled air with high accuracy.

Binson *et al.* conducted a study using an electronic nose system to detect lung cancer by analysing VOCs in exhaled air. The study included 22 patients with lung cancer and 40 healthy controls. Five gas sensors were used in the e-nose system (TGS2600, TGS2620, TGS822, TGS826, TGS2610), which are low-cost, fast-response, and low-power. Data from the sensors were analysed using three classification algorithms: linear discriminant analysis (LDA), k-Nearest Neighbours (kNN), and SVM. The LDA algorithm achieved the best results, with a classification accuracy of 93.14%, a sensitivity of 88.63%, a specificity of 95.62%, and an area under the receiver operating characteristic curve (ROC) curve of 0.98. The authors indicated that the e-nose system shows excellent potential for rapid and non-invasive detection of lung cancer. However, further research is needed on sensor stability and increasing sample representativeness [67].

2.1.2. Respiratory diseases

Respiratory diseases encompass a wide range of conditions affecting the airways and lungs. These include asthma, chronic obstructive pulmonary disease (COPD), respiratory infections, cystic fibrosis, and lung cancer. COPD and asthma are common respiratory diseases with a significant impact on public health. According to the WHO, COPD causes about 3.23 million deaths worldwide each year. Asthma affects more than 262 million people and leads to around 461,000 deaths annually. COPD makes it difficult to breathe and often causes cough and shortness of breath [77]. Asthma is characterised by reversible narrowing of the airways, attacks of shortness of breath, and wheezing, frequently triggered by allergens, infections, or physical exertion. Respiratory infections, both viral and bacterial, can cause bronchitis and pneumonia, and in some cases, lead to chronic changes in lung tissue [78]. Early diagnosis of COPD and asthma allows for the implementation of appropriate treatment and prevention, significantly reducing the frequency of exacerbations and limiting permanent lung damage. Furthermore, it allows for individualised therapy and patient education, improving quality of life and respiratory function.

Chronic respiratory diseases such as COPD and asthma are associated with chronic lung inflammation, which increases susceptibility to tissue damage and the development of lung cancer. People with COPD are particularly susceptible to lung cancer, while asthma, although less likely to lead directly to cancer, can promote changes in lung tissue due to long-term inflammation. Methods for detecting and distinguishing these diseases based on exhaled breath analysis are gaining popularity and hold promise for non-invasive diagnostics [2], [61], [62].

The study by Fens *et al.* involved 100 patients: 21 with fixed airways obstruction (fixed asthma), 39 with reversible airways obstruction (classic asthma), and 40 patients with COPD (GOLD stages II-III). A Cyranose 320 electronic nose, equipped with an array of 32 carbon-black polymer sensors responding with a change in resistance to volatile organic compounds, was used to record respiratory profiles. The acquired signals created characteristic "breath fingerprints," which were then analysed using principal component analysis for dimensionality reduction and canonical discriminant analysis for patient classification. The results showed that the method effectively distinguished between fixed asthma and COPD, as well as classic asthma, achieving accuracies of 88% and 83%, respectively, and high AUC values above 0.93, indicating the eNose's significant potential as a non-invasive tool supporting the differential diagnosis of obstructive lung diseases [62].

A study by de Vries *et al.* evaluated whether exhaled breath analysis using the e-nose scan detect early lung cancer in patients with COPD. A total of 682 patients with COPD and 211 patients with lung cancer participated. Within 2 years of study entry, 37 (5.4%) COPD patients developed lung cancer. The eNose SpiroNose (metal-oxide semiconductor) was integrated with a pneumotachograph (SpiroNose; Breathomix) for measurements, allowing for the collection of breath samples in a clinical setting. The collected data - subjected to advanced processing, correction for ambient air effects, and principal component analysis - were then classified using LDA and evaluated using ROC curves. Distinguishing patients with COPD from those with lung cancer achieved an AUC of 0.89 and predicting which COPD patients would develop cancer within 2 years of enrolment, the model achieved an AUC of 0.90 with an accuracy (cross-validation) of 87%. The results suggest that the eNose may be a non-invasive tool for not only distinguishing COPD from lung cancer but also for early detection of lung cancer in patients with COPD [61].

2.1.3. Chronic kidney disease

Chronic kidney disease (CKD) is a condition that affects the kidneys' ability to function correctly. In the early stages, it often shows no symptoms, but over time it can cause fatigue, blood in the urine, and swelling, especially in people with diabetes, high blood pressure, or those taking certain long-term medications. Because CKD allows toxins to accumulate in the body, it alters the composition of blood, urine, saliva, and even breath - changes that can be monitored to help detect and track the disease and the effectiveness of the haemodialysis. New technologies, such as AI-powered electronic noses, are making it possible to do this in a simple, non-invasive, and affordable manner [79].

Guo *et al.* developed an electronic nose system enhanced with machine learning algorithms to detect breath samples from healthy individuals as well as from those with diabetes, kidney disease, or respiratory inflammation. The system can also be used to evaluate the effectiveness of haemodialysis. In their study, they used an array of 12 metal-oxide semiconductor gas sensors housed in a steel measurement chamber. To test the system's accuracy in classifying different breath samples, training and test sets were randomly selected for each disease group, and features were extracted using PCA. Classification was then performed using a kNN with k = 5. The system achieved an average classification accuracy of 80.15% for samples taken before haemodialysis and 82.0% for samples after haemodialysis, demonstrating its potential for non-invasive monitoring of treatment effectiveness. When distinguishing between healthy breath samples and those from individuals with diabetes, kidney disease, or respiratory inflammation, the e-nose demonstrated high sensitivity, at 87.67% for diabetes, 86.57% for kidney disease, and 70.20% for airway inflammation. Similarly, it showed high specificity at 86.87%, 83.47%, and 75.07%, respectively. These results indicate strong performance in detecting diabetes and kidney disease, with slightly lower accuracy in detecting respiratory inflammation [80].

Another study on the effectiveness of dialysis was conducted by Jayasree *et al.* The authors proposed a system for detecting ammonia in the exhaled breath of patients with renal failure using MOS sensors and an SVM classifier. Analysis of samples from 40 patients, both before and after dialysis, enabled the extraction of geometric features such as rise time, peak time, and maximum voltage. The best results were obtained with the TGS2444 sensor, achieving a classification accuracy of 88% using all three features. The MQ137 and MQ135 sensors correctly classified most of the predialysis samples, but the algorithm performed less well in classifying the postdialysis group [81].

2.1.4. Use of Breath Tests in Clinical Diagnostics

2.1.4.1. Helicobacter pylori – Urea Breath Test (UBT)

Helicobacter pylori is a bacterium that colonises the gastric mucosa and can lead to chronic inflammation, gastric and duodenal ulcers, and increase the risk of developing stomach cancer. The urea breath test (UBT) involves the ingestion of urea labelled with a carbon isotope (^{13}C or ^{14}C). The bacterium breaks down urea into ammonia and carbon dioxide, which are released into the exhaled air and measured using a detector. The test is non-invasive, rapid, and highly sensitive, and is used both to detect infections and to monitor the effectiveness of eradication therapy [82].

2.1.4.2. Carbohydrate Intolerances and SIBO – Hydrogen Tests

Lactose, fructose, or sorbitol intolerances, as well as small intestinal bacterial overgrowth (SIBO), cause gastrointestinal symptoms such as bloating, diarrhoea, and abdominal pain. Breath tests involve consuming a specific sugar and then measuring the concentration of hydrogen and/or methane in exhaled air over the following hours. Gut bacteria ferment unabsorbed sugars, producing gases that are exhaled. An increase in hydrogen or methane above normal levels indicates digestive disorders or the presence of SIBO. The test is safe, non-invasive, and commonly used to diagnose intestinal disorders [83], [84].

2.1.4.3. Asthma – Fractional Nitric Oxide Measurement

Measuring fractional nitric oxide (FeNO) in exhaled air enables the assessment of the degree of eosinophilic bronchitis, a characteristic of allergic asthma and can be used as a complementary diagnostic tool according to the Official Clinical Practice Guideline developed by the American Thoracic Society (ATS) [85]. Higher FeNO values indicate active inflammation and aid in selecting the appropriate inhaled corticosteroid therapy. The test is non-invasive, quick, and repeatable, and its results support both diagnosis and monitoring of therapy effectiveness [13], [86].

2.1.5. Other diseases that can be detected by breath testing

Diagnosing diseases based on exhaled air is a rapidly evolving field, and researchers are seeking correlations between various diseases and the volatile organic compounds present in breath. In addition to the diseases presented in this chapter, researchers are particularly interested in detecting diabetes, metabolic syndrome, and ketosis through the analysis of specific biomarkers in breath. The most important marker of ketosis is acetone [87], [88], whose concentration increases in uncontrolled diabetes and during ketosis. Other volatile organic compounds, such as isoprene [18], [89] and aldehydes, may indicate oxidative stress, lipid disorders, and early symptoms of metabolic syndrome [90]. Diabetes detection using breath analysis by the e-nose system supported by machine learning algorithms is discussed in detail in Section 2.3.

Additionally, work is underway on the detection of diseases from breath, such as neurodegenerative diseases (Alzheimer's disease, Parkinson's disease), cardiovascular diseases, SARS-CoV-2, and halitosis. Details are discussed in Chapter 4 of the book Exhaled Breath Analysis [91].

Breath analysis has broad and diverse applications in scientific research, and its current use in clinical practice highlights both the potential and the need for further development of this field. Advances in e-nose systems and artificial intelligence algorithms create a real possibility of introducing additional clinically validated, non-invasive disease detection methods that can significantly improve patient diagnosis and monitoring.

2.2. Recent Achievements of Exhaled Breath Analysis at the Research Stage - Artificial Intelligence and Machine Learning Algorithms

CHAPTER

7

Recent achievements of exhaled breath analysis at the research stage—Artificial intelligence and machine learning algorithms

Anna Paleczek

AGH University of Krakow, Biomarkers Analysis LAB, Institute of Electronics, Krakow, Poland

7.1 Introduction

Breath sample analysis is a complex task. Samples can be measured both by advanced methods such as gas chromatography and mass spectrometry as well as by using e-noses (commercial and specially manufactured for specific studies). In each of these cases, a large amount of multidimensional data is created, which makes it impossible to analyze with the naked eye and diagnose the disease or draw conclusions from it. In many cases, data are even impossible to visualize in 2D and 3D space. For this reason, data preprocessing methods and machine learning algorithms are used, which enable reducing the dimensionality of data, data analysis and visualizations, disease classification, and, increasingly often, the interpretation of algorithm decisions (Fig. 7.1).

The individual steps of the breath sample processing pipeline are as follows:

1. **Data collection**—Training machine learning algorithms in medical cases is challenging due to ethical and regulatory constraints, patient recruitment, and logistical issues with collecting exhaled air samples, which can be performed online or offline. Accurate breath analysis with AI/ML algorithms requires comprehensive patient data to focus on correlation VOC profiles or specific compounds with diseases and blood parameters.
2. **Data preprocessing**—Data from gas sensors are most often the results of measurements of electrical values affected by noise and drift; therefore, the use of preprocessing techniques significantly improves the effectiveness of regression or signal classification methods. One of the stages of data preprocessing is normalization (standardization).
3. **Feature engineering**—This is one part of the breath analysis data processing pipeline. During this stage, based on the course of the sensor response to the

Exhaled Breath Analysis. https://doi.org/10.1016/B978-0-443-33796-3.00005-2

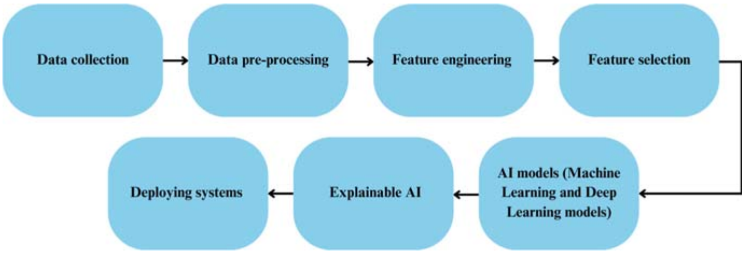


FIGURE 7.1

General Block Diagram of AI/ML Breath Analysis Systems.
Presents a general block diagram showing the procedure of analyzing breath samples from taking a sample from people to obtaining a visualization and final diagnosis, and in some cases, to deploying the solution to web, IoT, or embedded platforms.

breath sample, features are determined using signal processing methods and domain knowledge. The use of feature engineering increases the efficiency of the algorithms and allows the conversion of multidimensional time series data into a one-dimensional feature vector, which can be further used to train machine learning algorithms. In addition, researchers use clinical data from patients that can affect the composition of the breath, such as age, gender, and smoking status, which increase the efficiency of the algorithms.

4 **Feature selection**—The aim of this step is to emphasize the importance of feature selection in data processing, especially when handling a large number of features from multiple sensors. Using too many features can lead to overfitting and poor generalization, so dimensionality reduction techniques like PCA and LDA, as well as other methods such as ICA and KPCA, are employed to reduce the feature space. These techniques aim to retain essential information or maximize class separation. In addition, recursive feature elimination, greedy search, and regularization methods like Lasso and its variations (e.g., SGL) are used in breath analysis tasks, showing their effectiveness in improving model performance and identifying important features.

5 **AI models (Machine Learning and Deep Learning models)**—This part is the most complex stage of the pipeline. It consists of the appropriate selection of the algorithm for the task, depending on the type of data and the availability of labels. Most often, several algorithms from the same type of learning are tested and the best one is selected based on the appropriate metrics. During training, it is important to properly interpret its course as well as metrics to obtain the best model and avoid underfitting or overfitting the model.

6 **Explainable AI**—AI algorithms often operate as black boxes, providing results like disease diagnoses without insight into their decision-making process. This lack of transparency is risky, especially in medicine, highlighting the need for

explainable AI to identify errors and discover new disease factors. Understanding feature impact improves trust among doctors and patients and enables sensor optimization in breath-based disease detection. This can minimize device size, reduce costs and energy consumption, and enhance accessibility.

7. **Deploying systems**—Many e-nose systems are developed primarily on data collected online and then analyzed using machine learning algorithms. To create a portable, widely available e-nose device for diagnosing diseases based on the analysis of exhaled air, it is necessary to deploy the entire pipeline, including preprocessing and model interference to enable real-time diagnosis. For this purpose, various platforms are used, often based on the Internet of Things technology.

7.2 Data collection

To train machine learning algorithms, it is necessary to have a large data set. In cases involving data from humans, that is, in medical cases, including the analysis of exhaled air, this is often a difficult task, requiring the consent of the bioethics committee, preparation of an appropriate protocol, conducting research in accordance with the Declaration of Helsinki, local statutory requirements, as well as compliance with personal data protection regulations (e.g., GDPR) [1]. It is also often difficult to find the right group of patients with specific health parameters or rare diseases. Exhaled air can be collected online (directly to the device) [2] and offline (into specialist bags, e.g., Tedlar^{O} [3–6]), which is also a limitation and results in difficulty in analyzing breaths. For online measurement, it is necessary for patients to come to a scientific laboratory or transport the device to a medical facility, which is often difficult to organize. By contrast, in the case of offline measurements, the breath sample can be collected in a special bag, but there are also limitations in its storage and transport, which is why the analysis device should be located close to the breath collection point, for example, a hospital.

In addition to the breath sample, it is necessary to collect data about the patient and their medical history to train machine learning algorithms. Patient data often include venous or capillary blood test results and also imaging test results [7–10]. In addition, data on medications taken are collected as is demographic data of the patients (age, gender, body weight, medical history). In the case of breath sample analysis, it is also important to have information on whether the patient is a smoker or former smoker and obtain data relating to recent intake of food and fluids [7–9]. Such external factors can have a key impact on the results of breath tests.

There are several ways to assess the physiological state of the body through breath tests and their analysis using machine learning algorithms:

- Analysis of the correlation of the entire respiratory profile (VOCs profile) with a given disease entity [11–16],

- Detection of a specific compound from the breath and association of its presence with the disease [17–20],
- Detection of a specific compound (several defined compounds) in the breath and correlation of its/their concentrations with a blood parameter, for example, blood glucose level [21,22] or cholesterol level [10].

Paying great attention to collecting breath samples and patient information is a crucial stage in creating machine learning algorithms.

7.3 Data preprocessing

Data preprocessing is the first step in preparing the collected sensor data for further processing using machine learning and AI algorithms. Data from gas sensors are most often measured electrical values that are burdened with measurement errors and noise or drift related to the nature of the sensor layers. Data quality significantly affects the training process and the efficiency of models. Researchers use various signal processing techniques such as filtering or normalization to deal with the abovementioned problems to prepare data for the next step of the data processing pipeline. The most popular techniques are discussed in this subsection.

7.3.1 Filtering

One of the most popular filtering methods is the window-mean filter (given by Eq. (7.1)). A window-mean filter is a smoothing technique that replaces each data point in a time series or signal with the mean of neighboring values within a defined window. This helps reduce noise and fluctuations, preserving trends while averaging out short-term variations [23]. Assume that x_k (where $k \in$ *[0,M]*) is the raw signal and N is the window length; the filtered signal value y_i is calculated using Eq. (7.1)., where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values in the filter window, respectively.

$$y_i = \left(\sum_{k=i-N+1}^{i} \frac{x_k - x_m ax - x_m in}{(N-2)} \right), i = N-1, N, \ldots, M \tag{7.1}$$

A similar method was used by Polaka et al.—the median filter differs from the mean filter in that, instead of replacing each point with the mean of the values in the window, it replaces it with the median [24]. The median filter is better at removing impulse noise (so-called "salt-and-pepper noise") because the median is not as susceptible to extreme values as the mean.

7.3.2 Baseline subtraction (sensor drift reduction)

Sensor drift refers to a gradual change in sensor sensitivity over time caused by factors like sensing material or substrate aging, slow material morphological evolution observed in long-term tests (as shown in Fig. 7.2) [25]. This slow process impacts

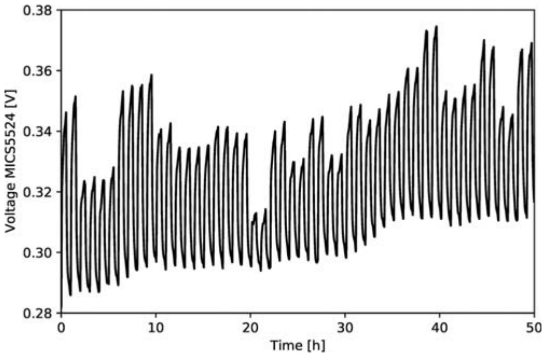


FIGURE 7.2

Long-term stability test of MICS5524 gas sensor.

Creative Commons Attribution (CC BY) license [27].

signal accuracy, requiring correction and continuous recalibration to maintain sensor performance and cannot be omitted during analysis, especially during the creation of medical devices for diagnosis based on exhaled breath testing.

The possibility of using DWT to filter signals from gas sensors and eliminate drift was demonstrated by Zuppa et al. [25]. They demonstrated the effectiveness of discrete wavelet transform (DWT) in recovering sensor signals affected by low-frequency drift. By decomposing signals into low- and high-frequency components at multiple scales, DWT isolates the drifts while preserving the signal trend. The authors showed that removing drift using DWT helps increase cluster separation in principal component analysis (PCA). Moreover, Ye et al. used DWT to remove noise in the resistance signals [22].

One of the techniques for sensor response drift minimization is baseline normalization. Ye et al. proposed defining normalized resistance (R_i) as a ratio between resistance change (Eq. 7.2.) where R_{aroma} is measured sensor resistance [22].

$$R_i = \frac{(R_a\text{roma} - R_b\text{aseline})}{R_b\text{aseline}} \tag{7.2}$$

Another technique proposed by Binson et al. [26] was the manipulation of the baseline expressed by the mathematical formula 7.3.

$$\text{SR}\frac{\text{baseline}}{(S, D, T)} = \text{SR}\frac{}{(S, D, T)} - \frac{1}{\text{BN}}\sum_{t=1}^{\text{BN}}\text{SR}(S, D, T) \tag{7.3}$$

where

$\text{SR}\frac{\text{baseline}}{(S, D, T)}$ the sensor response after baseline manipulation to sameple $S(= 1, 2, ., \text{Bn})$ from sensor $D(D = 1, 2, 000, \text{Dn})$ at time $T(T = 1, 2, .)$

A different approach was used by Paleczek et al. The authors used curve fitting to the minimum values of the response of the sensor during the purge stage (response in pure air) [27] shown in Fig. 7.3. Another interesting method that enables real-time baseline tracking was proposed by Wang et al. Its advantage is that real-time baseline compensation and use in portable analyzers are possible [28]. Martin et al. proposed creating a calibration curve constructed by measuring the sensor response to the same concentrations in synthetic gas mixtures at the start of sensor use and after 1 year [1].

7.3.3 Data normalization and standardization

Since data collected from different sensors are usually of different scales, one of the stages of data preprocessing is normalization and standardization. These techniques enable comparison of data collected at different scales and facilitate statistical analyses, as well as training machine learning algorithms that are sensitive to different feature scales. Models can perform poorly if features have different scales, as large values could dominate the learning process [29].

The first of the most popular methods, which is not only used in the analysis of breaths and gas sensor data, is normalization. By using formula 7.4., the values are transformed to the range [0, 1] [30].

$$x^{'} = \frac{X - x_{\min}}{(x_{\max} - x_{\min})} \tag{7.4}$$

where

- x is the original value,
- $x_{\min}$ is the minimum value in the dataset,
- $x_{\max}$ is the maximum value in the dataset,

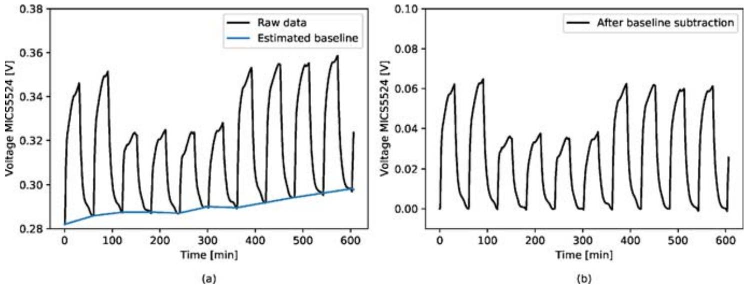


FIGURE 7.3

(a) Sensor response before baseline subtraction and (b) sensor response after baseline subtraction using a curve-fitting method.

Creative Commons Attribution (CC BY) license [27].

- x' is the normalized value

The standard scaler method transforms data to have a mean of 0 and a standard deviation of one by applying feature-wise standardization (given by formula 7.5). This method also helps to effectively train the model particularly in algorithms sensitive to feature scaling (e.g., support vector machines, k-nearest neighbors, and gradient descent) and helps ensure that features contribute equally to model training by adopting a similar scale [29].

$$z = \frac{(x - \mu)}{\sigma} \tag{7.5}$$

where

- x is the original value,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset,
- z is the standardized value.

Another method, used by Binson et al., is called auto-scaling and involves two steps: centering and scaling. In the centering step, the data from each sensor (or column) are normalized by subtracting its mean, resulting in variables with a mean of zero. The scaling step then adjusts the data to ensure all variables have the same scale, providing equal weight to each parameter [31].

7.3.4 Detection and handling of outliers

The presence of outliers also has a negative impact on the training and effectiveness of machine learning algorithms. Identifying and removing them improve the generalization of the model and statistical analysis. Statistical tests, visualization techniques (like box plots or scatter plots), and machine learning algorithms (like isolation forests or clustering methods) are used to detect outliers, which can then be deleted or replaced with other values (imputed). Polaka et al. [24] used principal components analysis and calculated orthogonal and score distances and then evaluated them by the method presented by Rodionova and Pomerantsev [32] to identify outliers.

7.3.5 Handling imbalanced data

In the case of medical data, imbalanced data are often observed. It is most often observed in the case of data classification when the class sizes are not close to each other. This can cause problems in training machine learning algorithms because they can favor the majority class. Therefore, it is important to analyze the class sizes and the distributions of the variables [33]. There are several different techniques to deal with class imbalance:

- Undersampling—reducing data samples from majority class.

- Oversampling—creating more examples for minority class.

The disadvantage of undersampling is the limitation of the number of samples in the data set, which can affect the efficiency of the algorithms and their ability to generalize the sample, while the disadvantage of oversampling is the difficulty of generating samples that effectively represent the minority class and do not introduce unnecessary noise. One such technique, often used in the analysis of breath sample data, is the synthetic minority oversampling technique (SMOTE). In this method, a given number of neighbors of the original sample, for example, $k = 5$, are identified, and based on this sample and random neighbors, SMOTE interpolates new points and adds them to the data. This method was used by Chen et al. [34], while a modification of the method, ADASYN-SMOTE, was used by Kapur et al. [7]. It differs from the basic version in that ADASYN (Adaptive Synthetic Sampling) adaptively adjusts the number of generated synthetic points to the local data density, which can improve the efficiency of the model in classification. El-Magd et al. [33] used support vector machine (SVM) to compute the class weighting which is later used to balance classes during training.

In addition to these techniques, it is important to use appropriate metrics which take into account class imbalance to evaluate classification efficiency.

7.3.6 Temperature and humidity compensation

In e-nose systems, additional sensors are often used to measure the temperature and humidity of the sample because the gas sensors used (most often MOX) are prone to change their responses depending on gas conditions and not only on compounds present in gas mixtures [22,27,35]. In the case of breath sample measurements, the influence of relative humidity is the most important factor that cannot be omitted as relative humidity in breath samples is higher than 89% [36].

7.4 Feature engineering

Data from e-noses often takes the form of waveforms such as those shown in Figs. 7.2 and 7.3. In the study of breath samples using gas sensors, we most often distinguish two stages: purging of sensors (using room air, clean synthetic air, etc.) and a sample dosing stage. The result of measuring signals from several sensors (sensor array) is multivariate time series data. Such data can be processed using, for example, recurrent neural networks, while to use classic machine learning or deep learning algorithms, applying feature engineering is necessary. The most used feature engineering techniques in breath measurements using sensor technology use the R_0 (sensor response in purge stage) and R_G (sensor response in sample dosing stage) values read from the graphs, as shown in Fig. 7.4. The response of the sensor R_0 and R_G can be a different electrical parameter, for example, conductance [1], resistance [37], voltage [22], or digital, depending on the measurement method used.

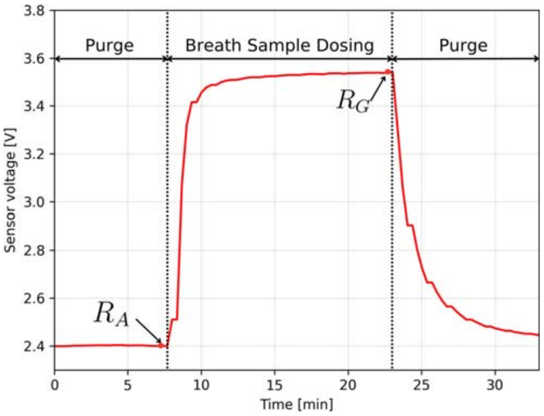


FIGURE 7.4

Sensor response curve.

Reproduced from Ref. [10] CCBY 4.0.

The most used basic feature engineering methods with equations based on R_0 and R_G values are listed in Table 7.1.

In addition to basic feature engineering techniques, researchers propose more advanced signal processing techniques to extract as much information as possible from the results of breath measurements using gas sensors.

Kapur et al. designed a GlucoBreath system, in which they used a wide range of generated features. In addition to the basic features similar to those listed in Table 7.1, they also used calculated statistical values from parameters such as curve

Table 7.1 Basic feature engineering techniques.

Technique	Equation	References
Minimum value of curve	min(curve)	[2]
Average value of curve	average(curve)	[2,31]
Maximum value of curve	max(curve)	[2,31]
Mean value of the last N time points to characterize the sensor response after saturation	$\frac{1}{N}\sum_{n=1}^{N} \mathrm{RGN}$	[2]
Area under the curve	$\int_{\mathrm{Ro}}^{\mathrm{RG}} f(t)\mathrm{dt}$	[2]
Time of maximum sensor response	time(max(curve))	[31]
Sensor response	$S = \frac{R_0}{R_G} - 1$	[37]
Sensor response	$S = R_G - R_0$	[1,38]
Sensor relative response	$S = \frac{R_G}{Ro}$	[38]

magnitude, first and second derivatives, as well as coefficients calculated using Fast Fourier Transform (FFT), Auto-regressive (AR) analysis and DWT. In addition, they used the phase, and five equal distance intervals were created from the sensor's response voltages for a breath sample and the slope and integral over these intervals were calculated. The authors also analyzed the shape of the signals by calculating the skewness, kurtosis, and entropy of the signal curve [7].

Hao et al. extracted features in the time domain, frequency domain, and statistics parameters. The authors obtained 14 time-domain features (e.g., maximum, minimum, mean, peak-to-peak value), 14 frequency-domain features (e.g., center of gravity frequency, frequency variance, power spectrum), and 10 statistical features (e.g., skewness, kurtosis, autocorrelation, information entropy) [30].

Sarno et al. used Principal component analysis (PCA) for feature extraction. Six features were derived from the sensors: CO gas, CO_2 gas, ketone gas, humidity, temperature, and VOC. PCA reduces data dimensions by creating new variables from linear combinations of the original features, preserving key data characteristics. This helps to identify and remove the less influential features, enhancing class separation [13].

A novel method for feature extraction was proposed by El-Magd et al. [33]. They used pretrained CNNs such as ResNet 18, ResNet 34, Resnet 50, AlexNet, and GoogleNet. The use of pretrained networks enables parameter extraction and the classification of even small datasets.

A gated recurrent unit-based autoencoder (GRU-AE) was used as a feature extraction method by Lu et al. in a proposed GRU-AE-MSEP framework. A GRU-AE was used to create detailed feature representations for effective classification. The autoencoder consists of two main parts: an encoder and a decoder. The encoder compresses high-dimensional multichannel data into a compact representation. The decoder then reconstructs the original high-dimensional data from this compressed form, aiming to minimize the reconstruction error [39].

In addition to features extracted from sensor measurements, researchers also use patients' medical data to train algorithms. Kort et al. [9] trained models using clinical features, breath features, and a combination of these two sources of information in their study. Their results showed that the model trained only on clinical data showed the lowest efficiency (sensitivity 53.9%). The model trained only on breath data showed a significantly higher sensitivity value of 88.2%. However, combining the two sets increased the sensitivity to 94.7%. The most commonly used clinical parameters are listed in Table 7.2.

7.5 Feature selection

The next step in data processing is feature selection. It is an important step in the pipeline because if we extract n features from k sensors it gives us k*n features. More features do not necessarily mean a better algorithm, because this algorithm is likely to become prone to overfitting and may learn irrelevant information present

Table 7.2 Clinical parameters used as additional parameters for training models for breath data analysis.

Parameter	References
Age	[7–9]
Gender	[7,9]
Blood pressure	[7]
SPO_2 (oxygen level in blood)	[7]
Heart rate	[7]
BMI	[9,10]
Smoking status (current smoker, ex-smoker, nonsmoker) and smoking history	[8,9]

in the data and may also have difficulty generalizing relationships between data [23]. To prevent these problems, techniques such as dimensionality reduction (e.g., PCA), feature selection, regularization (e.g., L1 and L2), or removing irrelevant features are used.

One of the most commonly used feature selection methods is dimensionality reduction [40,41] of the dataset with methods such as PCA and Linear Discriminant Analysis (LDA). PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are dimensionality reduction techniques used in data analysis, but they have different goals.

- PCA looks for directions in which the data have the highest variance, regardless of class labels. Its goal is to reduce the dimension of the data while preserving as much information as possible [40,41].
- LDA focuses on maximizing separation between classes. In addition to dimensionality reduction, LDA takes into account labels and tries to find directions that best separate the data into different classes [40,42].

In addition, ICA (independent component analysis) and KPCA (Kernel Principal Component Analysis) are used in sensor data analysis.

- ICA looks for independent components in the data, assuming that the data are a combination of sources that are statistically independent. It is often used in signals, for example, to separate sound or image sources [40,43].
- KPCA is a nonlinear version of PCA that uses the so-called kernel trick to find principal components in higher-dimensional spaces. This allows it to better handle nonlinear dependencies in the data [40,44,45].

In one study, Binson et al. used PCA to reduce data from five sensors into two components in the problem of detecting lung cancer from exhaled air [31]. In another study [26], in which samples from 218 people were analyzed to detect pulmonary diseases (lung cancer, chronic obstructive pulmonary disease (COPD), and asthma), Binson et al. compared the effectiveness of the abovementioned dimensionality reduction methods combined with machine learning algorithms. Their results showed that KPCA was the most effective feature selection technique in the task of classifying pulmonary disease.

Kapur et al. performed feature selection using the Recursive Feature Elimination (RFE) method which involves iteratively removing the least important features to find the optimal feature subset for the model. The process was performed iteratively until a satisfactory accuracy value was achieved [7].

Polaka et al. used a Greedy Search algorithm that performs a stepwise search to find the best feature subset. In their study, forward selection was used, starting with an empty set and adding features until further additions reduced performance [2].

Feature selection methods such as Lasso, Group Lasso, and their modification as Sparse Group Lasso (SGL) were compared by Liu et al. [23]. Their research showed that without using feature selection methods, the accuracy levels of classical machine learning models such as SVM, KNN, LogitBoost, and NB were several percent lower than with SGL. Models trained with SGL also showed higher efficiency than models trained on features selected with Lasso and Group Lasso methods. The authors noted that this method, due to feature identification, also enables identification of the most important sensors [23].

7.6 AI models for exhaled breath analysis for medical purposes

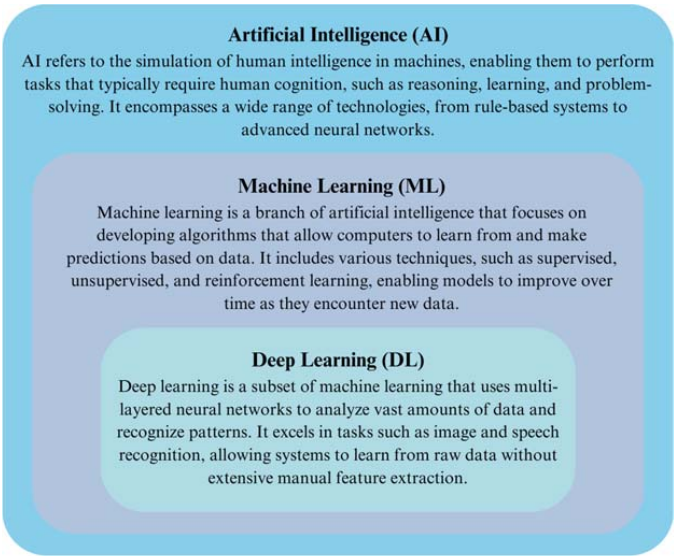
Artificial intelligence (AI) is a very broad concept; it is a field of computer science dealing with the creation and development of systems and algorithms capable of performing tasks that normally require human intelligence, such as the recognition of images, natural language processing, or making decisions based on data.

7.6.1 Types of the AI algorithms

Artificial intelligence algorithms can be divided on the basis of their complexity into machine learning algorithms (ML) and deep learning algorithms (DL). AI types and the characteristics of each field are shown in Fig. 7.5.

In addition to dividing AI algorithms by complexity, we can also divide algorithms by the task they are assigned and the data they provide [46]. There are four main types (Fig. 7.6):

- **Supervised learning**—the algorithm receives not only a set of input data, but also output data, and on this basis, it learns features and rules that are characteristic of a given class. This is the most common case of machine learning. This approach is widely used for tasks like classification (e.g., recognizing if a patient has a given disease) and regression (e.g., predicting health parameters such as blood glucose level).



Artificial Intelligence (AI)

AI refers to the simulation of human intelligence in machines, enabling them to perform tasks that typically require human cognition, such as reasoning, learning, and problem-solving. It encompasses a wide range of technologies, from rule-based systems to advanced neural networks.

Machine Learning (ML)

Machine learning is a branch of artificial intelligence that focuses on developing algorithms that allow computers to learn from and make predictions based on data. It includes various techniques, such as supervised, unsupervised, and reinforcement learning, enabling models to improve over time as they encounter new data.

Deep Learning (DL)

Deep learning is a subset of machine learning that uses multi-layered neural networks to analyze vast amounts of data and recognize patterns. It excels in tasks such as image and speech recognition, allowing systems to learn from raw data without extensive manual feature extraction.

FIGURE 7.5

Different levels of complexity in AI algorithms.

- **Unsupervised learning**—involves processing input data and learning rules defining them without the help of target output data. These are most often clustering algorithms.
- **Semisupervised learning**—uses a dataset that contains a small amount of labeled data and a larger amount of unlabeled data. The model uses both types of data to learn, which can improve accuracy and generalization compared with unsupervised learning alone, especially when labeling is expensive or time-consuming.
- **Reinforcement learning**—the system does not receive input or output data but only certain information about the environment (a "reinforcement signal" that may be positive (reward) or negative (punishment)) in which it is located and learns to take actions and make decisions by striving to maximize the reward.

In the case of breath analysis, the best known are classic supervised machine learning algorithms (Fig. 7.7), supervised deep learning algorithms (Fig. 7.8), and unsupervised learning (Fig. 7.9).

The operating principle of the algorithms from each of the above groups, their effectiveness, and application in the analysis of exhaled air for medical purposes are discussed below.

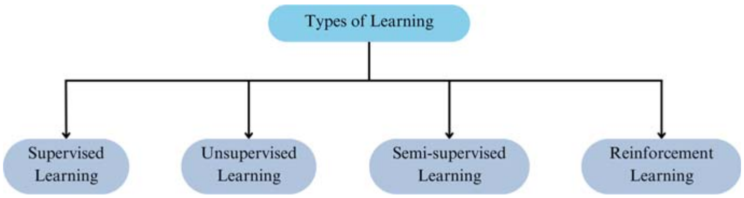


FIGURE 7.6

Different levels of complexity in AI algorithms.

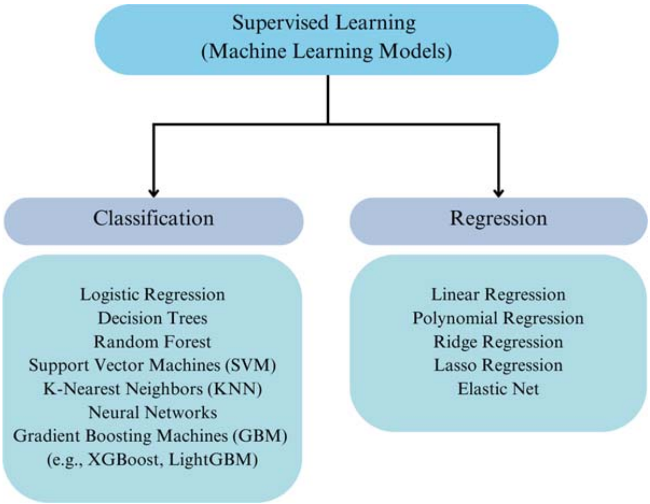


FIGURE 7.7

Supervised machine learning models.

7.6.2 Performance evaluation of AI algorithms

To evaluate the effectiveness of algorithms, we will use different metrics depending on whether the problem is classification or regression. The most used metrics are listed in Table 7.3.

Abbreviations are as follows:

- TP—True Positives
- TN—True Negatives
- FP—False Positives
- FN—False Negatives

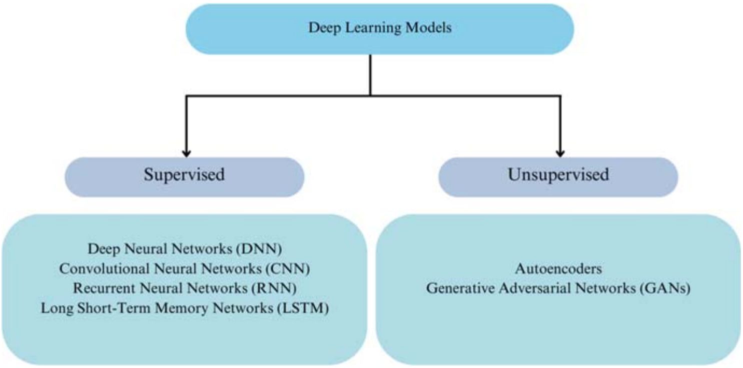

Deep Learning Models
Supervised
Unsupervised
Deep Neural Networks (DNN)
Convolutional Neural Networks (CNN)
Recurrent Neural Networks (RNN)
Long Short-Term Memory Networks (LSTM)
Autoencoders
Generative Adversarial Networks (GANs)

FIGURE 7.8

Deep learning models.

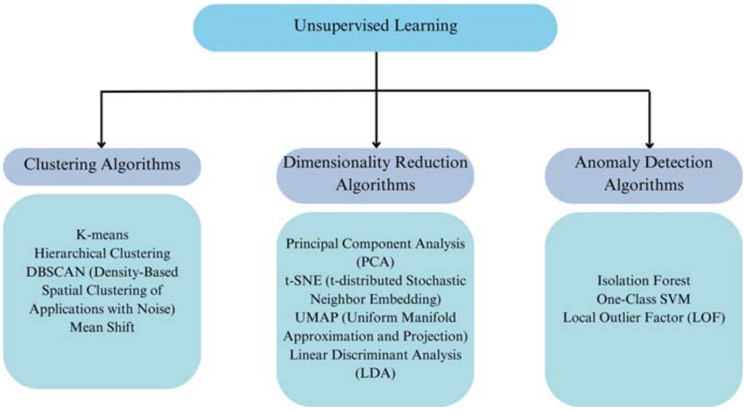

Unsupervised Learning
Clustering Algorithms
Dimensionality Reduction Algorithms
Anomaly Detection Algorithms
K-means
Hierarchical Clustering
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
Mean Shift
Principal Component Analysis (PCA)
t-SNE (t-distributed Stochastic Neighbor Embedding)
UMAP (Uniform Manifold Approximation and Projection)
Linear Discriminant Analysis (LDA)
Isolation Forest
One-Class SVM
Local Outlier Factor (LOF)

FIGURE 7.9

Examples of unsupervised algorithms.

- Actual value of the target variable
- Predicted value of the target variable
- n—Number of data points

In addition to classical metrics calculated using formulas, graphical representations of the performance of algorithms are also used, such as the confusion matrix (Fig. 7.10a) or the ROC curve (Fig. 7.10b).

Table 7.3 Metrics for classification and regression tasks [1,33,46,51,57,84,85].

Metric	Equation
Accuracy	$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$
Precision (positive predictive value)	$\text{Precision} = \frac{TP}{(TP+FP)}$
Recall (sensitivity or true positive rate)	$\text{Recall} = \frac{TP}{(TP+FN)}$
F1 score	$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Specificity (true negative rate)	$\text{Specificity} = \frac{TN}{(TN+FP)}$
False positive rate (FPR)	$\text{FPR} = \frac{FP}{FP+TN}$
False negative rate (FNR)	$\text{FNR} = \frac{FN}{FN+TP}$
Balanced accuracy	$\text{Balanced Accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2}$
Area under the ROC curve (AUC-ROC)	Illustrated in Fig. 7.10b
Matthews correlation coefficient (MCC)	$MCC = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Mean absolute error (MAE)	$\text{MAE} = \frac{1}{n\sum_{(i=1)}^{n} \lvert y_i - (y_i)^{\square} \rvert}$
Mean squared error (MSE)	$\text{MSE} = \frac{1}{n} \sum_{(i=1)}^{n} \left(y_i - (y_i)^{\square}\right)^2$
Root mean squared error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{n\sum_{(i=1)}^{n} (y_i - (y_i)^{\square})^2}}$
Mean absolute percentage error (MAPE)	$\text{MAPE} = \frac{1}{n\sum_{(i=1)}^{n} \left\lvert \frac{y_i - (y_i)^{\square}}{y_i} \right\rvert} \times 100$
R-squared (coefficient of determination)	$R^2 = 1 - \frac{\sum_{(i=1)}^{n} \left(y_i - (y_i)^{\square}\right)^2}{\sum_{(i=1)}^{n} \left(y_i - y^{\square}\right)^2}$
Adjusted R-squared	$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$

7.6.3 Dealing with problems in algorithms training

The goal of training a machine learning algorithm is to teach it to recognize patterns in training data and then generalize this ability to unseen examples. Model capacity refers to its ability to fit complex functions and relationships in data. A model is considered optimal when its capacity matches the complexity of the problem [47]. Two main issues during training are [46]:

- Underfitting occurs when the model fails to minimize the error sufficiently during training, often due to having too small a network size or capacity.
- Overfitting happens when there is a large discrepancy between the training and test error (the model memorizes the training data and cannot generalize to new data). This is typically due to too few training examples or too complex a model.

Several solutions can help prevent these issues:

- Adjusting the hyperparameters—Increasing the model size can enhance capacity and prevent underfitting but can also lead to overfitting if too large [7,34].

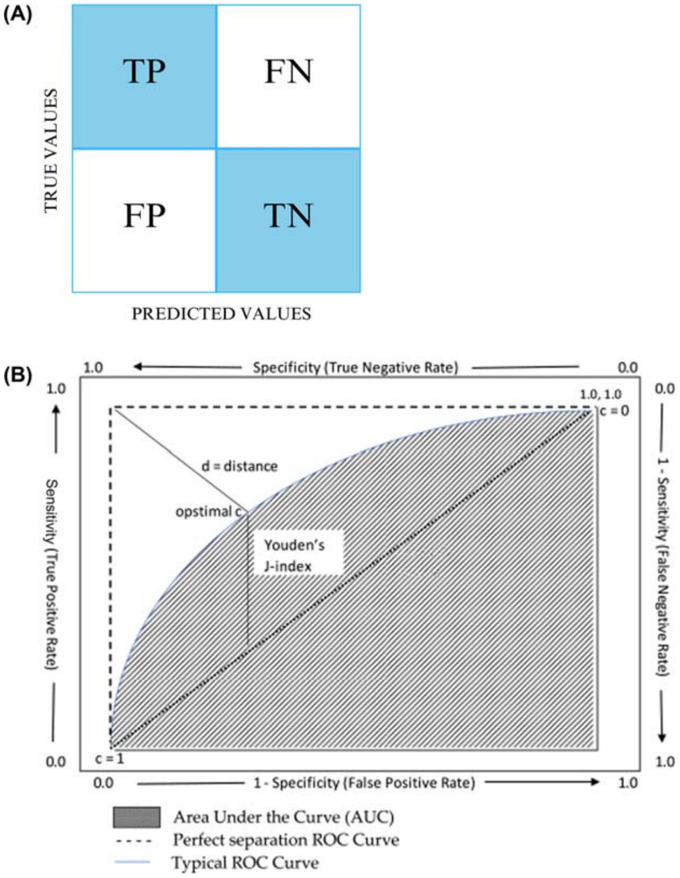

(A)
TRUE VALUES
TP
FN
FP
TN
PREDICTED VALUES
(B)
1.0
Specificity (True Negative Rate)
0.0
1.0
Sensitivity (True Positive Rate)
0.0
1.0, 1.0
c = 0
d = distance
opstimal c
Youden's J-index
c = 1
0.0
1 - Specificity (False Positive Rate)
1.0
0.0
1 - Sensitivity (False Negative Rate)
1.0
Area Under the Curve (AUC)
Perfect separation ROC Curve
Typical ROC Curve

FIGURE 7.10

Graphical representations of the performance of AI algorithms. (a) Confusion matrix, (b) ROC curve.

Reproduced from Ref. [83] with permissions.

- Increasing the number of training examples—More data helps prevent overfitting by providing varied examples for the model to learn from. If more data is not available, data augmentation techniques can be used [48,49].
- Weight regularization—This technique reduces model complexity by restricting the weights to small values, promoting a more regular weight distribution. It only applies during training, improving model performance on test data [47].

- Cross-validation—This technique involves splitting the data into multiple folds and training the model on different combinations of these folds, testing it on the remaining fold each time. Cross-validation helps ensure that the model generalizes well by evaluating its performance across different subsets of data and reducing the risk of overfitting [50].

7.6.4 The most commonly used algorithms in the analysis of human breath samples for disease diagnosis

7.6.4.1 Support Vector Machines

Support Vector Machines is one of the most popular algorithms used for the classification of exhaled air for disease detection. The operation of this algorithm is that it constructs a hyperplane or set of hyperplanes that maximize the separation of features between classes. In addition to classification, this algorithm can be used for Support Vector Regression (SVR) tasks, in which it tries to find a function that best reflects the fit (within a tolerance margin) to the data [26,51,52].

Liu et al. compared the performance of different machine learning algorithms (LogitBoost, KNN, Naïve Bayes) to classify breath samples taken from 46 patients with lung cancer, 36 healthy volunteers, and five patients with benign pulmonary diseases. Data were collected from 19 sensors included in the designed e-nose. The results showed that SVM with the adopted Gaussian radial basis kernel had the highest accuracy, especially when data were preprocessed with the SGL method [23].

High efficiency of SVM algorithm for the classification of respiratory diseases was also presented by Mahdavi et al. To develop the algorithm for COPD detection, they used breath samples collected from 34 healthy individuals and 33 chronic obstructive pulmonary disease (COPD) patients. The algorithm combined with a modern feature selection method achieved 80.60% accuracy, 78.79% sensitivity, and 82.35% specificity [50].

A similar study was conducted by Binson et al. They examined breath samples from 32 lung cancer patients, 38 COPD patients, and 72 healthy controls including smokers and nonsmokers, which were used to create separate algorithms for COPD and lung cancer detection. Interestingly, for lung cancer detection, KNN was the better algorithm, achieving 91.3% accuracy, 84.4% sensitivity, and 94.4% specificity, while for COPD detection, SVM performed best, achieving 90.9% accuracy, 81.6% sensitivity, and 95.8% specificity [53].

Tirzīte et al. used the commercial e-nose Cyranose 320 [54] to analyze the breath of patients with histologically or cytologically verified lung cancer, healthy volunteers, and patients with other lung diseases (e.g., chronic obstructive pulmonary disease (COPD), asthma, pneumonia, pulmonary embolism, benign lung tumors). In this case, SVM also achieved very high efficiency in detecting lung cancer (98.8%) and in classifying healthy versus other lung disease (87.3%) [8].

7.6.4.2 K-Nearest Neighbors

K-Nearest Neighbors is a machine-learning algorithm based on the nearest neighbor's method. It is an unsupervised or supervised algorithm (depending on the application) that classifies data or predicts a value based on the distance to other data in the set. Data are represented as points in a multidimensional space, and the distances between points are calculated using various distance metrics such as Euclidean [55], Manhattan, and Cosine. Then, *k* neighbors closest to the new point are selected, and depending on whether it is classification or regression, the point is assigned to the class that is most frequently represented among *k* neighbors (majority voting) or the result is the mean value or median value of *k* neighbors. This algorithm is simple to implement and does not require training, but it is sensitive to noise and outliers.

Martin et al. used the k-NN algorithm to classify synthetic mixtures (lung cancer biomarkers) added to breaths collected from healthy individuals. Their study showed that with increasing concentrations of added biomarkers, the algorithm achieved higher efficiency. The authors chose k-NN due to its simplicity and effectiveness as a nonparametric method, suitable for non-normal and/or heteroscedastic datasets. They pointed out that k-NN is a good choice for small datasets, which are most often found in medical data, especially breath analysis [1].

Guo et al. designed an e-nose system to detect several diseases such as diabetes, airway inflammation, and renal disease. For renal disease, they also collected patients' breaths before and after hemodialysis. The authors used PCA algorithm to create features and then classified data using k-NN. They achieved high accuracy in each case—over 80% accuracy in the classification of renal failure samples before and after treatment, the sensitivity of diabetes detection was 87.67% and the specificity was 86.87%, for renal failure detection, metrics were 86.57% and 83.47%, respectively, and for airway inflammation, the algorithm achieved sensitivity of 70.20% and specificity of 75.07%. These results show that the k-NN algorithm combined with PCA can be used to detect various diseases in exhaled air as well as to evaluate the effectiveness of dialysis [12].

k-Nearest Neighbors algorithm was also used by Smieja et al. to detect diabetes in exhaled air. They developed a custom, portable e-nose system with 6 VOC sensors. The system achieved an accuracy of 0.936, a precision of 1, and a recall of 0.875 in classifying diabetes and healthy samples taken from 28 individuals [55].

7.6.4.3 Tree models

Decision trees are one of the most popular machine learning algorithms that divide data into smaller subsets based on features, creating a tree-like structure. Each node corresponds to a division based on a given feature, and the leaves represent decisions or predictions. The algorithm strives to minimize entropy or Gini Impurity. Decision trees are easy to interpret because it is possible to prepare a graphical representation of them and see how the algorithm determines the boundaries of the division into classes. In the case of analysis of sensor data from breath measurements, it is possible to analyze what value measured by a given sensor (which may correspond to the concentration of a biomarker) is the decision boundary between the patient's health state. However, decision trees are susceptible to overfitting and noise.

An extension of classic decision trees is random forest models. This is a composition of many decision trees. The decisions/predictions of individual trees in a random forest are components of the final decision of the model—average (for regression) or majority voting (for classification). Compared with classical decision trees, RF is less susceptible to overfitting but is more difficult to visualize and interpret as well as being slower and more computationally complex [56].

Yang et al. compared different machine learning models in breath cancer prediction based on breath data collected using Cyranose 320. They used algorithms such as SVM with different kernels, kNN, Naive Bayes, Decision Tree, and Random Forest. In their results, the tree algorithms significantly outperformed the other algorithms. Both DT and RF showed similar results of accuracy at 91% and Random Forest showed 1% better AUC [57].

7.6.4.4 Gradient boosting models

An extension of classic machine learning methods is gradient boosting models, in which base models such as decision trees are built iteratively, and each subsequent model learns from the errors of the previous models. During training, the loss function is minimized using the gradient of errors between the prediction and the true values [26,27,58,59]. There are several varieties of models based on gradient boosting. Gradient-boosted algorithms are becoming increasingly popular in breath analysis [26,27,30,59]. The choice of the appropriate algorithm depends on the problem and should be made taking into account the advantages and disadvantages of each of them as well as their suitability for the problem.

- GBoost (Gradient Boosting) is a classic version of gradient boosting [60,61].
- XGBoost (eXtreme Gradient Boosting) is an improved version of classic GBoost, which could use L1 and L2 regularization, which allows for the reduction in overfitting and also optimizes calculations by parallelizing them. It is faster than the classic version [58].
- AdaBoost (Adaptive Boosting) uses stumps as base models, which have weights depending on their accuracy. It is more accurate in difficult cases because it increases their weights in subsequent iterations, but it is more susceptible to noise and overfitting [62].
- CatBoost is a model specially optimized for working with categorical data and does not require their initial preprocessing, it also copes well with missing data [63–66].
- LightGBM (Light Gradient Boosting Machine) is a modification created for the efficient (in time and memory) processing of very large data sets. It uses a leaf-wise growth strategy, which builds deeper trees in places of greater heterogeneity, which increases accuracy but also the risk of overfitting [67].

Glucobreath, a device designed by Kapur et al. [7] was used to collect the breath of both diabetes sufferers and healthy controls. The authors decided to compare different machine learning algorithms to detect diabetes in exhaled air. The XGBoost and GBoost algorithms achieved the highest accuracy with levels of 0.845 and 0.846, respectively, while the combination of these two algorithms achieved 0.969 accuracy in diabetes prediction. The authors reported that these algorithms performed better because they ensemble learning that combines multiple weaker learners to create a stronger model. The additional advantage is their ability to identify the most important features, prevent overfitting, and improve generalization.

Binson et al. compared the performance of Random Forest, AdaBoost, and XGBoost in the classification of breath samples taken from 199 participants (healthy, COPD, and lung cancer). The samples were measured using a custom e-nose system designed by the authors. XGBoost outperformed other algorithms achieving accuracy levels of 79.31% and 76.67% in predicting lung cancer and COPD, respectively [68].

Paleczek et al. used the LightGBM model to predict total cholesterol level based on patient breath. The performance of the model was compared with other most common machine learning regression models and achieved across the entire measurement range and for the norm range ≤200 mg/dL achieving MAPE 13.7% and 8%, respectively [10].

Ye et al. trained Gradient Boosted Trees (GBT), SVM, and RF to classify three ranges of blood glucose levels based on exhaled air. The best accuracy was achieved by GBT at 90.4%. The authors also trained a regressor version of each model to estimate the exact BLG. In this case, the GBT Regressor also outperformed other algorithms and achieved R^2 0.873 and mean average error 0.77 mmol/L [22].

7.6.4.5 Artificial Neural Networks (ANN)

Artificial Neural Networks are a machine learning model that was created drawing on inspiration from biological neural networks. It consists of an input layer, hidden layers, and an output layer that generates results. Neurons in these layers are connected by weights that determine the signal strength between them. During training, the network iteratively modifies its weights using the backpropagation algorithm to minimize the loss function.

Waltman et al. trained ANN to detect prostate cancer in exhaled air profiles measured by the commercially available electronic nose device, Aeonose (The eNose Company). The model achieved accuracy of 75% in detecting prostate cancer [69].

An ANN model was also used by Ooko et al. in their custom-designed device for respiratory disease detection. The trained model had four layers—an input layer, two dense layers (20 and 10 neurons) and an output layer. The model predicts respiratory diseases with an accuracy of 95.4%.

Accuracy of ANN and gradient-boosted decision trees (GBDT) in lung cancer prediction was compared by Temerdashev et al. They analyzed breath samples collected from 112 lung cancer patients and 120 healthy individuals using gas chromatography-mass spectrometry (GC-MS). ANN model achieved higher performance than GBDT (82%−88% sensitivity and 80%−86% specificity on test data) [70].

7.6.4.6 Convolutional Neural Networks (CNN)

The convolutional neural network algorithm contains convolutional layers between the input and output layers that perform signal convolution with the chosen kernel. Weights are values in the filter matrix that determine how the filter interacts with a piece of input data. They are updated during network training using a backpropagation algorithm to minimize network error [47].

Lee et al. analyzed 181 clinical breath samples (from 74 healthy controls and 107 lung cancer patients) using an e-nose system that contained 10 semiconductor metal oxide (SMO), one photoionization detector (PID), and nine electrochemical (EC) gas sensors. Input data for the 1D-CNN model included the normalized response as a 19 × 2400 matrix. The authors compared the 1D-CNN model with Multilayer Perceptron (MLP) and recurrent neural network (RNN) models. The best accuracy of 92% was achieved by CNN model, while MLP and RNN achieved 85% and 83% accuracy, respectively [71].

Aulia et al. used an e-nose with 20 semiconductor gas sensors to analyze breath samples collected from 30 healthy people and 40 COPD subjects. The obtained data were processed using five AI algorithms: RF, ANN, CNN, gated recurrent unit (GRU), and graph convolutional network GCN. The adjacency matrix was computed using the Pearson correlation coefficient (PCC) to construct the GCN feature map. To enhance distinguishing accuracy, frequency components of the sensor response and PCA are utilized during data preprocessing. The GCN processes the input graph of sensor features to produce classification results. In this graph, each sensor's time and frequency components serve as nodes and the edges connect each node only to its neighboring nodes. Consequently, the dataset forms an undirected graph. The GCN model using the frequency dataset achieves a maximum accuracy of 94.8%. When combined with PCA for data preprocessing, the GCN model delivers improved performance, achieving an accuracy of 97.5%, a precision of 97.2%, a recall of 97.4%, and an F1-score of 97.5% [72].

CNN models were also used by El-Magd et al. [33] to predict COPD using data collected from an e-nose system. They used pretrained CNN: ResNet 18, ResNet 34, ResNet 50, AlexNet, and GoogleNet. Using a pretrained model and transfer learning improves classification results on small data sets. The classification models had a flattening layer or a global average layer (GAP) as the first layer, and the second is a linear layer. All five CNNs based on pretrained models achieved more than 93% accuracy in the test set, and furthermore, Resnet 50 and GoogleNet achieved 100% classification accuracy.

Another modification of CNN models, correlational neural network (CORNN), was used by Bhaskar et al. to predict diabetes based on the analysis of exhaled breath samples [73]. In CORNN architecture, correlation layers are used instead of convolutional layers. The authors compared different models and classifiers. The best results were achieved by CORNN with an MLP classifier (accuracy 97.37%) and CORNN with an SVM classifier (accuracy 98.02%). The CORNN model with both classifiers performed better than the classic CNN model with the same classifiers.

7.7 Interpretability

AI algorithms are most often used as black boxes, where we provide input data and expect a result at the output [74], which in the case of medical data is a diagnosis of a disease or a prediction of some health parameter. Making a diagnosis and making therapeutic decisions blindly is risky and not reliable. Therefore, not only in medicine but also in other fields where AI algorithms are used, is it necessary to use explainable artificial intelligence. Analysis of the impact of features on the decision of the model allows not only the identification of model errors but also the potential to find/discover new factors that affect the development of the disease [75]. The ability to explain the model's decision increases the trust of doctors and patients in technology.

In the case of disease detection based on breath, it is possible to identify the sensor that is most important for the model to make a decision and thus select new biomarkers for a given disease state [75] or limit the number of sensors in the e-nose, which leads to its minimization, reduction in energy consumption, reduction in production costs and, therefore, enables an increase in the availability of the diagnostic device [76].

There are many different methods of explaining machine learning models. It is possible to explain models globally, in other words to analyze the model's operation as a whole, as well as locally, where the model's decision is explained for individual observations [77,78].

One of the simplest methods used for model explanations is model decomposition. Such methods are used in decision trees in which we can trace the exact path of the model's decision, in linear models, regression coefficients determine the impact of features on the model's decision, and in models using attention maps, it is possible to analyze these coefficients [74].

The second group is model agnostic methods. They can be used even in complex models. The most popular of these methods are as follows:

- LIME (Local Interpretable Model-Agnostic Explanations)—in this technique, local linear models are created that explain the decision of complex models at specific points [74,77,78].
- SHAP (Shapley Additive Explanations) comes from Shapley's game theory, in which it is calculated how to fairly divide the profit between several players. In XAI, players are individual features. This method allows the analysis of the significance of features on the entire set as well as on a single example [79].
- Partial Dependence Plots (PDP) consist of calculating the influence of one (or more) features on the model decision, assuming that the others are constant [79].

Another group of model interpretability methods collects methods that rely on data perturbations, that is, their change to determine their influence on the model decision. Such methods include feature importance analysis or counterfactual explanations [79].

Kapur et al. used the SHAP method to analyze the influence of features on the models' decisions. Plot highlighted that key physiological features, such as age, blood pressure (BP), heartbeat, SPO2, and most FFT features, play a significant role in the classification process for diabetes detection. In addition, for sensor voltage, the primary contributors to diabetes prediction were TGS826, TGS2603, TGS2610, and TGS2620. These sensors are particularly sensitive to VOCs such as ammonia, LP gas, propane, butane, alcohol, organic solvent vapors, amine series, and sulfurous gases [7].

Paleczek et al. used feature importance analysis to identify key sensors for total cholesterol level prediction using the LightGBM regressor. The analysis of key features showed that the most significant predictors were the responses of the TGS1820, AL-03P, TGS2620, and MQ3 sensors, which primarily detect acetone, ethanol, and other volatile organic compounds (VOCs) [10].

7.8 AI/ML deployment in e-nose systems

Most of the research on breath analysis and disease detection using e-nose and AI algorithms is conducted offline, that is, breath samples are measured, data are recorded and then analyzed and used to train machine learning algorithms and disease diagnosis. However, a few of these systems are feasible for field use and online diagnosis.

Kapur et al. designed the Glucobreath [7] device, which combined the VOC-Analyzer microcontroller's WiFi-enabled wireless communication with external entities, such as the InfluxDB cloud database, via the MQTT protocol. This protocol allowed time-series data to be streamed to the cloud-based InfluxDB server, and the data were displayed using Grafana. The authors also prepared a WebUI for patient data entry and model prediction display.

Ooko et al. designed an e-nose device based on TinyML technology. TinyML is about preparing machine learning models on devices with limited computational power, memory, and energy resources, such as microcontrollers and IoT devices. TinyML allows data to be processed locally on the device, which reduces the need to send data to the cloud and improves the performance and the energy efficiency of the system, enables real-time diagnosis and increases the portability of the device. The authors deployed the model on Arduino Nano 33 BLE sense [80].

The e-nose system developed for the Raspberry Pi using RPi IDE and Python was proposed by Evangelista et al. An Android app, created on the Basic4Android platform, displayed glucose predictions and supports data storage, deletion, and analysis for diabetes classification. Data were collected through breath samples processed by Raspberry Pi and analyzed using machine learning techniques like CNN and SVM. Results are validated against standard diabetes tests and repeated across participant groups to ensure model reliability [81].

Another system was proposed by Tiele et al. They designed a simple mobile app and developed using Blynk that communicates with the device via Wi-Fi and supports control through USB or Wi-Fi. The app, compatible with IoT hardware like

Arduino and ESP32, enables features sampling and analysis for real-time data monitoring and graphing. The ESP32 microcontroller was chosen for its low-cost, low-power capabilities and integrated Wi-Fi and Bluetooth. Real-time readings from sensors like SCD30 and CCS811 are displayed and plotted [82].

7.9 Conclusions

In the process of creating solutions using artificial intelligence algorithms, extremely important steps are preprocessing the data, creating new features, training various algorithms, and then evaluating performance and deployment. Training most AI models requires having a large data set, which is often difficult or impossible in the case of medical data, which means that creating such diagnostic devices with AI is limited. Working with data from measuring breath samples, researchers must deal with baseline drift, noise, environmental pollution, and the impact of the temperature and high humidity of human breath. Proper data preparation requires knowledge of signal processing methods. Many different AI algorithms are used in the analysis of breath; often, these are supervised algorithms for disease classification. This chapter shows that classic machine learning algorithms are most often used, which achieve high levels of accuracy. An important element of data processing is also the analysis of the impact of features on the decisions of the model. Currently, many advanced methods are known that enable the interpretation of decisions of practically every AI model, but they are not popular in the analysis of breath samples. In addition to training an effective algorithm, the challenge is its deployment on portable devices. The algorithm's computational efficiency and complexity, energy consumption, and implementation of appropriate communication as well as processing time are important. In the case of breath analysis, a standard for measurement or analysis of samples has not yet been developed, which makes experiments difficult to reproduce. In addition, despite the large number of recorded sensor responses, it is impossible to compare or use them between two different devices or systems because their responses strongly depend on the mechanical conditions of the measurement chamber, such as flow or volume, as well as the measurement electronics. These problems significantly limit access to data that is an essential element of AI algorithms and the sharing of it between research groups.

References

[1] J.D.M. Martin, F. Claudia, A.C. Romain, How well does your e-nose detect cancer? Application of artificial breath analysis for performance assessment, J. Breath Res. 18 (2) (2024) 026002, https://doi.org/10.1088/1752-7163/AD1D64.

[2] I. Polaka, et al., The detection of colorectal cancer through machine learning-based breath sensor analysis, Diagnostics 13 (21) (2023) 3355, https://doi.org/10.3390/DIAGNOSTICS13213355/S1.

[3] J. Beauchamp, J. Herbig, R. Gutmann, A. Hansel, On the use of Tedlar® bags for breath-gas sampling and analysis, J. Breath Res. 2 (4) (2008) 046001, https://doi.org/10.1088/1752-7155/2/4/046001.

[4] L.J. McGarvey, C.V. Shorten, The effects of adsorption on the reusability of Tedlar® air sampling bags, AIHAJ - Am. Ind. Hygiene Assoc. 61 (3) (2000) 375–380, https://doi.org/10.1080/15298660008984546.

[5] P. Mochalski, J. King, K. Unterkofler, A. Amann, Stability of selected volatile breath constituents in Tedlar, Kynar and Flexfilm sampling bags, Analyst 138 (5) (2013) 1405–1418, https://doi.org/10.1039/C2AN36193K.

[6] F.J. Gilchrist, et al., The suitability of Tedlar bags for breath sampling in medical diagnostic research, Physiol. Meas. 28 (1) (2006) 73, https://doi.org/10.1088/0967-3334/28/1/007.

[7] R. Kapur, et al., GlucoBreath: an IoT, ML, and breath-based non-invasive glucose meter, IEEE Acc. 12 (2024) 59346–59360, https://doi.org/10.1109/ACCESS.2024.3392015.

[8] M. Tirzite, M. Bukovskis, G. Strazda, N. Jurka, I. Taivans, Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis, J. Breath Res. 11 (3) (2017) 036009, https://doi.org/10.1088/1752-7163/AA7799.

[9] S. Kort, et al., Diagnosing non-small cell lung cancer by exhaled breath profiling using an electronic nose: a multicenter validation study, Chest 163 (3) (2023) 697–706, https://doi.org/10.1016/J.CHEST.2022.09.042.

[10] A. Paleczek, et al., Noninvasive total cholesterol level measurement using an E-nose system and machine learning on exhaled breath samples, ACS Sens. (2024), https://doi.org/10.1021/ACSSENSORS.4C02198.

[11] T.D.C. Minh, D.R. Blake, P.R. Galassetti, The clinical potential of exhaled breath analysis for diabetes mellitus, Diabetes Res. Clin. Pract. 97 (2) (2012) 195–205, https://doi.org/10.1016/j.diabres.2012.02.006.

[12] K. Yan, D. Zhang, A novel breath analysis system for diabetes diagnosis, IEEE Comput. Soc. (2012), https://doi.org/10.1109/icch.2012.6724490.

[13] R. Sarno, S.I. Sabilla, D.R. Wijaya, Electronic nose for detecting multilevel diabetes using optimized deep neural network, Eng. Lett. 28 (1) (2020).

[14] D. Guo, D. Zhang, N. Li, L. Zhang, J. Yang, Diabetes Identification and Classification by Means of a Breath Analysis System, 2010, https://doi.org/10.1007/978-3-642-13923-9_6.

[15] I. Oakley-Girvan, S.W. Davis, Breath based volatile organic compounds in the detection of breast, lung, and colorectal cancers: a systematic review, Cancer Biomarkers 21 (1) (2018) 29–39, https://doi.org/10.3233/CBM-170177.

[16] K.D. McCarthy, Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors, Breast Cancer Res. Treat. 29 (8) (2002) 729.

[17] P. Španel, K. Dryahina, D. Smith, Acetone, ammonia and hydrogen cyanide in exhaled breath of several volunteers aged 4–83 years, J. Breath Res. 1 (1) (2007) 011001, https://doi.org/10.1088/1752-7155/1/1/011001.

[18] M. Sun, et al., Continuous monitoring of breath acetone, blood glucose and blood ketone in 20 type 1 diabetic outpatients over 30 days, J. Anal. Bioanal. Tech. 8 (5) (2017) 1–8, https://doi.org/10.4172/2155-9872.1000386.

[19] M. Sun, et al., Determination of breath acetone in 149 Type 2 diabetic patients using a ringdown breath-acetone analyzer, Anal. Bioanal. Chem. 407 (6) (2015) 1641–1650, https://doi.org/10.1007/S00216-014-8401-8.

[20] S. Boumali, M.T. Benhabiles, A. Bouziane, F. Kerrour, K. Aguir, Acetone discriminator and concentration estimator for diabetes monitoring in human breath, Semicond. Sci. Technol. 36 (8) (2021) 085010, https://doi.org/10.1088/1361-6641/AC0C63.

[21] K. Yan, D. Zhang, Blood glucose prediction by breath analysis system with feature selection and model fusion, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2014 (2014) 6406–6409, https://doi.org/10.1109/embc.2014.6945094.

[22] Z. Ye, J. Wang, H. Hua, X. Zhou, Q. Li, Precise detection and quantitative prediction of blood glucose level with an electronic nose system, IEEE Sens. J. 22 (13) (2022) 12452–12459, https://doi.org/10.1109/JSEN.2022.3178996.

[23] B. Liu, et al., Lung cancer detection via breath by electronic nose enhanced with a sparse group feature selection approach, Sensor. Actuator. B Chem. 339 (2021) 129896, https://doi.org/10.1016/J.SNB.2021.129896.

[24] I. Polaka, et al., Modular point-of-care breath analyzer and shape taxonomy-based machine learning for gastric cancer detection, Diagnostics 12 (2) (2022) 491, https://doi.org/10.3390/DIAGNOSTICS12020491/S1.

[25] M. Zuppa, C. Distante, K.C. Persaud, P. Siciliano, Recovery of drifting sensor responses by means of DWT analysis, Sensor. Actuator. B Chem. 120 (2) (2007) 411–416, https://doi.org/10.1016/J.SNB.2006.02.049.

[26] V.A. Binson, M. Subramoniam, Y. Sunny, L. Mathew, Prediction of pulmonary diseases with electronic nose using SVM and XGBoost, IEEE Sens. J. 21 (18) (2021) 20886–20895, https://doi.org/10.1109/JSEN.2021.3100390.

[27] A. Paleczek, D. Grochala, A. Rydosz, Artificial breath classification using XGBoost algorithm for diabetes detection, Sensors 21 (12) (2021) 12, https://doi.org/10.3390/s21124187.

[28] T. Wang, et al., Portable electronic nose system with elastic architecture and fault tolerance based on edge computing, ensemble learning, and sensor swarm, Sensor. Actuator. B Chem. 375 (2023) 132925, https://doi.org/10.1016/J.SNB.2022.132925.

[29] C. Avian, M.I. Mahali, N.A.S. Putro, S.W. Prakosa, J.S. Leu, Fx-net and PureNet: convolutional neural network architecture for discrimination of chronic obstructive pulmonary disease from smokers and healthy subjects through electronic nose signals, Comput. Biol. Med. 148 (2022) 105913, https://doi.org/10.1016/J.COMPBIOMED.2022.105913.

[30] L. Hao, G. Huang, An improved AdaBoost algorithm for identification of lung cancer based on electronic nose, Heliyon 9 (3) (2023) e13633, https://doi.org/10.1016/j.heliyon.2023.e13633.

[31] V.A. Binson, M. Subramoniam, L. Mathew, Prediction of lung cancer with a sensor array based e-nose system using machine learning methods, Microsyst. Technol. 30 (11) (2024) 1421–1434, https://doi.org/10.1007/S00542-024-05656-5/TABLES/4.

[32] O.Y. Rodionova, A.L. Pomerantsev, Detection of outliers in projection-based modeling, Anal. Chem. 92 (3) (2020) 2656–2664, https://doi.org/10.1021/ACS.ANALCHEM.9B04611/ASSET/IMAGES/LARGE/AC9B04611_0007.JPEG.

[33] L.M.A. El-Magd, G. Dahy, T.A. Farrag, A. Darwish, A.E. Hassnien, An interpretable deep learning based approach for chronic obstructive pulmonary disease using explainable artificial intelligence, Int. J. Inf. Technol. 1 (2024) 1–16, https://doi.org/10.1007/S41870-023-01713-W/TABLES/9.

[34] K.C. Chen, S.W. Kuo, R.H. Shie, H.Y. Yang, Advancing accuracy in breath testing for lung cancer: strategies for improving diagnostic precision in imbalanced data, Respir. Res. 25 (1) (2024) 1–10, https://doi.org/10.1186/S12931-024-02668-7/FIGURES/4.

[35] N. Dufour, et al., Increasing the sensitivity and selectivity of Metal Oxide gas sensors by controlling the sensitive layer polarization, Proc. IEEE Sens. (2012), https://doi.org/10.1109/ICSENS.2012.6411463.

[36] L. Ferrus, H. Guenard, G. Vardon, P. Varene, Respiratory water loss, Respir. Physiol. 39 (3) (1980) 367−381, https://doi.org/10.1016/0034-5687(80)90067-5.

[37] A.T. Güntner, I.C. Weber, S. Schon, S.E. Pratsinis, P.A. Gerber, Monitoring rapid metabolic changes in health and type-1 diabetes with breath acetone sensors, Sensor. Actuator. B Chem. 367 (2022) 132182, https://doi.org/10.1016/J.SNB.2022.132182.

[38] M. MacIel, S. Sankari, M. Woollam, M. Agarwal, Optimization of metal oxide nanosensors and development of a feature extraction algorithm to analyze VOC profiles in exhaled breath, IEEE Sens. J. 23 (15) (2023) 16571−16578, https://doi.org/10.1109/JSEN.2023.3288968.

[39] B. Lu, L. Fu, B. Nie, Z. Peng, H. Liu, A novel framework with high diagnostic sensitivity for lung cancer detection by electronic nose, Sensors 19 (23) (2019) 5333, https://doi.org/10.3390/S19235333.

[40] F. Anowar, S. Sadaoui, B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE), Comput. Sci. Rev. 40 (2021) 100378, https://doi.org/10.1016/J.COSREV.2021.100378.

[41] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433−459, https://doi.org/10.1002/WICS.101.

[42] P. Xanthopoulos, P.M. Pardalos, T.B. Trafalis, Linear Discriminant Analysis, 2013, pp. 27−33, https://doi.org/10.1007/978-1-4419-9878-1_4.

[43] C.J. James, C.W. Hesse, Independent component analysis for biomedical signals, Physiol. Meas. 26 (1) (2004) R15, https://doi.org/10.1088/0967-3334/26/1/R02.

[44] J.M. Lee, C.K. Yoo, S.W. Choi, P.A. Vanrolleghem, I.B. Lee, Nonlinear process monitoring using kernel principal component analysis, Chem. Eng. Sci. 59 (1) (2004) 223−234, https://doi.org/10.1016/J.CES.2003.09.012.

[45] T.J. Chin, D. Suter, Incremental kernel principal component analysis, IEEE Trans. Image Process. 16 (6) (2007) 1662−1674, https://doi.org/10.1109/TIP.2007.896668.

[46] A. Géron, Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems, pp. 547.

[47] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning.

[48] L. Liu, et al., Boost AI power: data augmentation strategies with unlabeled data and conformal prediction, a case in alternative herbal medicine discrimination with electronic nose, IEEE Sens. J. 21 (20) (2021) 22995−23005, https://doi.org/10.1109/JSEN.2021.3102488.

[49] J. Kim, et al., A novel pathway to construct gas concentration prediction model in real-world applications: data augmentation; fast prediction; and interpolation and extrapolation, Sensor. Actuator. B Chem. 382 (2023) 133533, https://doi.org/10.1016/J.SNB.2023.133533.

[50] H. Mahdavi, S. Rahbarpour, S.M. Hosseini-Golgoo, H. Jamaati, A single gas sensor assisted by machine learning algorithms for breath-based detection of COPD: a pilot study, Sens. Actuators A Phys. 376 (2024) 115650, https://doi.org/10.1016/J.SNA.2024.115650.

[51] K.C. Suresh, R. Prabha, N. Hemavathy, S. Sivarajeswari, D. Gokulakrishnan, M. Jagadeesh kumar, A machine learning approach for human breath diagnosis with

soft sensors, Comput. Electr. Eng. 100 (2022) 107945, https://doi.org/10.1016/J.COMPELECENG.2022.107945.

[52] R.S. Parte, A. Patil, A. Patil, A. Kad, S. Kharat, Non-Invasive Method for Diabetes Detection using CNN and SVM Classifier, International Journal of Scientific Research and Engineering Development 3, Available: www.ijsred.com. (Accessed 19 December 2021).

[53] V.A. Binson, M. Subramoniam, L. Mathew, Discrimination of COPD and lung cancer from controls through breath analysis using a self-developed e-nose, J. Breath Res. 15 (4) (2021) 046003, https://doi.org/10.1088/1752-7163/ac1326.

[54] Cyranose 320 Electronic Nose - Smart Smell Detection Sensors. https://www.sensigent.com/cyranose-320.html.

[55] P.J. Smieja, J. Lu, Q. Sun, X. Fu, T. Zhang, Exhaled breath analysis based diabetes detection with k-nearest neighbors classifier, in: Proceedings - 2023 16th International Symposium on Computational Intelligence and Design, ISCID, 2023, pp. 126—130, https://doi.org/10.1109/ISCID59865.2023.00037.

[56] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5—32, https://doi.org/10.1023/A:1010933404324/METRICS.

[57] H.Y. Yang, Y.C. Wang, H.Y. Peng, C.H. Huang, Breath biopsy of breast cancer using sensor array signals and machine learning analysis, Sci. Rep. 11 (1) (2021) 1—9, https://doi.org/10.1038/s41598-020-80570-0.

[58] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[59] A. Ogunleye, Q.G. Wang, Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease, in: IEEE International Conference on Control and Automation, 2018-June, ICCA, 2018, pp. 805—810, https://doi.org/10.1109/ICCA.2018.8444167.

[60] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artif. Intell. Rev. 54 (3) (2021) 1937—1967, https://doi.org/10.1007/S10462-020-09896-5/METRICS.

[61] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorob. 7 (DEC) (2013) 63623, https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX.

[62] R.E. Schapire, Explaining Adaboost, in: Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, 2013, pp. 37—52, https://doi.org/10.1007/978-3-642-41136-6_5/TABLES/2.

[63] W. Chang, X. Wang, J. Yang, T. Qin, An improved CatBoost-based classification model for ecological suitability of blueberries, Sensors 23 (4) (2023) 1811, https://doi.org/10.3390/S23041811/S1.

[64] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018). Available: https://github.com/catboost/catboost. (Accessed 14 March 2023).

[65] J.T. Hancock, T.M. Khoshgoftaar, CatBoost for big data: an interdisciplinary review, J. Big Data 7 (1) (2020) 1—45, https://doi.org/10.1186/S40537-020-00369-8/FIGURES/9.

[66] P.B. Dash, J. Nayak, C.R. Kishore, M. Mishra, B. Naik, Efficient ensemble learning based CatBoost approach for early-stage stroke risk prediction, Smart Innov. Syst. Technol. 317 (2023) 475—483, https://doi.org/10.1007/978-981-19-6068-0_46/COVER.

[67] Welcome to LightGBM's documentation! — LightGBM 4.0.0 documentation, Available: https://lightgbm.readthedocs.io/en/stable/. (Accessed 07 August 2024).

[68] V.A. Binson, M. Subramoniam, L. Mathew, Detection of COPD and Lung Cancer with electronic nose using ensemble learning methods, Clin. Chim. Acta 523 (2021) 231−238, https://doi.org/10.1016/J.CCA.2021.10.005.

[69] C.G. Waltman, T.A.T. Marcelissen, J.G.H. van Roermund, Exhaled-breath testing for prostate cancer based on volatile organic compound profiling using an electronic nose device (Aeonose^{TM}): a preliminary report, Eur. Urol. Focus 6 (6) (2020) 1220−1225, https://doi.org/10.1016/J.EUF.2018.11.006.

[70] A.Z. Temerdashev, E.M. Gashimova, V.A. Porkhanov, I.S. Polyakov, D.V. Perunov, E.V. Dmitrieva, Non-invasive lung cancer diagnostics through metabolites in exhaled breath: influence of the disease variability and comorbidities, Metabolites 13 (2) (2023) 203, https://doi.org/10.3390/METABO13020203.

[71] B. Lee, et al., Breath analysis system with convolutional neural network (CNN) for early detection of lung cancer, Sensor. Actuator. B Chem. 409 (2024) 135578, https://doi.org/10.1016/J.SNB.2024.135578.

[72] D. Aulia, R. Sarno, S.C. Hidayati, A.N. Rosyid, M. Rivai, Identification of chronic obstructive pulmonary disease using graph convolutional network in electronic nose, Indones. J. Electr. Eng. Comput. Sci. 34 (1) (2024) 264−275, https://doi.org/10.11591/ijeecs.v34.i1.pp264-275.

[73] N. Bhaskar, V. Bairagi, E. Boonchieng, M.V. Munot, Automated detection of diabetes from exhaled human breath using deep hybrid architecture, IEEE Acc. 11 (2023) 51712−51722, https://doi.org/10.1109/ACCESS.2023.3278278.

[74] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI methods - a brief overview, Lect. Notes Comput. Sci. (2022) 13−38, https://doi.org/10.1007/978-3-031-04083-2_2/FIGURES/3, 13200 LNAI.

[75] M. Wieczorek, A. Weston, M. Ledenko, J.N. Thomas, R. Carter, T. Patel, A deep learning approach for detecting liver cirrhosis from volatolomic analysis of exhaled breath, Front. Med. 9 (2022) 992703, https://doi.org/10.3389/FMED.2022.992703/BIBTEX.

[76] A. Paleczek, A. Rydosz, The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: laboratory studies, Sensor. Actuator. B Chem. 408 (2024) 135550, https://doi.org/10.1016/J.SNB.2024.135550.

[77] R. Dwivedi, et al., Explainable AI (XAI): core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023), https://doi.org/10.1145/3561048/ASSET/160EFD77-21FC-4899-B447-110C1806F0DB/ASSETS/GRAPHIC/CSUR-2021-0681-F21.JPG.

[78] S. Razavi, Deep learning, explained: fundamentals, explainability, and bridgeability to process-based modelling, Environ. Model. Software 144 (2021) 105159, https://doi.org/10.1016/J.ENVSOFT.2021.105159.

[79] M.Z. Naser, An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: navigating causality, forced goodness, and the false perception of inference, Autom. Constr. 129 (2021) 103821, https://doi.org/10.1016/J.AUTCON.2021.103821.

[80] S.O. Ooko, J. Nsenga, Tiny machine learning (TinyML) based self diagnostic kit for respiratory diseases, in: ICCA 2023 - 2023 5th International Conference on Computer and Applications, Proceedings, 2023, https://doi.org/10.1109/ICCA59364.2023.10401673.

[81] G.G. Evangelista, A.C. Paglinawan, C.C. Paglinawan, Design of breath acetone detection for diabetes detection and classification using electronic nose, in: 2024 16th International Conference on Computer and Automation Engineering, ICCAE, 2024, pp. 708−713, https://doi.org/10.1109/ICCAE59995.2024.10569837.

[82] A. Tiele, A. Wicaksono, S.K. Ayyala, J.A. Covington, Development of a compact, IoT-enabled electronic nose for breath analysis, Electronics 9 (1) (2020) 84, https://doi.org/10.3390/ELECTRONICS9010084.

[83] P. McCarthy, Predicting trips to health care facilities: a binary logit and receiver operating characteristics (ROC) approach, Res. Transport. Econ. 103 (2024) 101411, https://doi.org/10.1016/J.RETREC.2024.101411.

[84] B. Selvaraj, E. Rajasekar, J.B. Balaguru Rayappan, Machine learning approaches: detecting the disease variants in human-exhaled breath biomarkers, ACS Omega 9 (1) (2024) 215−226, https://doi.org/10.1021/ACSOMEGA.3C03755/ASSET/IMAGES/LARGE/AO3C03755_0014.JPEG.

[85] V. Plevris, G. Solorzano, N.P. Bakas, M.E.A. Ben Seghier, Investigation of Performance Metrics in Regression Analysis and Machine Learning-based Prediction Models 8, 2022, https://doi.org/10.23967/ECCOMAS.2022.155.

2.3. Review of the algorithms used in exhaled breath analysis for the detection of diabetes

IOP Publishing J. Breath Res. 16 (2022) 026003 https://doi.org/10.1088/1752-7163/ac4916

Journal of Breath Research

OPEN ACCESS

RECEIVED
17 November 2021

REVISED
23 December 2021

ACCEPTED FOR PUBLICATION
7 January 2022

PUBLISHED
26 January 2022

PAPER

Review of the algorithms used in exhaled breath analysis for the detection of diabetes

Anna Paleczek* and Artur Rydosz

Institute of Electronics, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland

* Author to whom any correspondence should be addressed.

E-mail: paleczek@agh.edu.pl

Keywords: exhaled breath analysis, diabetes, algorithms, neural network, PCA, KNN, SVM

Abstract

Currently, intensive work is underway on the development of truly noninvasive medical diagnostic systems, including respiratory analysers based on the detection of biomarkers of several diseases including diabetes. In terms of diabetes, acetone is considered as a one of the potential biomarker, although is not the single one. Therefore, the selective detection is crucial. Most often, the analysers of exhaled breath are based on the utilization of several commercially available gas sensors or on specially designed and manufactured gas sensors to obtain the highest selectivity and sensitivity to diabetes biomarkers present in the exhaled air. An important part of each system are the algorithms that are trained to detect diabetes based on data obtained from sensor matrices. The prepared review of the literature showed that there are many limitations in the development of the versatile breath analyser, such as high metabolic variability between patients, but the results obtained by researchers using the algorithms described in this paper are very promising and most of them achieve over 90% accuracy in the detection of diabetes in exhaled air. This paper summarizes the results using various measurement systems, feature extraction and feature selection methods as well as algorithms such as support vector machines, *k*-nearest neighbours and various variations of neural networks for the detection of diabetes in patient samples and simulated artificial breath samples.

1. Introduction

Noninvasive methods of disease detection are increasingly the subject of research; in particular, researchers focus on the analysis of exhaled air. The first reports on the use of the diagnostic potential of breath come from the times of Hippocrates, who diagnosed uncontrolled diabetes and liver disease based on the smell of acetone emitted from the patient's mouth [1].

The main components of the exhaled air are nitrogen (78%–79%), oxygen (13%–16%) and carbon dioxide (4%) [2]. The remaining part of the breath profile consists of volatile organic compounds (VOCs). The composition of inhaled and exhaled air is shown in figure 1.

Currently, more than 3000 VOCs are identified in breath samples [3, 4], but due to the low concentration of VOCs in the breath and the use of various measurement methods, the exact number of VOCs in a single breath is not well defined. For example, Phillips *et al* have reported 204 VOCs [4], Smolinska *et al* have reported 300–500 VOCs [3] and Barash *et al* have identified more than 500 VOCs in each breath sample [5]. Moreover, Barash *et al* have shown that only some of the VOCs are common in different breath samples [5]. Each person has a different composition of breathing, somehow it could be said that the exhaled breath profile is an identifier similarly to fingerprints. Although there is a lack of confirmation of such statement so far. The exhaled breath profile depends also on the diet and on diseases [6, 7], which makes the exhaled breath analysis even more challenging. In addition, the composition of the exhaled VOCs is demanding to be investigated because these compounds in the breath are present in very low concentrations of a few parts per million (ppm), parts per billion and parts per trillion [7, 8].

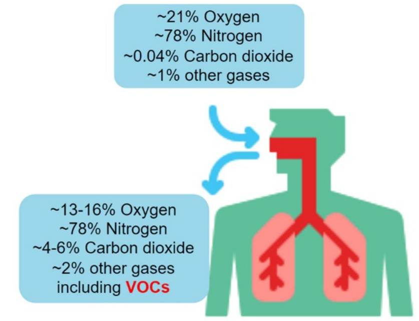


Figure 1. Composition of inhaled and exhaled air.

Table 1. Potential biomarkers of disease in breath. Reproduced from [40]. CC BY 4.0.

Disease	Biomarkers	References
Diabetes	Acetone	[6, 8, 29–33, 35–37, 41]
Diabetes	Isoprene	[42]
Asthma	Nitric oxide	[8, 15, 43]
Cystic fibrosis	Hydrogen cyanide	[44, 45]
Lung cancer	VOC pattern	[19, 20, 41]
Chronic kidney disease	Trimethylamine	[46]
Colorectal cancer	Methane	[47, 48]
Myocardial infarction	Pentane	[49, 50]
Obstructive sleep apnea	Pentane and Nitric oxide	[51]
Renal failure	Ammonia	[52, 53]

VOCs are divided into two categories: exogenous and endogenous, depending on their origin. The groups of chemical compounds are not separable; the presence of one chemical in the exhaled air can be of both exogenous and endogenous origin. Exogenous VOCs are emitted as a result of the influence of the external environment on the body, not only of those inhaled with air, such as air pollution and cigarette smoke, but also of medications and diet [9–11]. Endogenous VOCs, called biomarkers, are produced by cells in the body, e.g. in metabolic diseases, asthma and cancer [10–12]. The course of the disease and its advancement, treatment methods and medications also affect the composition of the exhaled air [13]. Biomarkers are used to diagnose diseases, predict the risk of developing a disease or monitor the course of treatment and assess the effectiveness of therapies used in diseases such as asthma [14–18], cancer [13, 19–26], chronic obstructive pulmonary disease [13, 24, 27, 28] and diabetes [6, 29–37]. Selected biomarkers of diseases in breath were shown in table 1. Due to the very low concentrations of VOC in exhaled air, researches have proposed the use of preconcentrators and micropreconentrators in systems designed for breath analysis [35, 38, 39] as an option that lead to increased limit of detection. However, the analysis becomes more complex, additional precautions and factors need to be taken into account.

The World Health Organization reports more than 500 million people with diabetes worldwide and more than 1.6 million diabetic deaths annually [54]. The number of patients is constantly increasing, and according to estimates in 1995 it was 7.4%, while the predicted value for 2025 is 9% [55].

The International Diabetes Federation (IDF) estimated 463 million adults diabetes in 2009, but in the latest report from 2021, this number increased by 16% and the current estimate is 537 million adults with diabetes. These numbers continue to rise and by 2045 it is projected to increase to 783 million adults living with diabetes. Almost half (44.7%) of diabetics is undiagnosed. Additionally, the IDF report shows that 10.6% of adults worldwide have high risk of developing type 2 diabetes due to impaired glucose tolerance. Early diagnosis plays an important role in the treatment of diabetes and in reducing additional health complications [56].

In 2021 we have celebrated the 100th anniversary of the insulin discovery, that is considered as a milestone in the diabetes treatment changing this fatal disease into chronic one. Nowadays, it is expected that the development of truly noninvasive method for diabetes treatment and management will be another milestone, and exhaled breath analysis is considered as an ideal option. Therefore, the main aim of this paper is to review the algorithms used to detect diabetes in exhaled air and evaluate the effectiveness of these algorithms. Researchers have presented different approaches to diabetes diagnosis. The diabetes and healthy classification [57–60] was mainly carried out, but due to the practical application, some researches proposed the type 1 (T1DM), type 2 (T2DM), and healthy classification [61–63], as well as healthy, prediabetes and diabetes classification [64]. In addition to using direct classification, a diagnosis based on predicted blood sugar concentration has also been proposed [65]. The researches were carried out the analysis on the breath samples taken from patients [57, 58, 60, 62, 64, 66–68], and simulated acetone concentrations [40, 68–70]. The algorithms presented in the papers have been developed mainly in MATLAB® programming and numeric computing environment [62, 63, 71] and in Python programming language with Keras interface for designing artificial neural networks (ANNs) [64]. The general diagram used in the design of breath analysis devices is shown in figure 2.

2. Diabetes mellitus

Diabetes mellitus (DM) is a chronic metabolic disease associated with impaired insulin secretion and/or function. Several types of diabetes are known, but

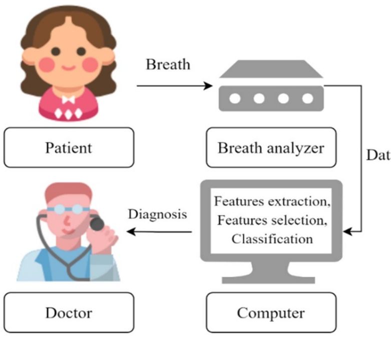


Figure 2. General diagram used in the design of breath analysis devices.

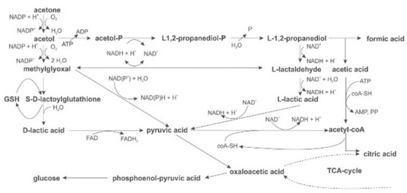


Figure 3. Acetone biosynthesis. Reproduced from [88]. © 2017 IOP Publishing Ltd CC BY 3.0.

the most common are type 1, type 2 and gestational DM [72]. Type 2 diabetes is detected in 90% of cases. People who are not physically active, are obese, are elderly, or have a family history of diabetes are at high risk of developing this type of diabetes [73, 74]. In this case, tissues become resistant to insulin and the pancreas cannot produce enough insulin for the body to function properly. A different mechanism causes type 1 diabetes, in which the patient's body stops producing insulin due to the autoimmunology aggression on beta-cells [75, 76]. The major difference from the patient perspective is that type 2 diabetes can be prevented, for example by appropriate prophylaxis, quitting smoking, changing diet, and incorporating physical activity [77], but at present there are no known methods of preventing type 1 diabetes [78]. Diabetes is manifested by hyperglycaemia, polyuria, polydipsia, weight loss (sometimes with polyphagia), and blurred vision. Untreated, it leads to many complications of the circulatory, nervous, and visual systems, as well as the development of a diabetic foot. For this reason, it is essential to carry out screening tests to detect the disease as early as possible and start treatment [74, 79].

Due to the high potential of using exhaled air for noninvasive detection of diabetes, researchers compared breath samples from diabetics and healthy individuals to identify differences and select potential diabetes biomarkers. Yan *et al* showed that isopropanol and 2,3,4-trimethylhexane, 2,6,8-trimethyldecane, tridecane and undecane in combination can be T2DM biomarkers [80]. In a study conducted by Nelson *et al* differences in exhaled acetone in diabetics and healthy infants were observed, and the content of exhaled isoprene was comparable in both groups [32], while Neupane *et al* suggested that isoprene could be used to detect hypoglycaemia in patients with type 1 diabetes [42]. Trefz *et al* also observed increased levels of isopropanol and isoprene in diabetics compared to the healthy control group [81]. Since the origin of isoprene in the breath is most often associated with cholesterol and fat levels [82] and the small amount of research confirming its relationship with diabetes, it has not been clearly recognized as a diabetes biomarker. For a long time, most research has focused on the correlation of acetone in the exhaled air and diabetes. Numerous research including comparisons of the composition of exhaled air show an increased concentration of acetone in diabetics [6, 29–36, 83]. The biochemical sources of acetone in the exhaled air and its metabolic relationship with diabetes are also now known [31, 84, 85].

Acetone is present in exhaled air when the body produces excess acetyl-CoA, as it was illustrated in figure 3. This molecule is formed in hepatocytes to which free fatty acids resulting from lipolysis have been pretransported. The second source of this molecule is the glycolysis process, i.e. the conversion of glucose into pyruvic acid in the cytoplasm of the cell, and then in the mitochondrion of the cell, the pyruvic acid is converted into acetyl-CoA, which is transferred to the Krebs cycle [84, 86]. Uncontrolled diabetes leads to an increase in free fatty acids and therefore excess acetyl-CoA, which does not end up in the Krebs cycle, but in the ketogenesis process [86, 87]. In this process, acetoacetyl-CoA is formed from two acetoacetyl-CoA molecules, which undergoes successive transformations, the product of which is acetoacetate. Subsequently, spontaneous, nonenzymatic decarboxylation of acetoacetate leads to the formation of acetone, which is absorbed into the bloodstream and excreted through the alveoli with exhalation [31, 84, 85].

Based on the literature review the exhaled acetone concentration vary in the range of 0.176–25 ppm, as it was shown in table 2. Although, the concentrations of acetone in the breath of healthy people were lower than for patients with diabetes.

3. Data acquisition systems

3.1. Gas sampling methods

The gas-sampling procedure is a crucial element of exhaled breath analysis. Generally, it can be divided into two subcategories, the direct and indirect methods, whereas the indirect method is the most popular.

Table 2. Acetone concentration in health and diabetes samples. Reproduced from [40]. CC BY 4.0.

Diabetic stage	Measured acetone concentration	References
T2DM	1.76–3.73 ppm	[31]
Healthy	0.22–0.80 ppm	[31]
Controlled diabetic	0.19–0.66 ppmv	[35]
Untreated T2DM	0.92–1.20 ppmv	[35]
Diabetes	1.25–2.5 ppm (or up to 25 ppm)	[36]
Healthy	0.2–1.8 ppm	[36]
T1DM	4.9 ppm	[83]
T2DM	1.5 ppm	[83]
Healthy	1.1 ppm	[83]
Diabetes	>1.8 ppmv	[89]
Healthy	<0.8 ppmv	[89]
T1DM	2.19 ppmv (mean)	[90]
Healthy	0.48 ppmv (mean)	[90]
Healthy	0.177–2.441 ppm	[91]
Healthy	0.176–0.518 ppm	[92]

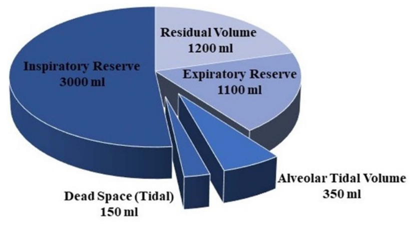


Figure 4. Breath fractional volumes for an average size healthy male at rest.

Table 3. Gas-collecting methods.

Method	System	References
Indirect	Air supplied directly to the device	[62, 63, 65, 66, 71]
Direct	Tedlar® bags	[57–60, 66, 67, 93, 94]
	Fluorinated ethylene propylene (FEP) bag	[83]
	Teflon sealed bag with saliva and moist filter	[68]

In the direct method, the gas sample is exhaled directly into the device [62, 63, 65, 66, 71], while in the indirect method it is collected in specially designed bags, for example Tedlar® bags [57–60, 67, 69, 93, 94], Teflon sealed bag with saliva and moist filter [68] and fluorinated ethylene propylene breath gas collection bag [83]. Tedlar® bags (Dupont de Nemours) which are made of polyvinylfluoride, are the most popular bags used in breath sampling studies [95]. The air in the human lungs is composed of death that has a volume of approximately 150 ml and approximately 350 ml of alveolar volume [7, 96]. Breath fractional volumes for an average size healthy male at rest [96] are shown in figure 4.

The most important compounds for the biomarkers analysis, such as VOCs, are in the endtidal part of exhalation. Endogenous particles concentrations are highest at the end of expiration, when the endtidal pressure of exhaled CO_2 reaches a plateau, therefore it is recommended to use capnometers or CO_2 sensors and consider the obtained results during preprocessing and selecting the data fragment to determine the sensor responses [41, 97]. Within the indirect method, patients, for example, were asked to blow into Tedlar® bags flushed with N_2 and heated at 40 °C for 1 h [66]. In the case of the use of Teflon sealed bags, the bags were flushed with dry air [68]. Silica gel dehumidifier is commonly used as hygroscopic material [57, 58, 66] and did not show an influence on diabetes detection [57]. Collected breath samples were stored at 4 °C until analysis [60].

Beauchamp *et al* compared the stability of VOCs after storing the breath sample in Tedlar® bags for 10 and 70 h. The result shows that the above 80% compounds in the sample remain stable within 10 h, but within 70 h the percentage of recovery is unacceptable for future breath analysis. They also proved that there is a diffusion of compounds through the surface of the bag. This is noticeable by the exponential decrease in relative humidity (RH), tending to the ambient level. Another evidence of diffusion is observed inside the bags the increased level of contaminants from the environment. Nevertheless, the decrease in humidity inside the bag can be an advantage for measurements that are sensitive to high humidity [95].

In another study, Righettoni *et al* also showed that the humidity inside the Tedlar® bag decreased during measurements. For this reason, they decided to heat the breath sample in the bag for 1 h at 40 °C. This allowed the RH to stabilize from about 90% to less than 30%. Moreover, the RH content inside the collecting bag has decreased tending to the humidity level contained in the ambient air [98].

The research carried out by Mansour *et al* showed that the respiratory temperature is the range of 31.4 °C–35.4 °C and 31.4 °C–34.8 °C, while the RH is 65.0%–88.6% and 41.9%–91.0%, for Halifa and Parisian participants, respectively [99]. Other studies by Ferrus *et al* showed that RH in human exhaled air is in the range of 89%–97% [100]. Table 3 shows different gas-collecting systems.

3.2. Systems for exhaled breath analysis

Currently, there are several known commercial devices (e-nose systems) on the market that are used to diagnose disease based on exhaled air, but none of them have been adapted to detect diabetes in the exhaled air. The most popular are FOX 4000 (AlphaMOS, France) [101, 102], FAIMS

Table 4. The most commonly used sensors in exhaled breath analysers for the detection of diabetes according to the literature review.

Sensor	Target gas/measured value	References
TGS4161	Carbon dioxide	[57, 59]
TGS822	VOCs, hydrogen, carbon monoxide, etc.	[57–59, 71]
TGS1820	Acetone	[40]
TGS8100	Air contaminants	[40, 113]
TGS2620	Alcohol, solvent vapors	[40, 57–59, 71, 113]
TGS825	Hydrogen sulphide	[57, 71]
TGS826	Ammonia, VOCs, hydrogen	[57, 58]
TGS2201	Nitric oxide, Nitrogen dioxide	[57]
TGS821	Hydrogen	[57, 58]
TGS2602	VOCs, hydrogen, ammonia, hydrogen sulphur	[57–59]
MiCS-5524	VOCs	[40, 64, 65]
MQ-2	Propane, hydrogen, methane	[60]
MQ-3	Alcohol	[40, 60, 62, 63]
MQ-5	Hydrogen, acetone, carbon monoxide, alcohol	[62, 63]
MQ-7	Carbon monoxide	[64, 65]
MQ-9	Carbon monoxide	[60]
MQ-135	Benzene, ammonia, carbon dioxide, nitric oxide	[60, 64, 65]
MQ-137	Ammonia	[60]
MQ-138	Toluene, acetone, ethanol, formaldehyde	[60, 64, 65]
HTG3515CH	Temperature, relative humidity	[58, 59]
DHT-22	Temperature, relative humidity	[64, 65]
SHT85	Temperature, relative humidity	[40]
Honeywell HIH 4000–002	Relative humidity	[60]

breath analyser (Owlstone Medical, UK) [103], Ketonix Bluetooth and USB noninvasive breath analyser (Ketonix AB, Sweden) [104, 105], Portable Breath Acetone Meter PBAM (Biosense™ Readout Health, USA) [106, 107], Cyranose Electronic Nose (Sensigent, USA) [108, 109] and Keyto Breath Sensor (Keyto, USA) [110].

Most often, breath analyser systems consist of matrices of several commercially available sensors sensitive to selected compounds [59, 60, 62–65, 71, 94, 111, 112]. Additionally sensors for the detection of RH [58–60, 64, 65] and temperature [58, 59, 64, 65] are also used in breath analysers. The most popular sensors, reported in the literature for the utilization in breath analysers for the detection of diabetes were are shown in table 4.

3.2.1. Metal oxide semiconductor sensors

The most popular gas sensors used in breath analysis are metal oxide semiconductor (MOS) and temperature modulated MOS (TM-MOS) sensor arrays specially optimized for the requirements of a breath analysis system by selecting sensors with increased sensitivity and selectivity to the potential biomarker of the detected disease [59, 93]. The MOS sensor resistance changes during interaction with oxidizing and reducing gases [71, 94, 113]. TM-MOS sensors showed greater efficiency in respiratory analysers than traditional MOS sensors [59]. The high RH of exhaled air affects the results of the measurements, especially those performed with the use of MOS sensors [63, 95, 114], which is why many groups of researchers use humidity absorbing systems and systems to measure humidity and temperature, which are used to compensate for their impact in the algorithms created.

3.2.2. Polymer sensor array

Polymer sensors consist of electrodes made of metal alloys, for example Pt/Pd, arranged on the surface of a ceramic substrate. The electrodes are deposited in the polymer layer synthesized by polymerization. Thickness of sensor can be adjusted by controlling the polymerization parameters such as amount of organic solvent. Detailed information on the production of polymer sensors was described by Yu *et al* [68]. The authors used a portable gas analysing system constructed from a conducting polymer sensor array (polypyrrole). The designed sensor array consists of Pt/Pd alloy electrodes placed on the surface of an alumina substrate coated by polypyrrole thin film sensors with different thicknesses obtained by chemical polymerization. The applied layers showed different levels of response to acetone and ethanol concentrations, which were controlled by the mass flow controller. The sensors were tested on breath samples of diabetics and healthy people as well as various concentrations of acetone and ethanol [68].

3.2.3. The proton exchange membrane fuel cell

Proton exchange membrane fuel cell (PEMFC) consists of a polymer electrolyte membrane that allows proton conductivity and transport of protons from the anode to the cathode. In the presence of gas, the potential for the working electrode changes [69, 115]. Jalal *et al* proposed a system for the acetone real-time monitoring composed of three-electrode fuel cell sensor. The PEMFC sensor was constructed from a sandwiched structure of the membrane electrode assembly. These sensors are characterized by long life, scalability, portability, good accuracy, and selectivity; however, they are susceptible to external factors such as pressure, humidity, and temperature. The sensors were tested on various concentrations of acetone [69].

3.2.4. MEMS (micro-electro-mechanical systems) cantilever sensor

The polymer-coated MEMS cantilever sensor consists of a cantilever beam with a rectangular cross section. It is arranged in a capacitor configuration with parallel plates. The cantilever plate is movable relative to the rigid substrate. Depending on the components of the gases that have been applied, the parameters of the cantilever such as the cantilever mass, stiffness, and surface stress condition change due to the sorption of particles on the polymer surface that covers the cantilever. This causes the cantilever to bend. The value of the deflection depends on the concentration of the compound to which the coating is selective, and it can be measured in the static mode sensing, i.e. with a laser reflectometer or by measuring capacity changes. Another measurement method is the dynamic mode sensing, which involves electrostatic induction of the cantilever by applying AC voltage to the capacitor plates and measuring electronically or by a laser Doppler vibrometer the change in resonance frequency caused by the presence of gases [70, 116]. Gupta *et al* designed a polymer-coated MEMS cantilever sensor. The authors selected polymers for the cantilever coating using fuzzy C-means clustering and fuzzy subtractive clustering methods. The sensors were tested on simulated artificial breath [70].

3.2.5. Gas chromatograph coupled to a mass spectrometer

In a gas chromatograph (GC) coupled to an mass mpectrometer (MS) system, the GC enables the separation of the analyzed mixture into its components over time. Separation of the mixture components takes place due to the differences in the migration rates of the individual components of the mixture through the chromatographic column. The mass spectrometer then records their mass spectra, on the basis of which each of the components of the separated mixture can be identified by measuring the mass to charge ratio [46, 117]. Siegel *et al* used an Agilent 7890A GC coupled to an Agilent 5975C mass spectrometer to analyze breath samples collected in Tedlar bags from 56 type 1 diabetes patients. To obtain as many as possible unidentified trace components, the automated mass spectral deconvolution and identification system was applied, and the components have been automatically identified using the SpectConnect server at the University of Georgia. They decided to apply manual prescreening to exclude components that were observed in less than 50% of hypo or total samples, eliminating the influence of known contaminants from components present in the measurement equipment or Tedlar bags and components with low signal-to-noise ratio [67].

4. Preprocessing

The measured responses are commonly filtered, normalized [40, 42, 47, 50, 69, 76, 77], and preprocessed using baseline manipulation [57, 94, 118] and/or baseline subtraction [57–59, 94, 119]. Figures 5 and 6 show the baseline fitting and subtraction. Baseline manipulation is used to enhance contrast, compensate for drift, and scale the data [24, 57]. An interesting method for computing the sensor response after a baseline manipulation process is given in [24], where the baseline manipulation is expressed by equation (1):

$$\mathrm{SR}^{\mathrm{baseline}}_{(S,D,T)} = \mathrm{SR}_{(S,D,T)} - \frac{1}{B_N}\sum_{t=1}^{B_N}\mathrm{SR}_{(S,D,t)}, \qquad (1)$$

where, $\mathrm{SR}_{(S,D,T)}$ and $\mathrm{SR}_{(S,D,t)}$ are sensors responses to sample S from D sensor response data, at time T and time t, B_N is stabilized data.

Further, data normalization is applied to eliminate the fluctuation caused by analyte concentration and oxygen pressure [57, 118]. Raw response normalization can be performed in two different ways, feature wise (column wise normalization) and sample wise (row wise normalization) [119]. When humidity and temperature sensors were used in their analysis system, the algorithm was designed to compensate or handle variations in collected breath samples caused by humidity and the presence of alveolar air [59]. To reconstruct the signal and compensate for noises whose source could be unstable voltage or variable temperature and humidity, the discrete wavelet transform was applied to the Z-normalized signal [64].

4.1. Features extraction

The most popular approach is to use preprocessed raw data obtained from sensors to train machine learning models. To increase the performance of breath analysis systems, feature extraction algorithms are commonly used before training selected classification or regression models [60, 63, 93, 119]. Commonly used methods to extract features from raw sensor data are, for example, the calculation value of maximum response, integral/area under the sensor's response curve (AUC), maximum derivative, and extreme response. The use of these methods significantly increases the dimension of features [60, 93, 119]. A more advanced feature extraction and reduction technique is the singular value decomposition (SVD) algorithm, which decomposes the input matrix into orthogonal eigenvectors and eigenvalues [63].

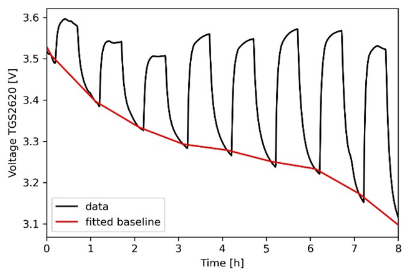


Figure 5. Data before the baseline subtraction with the fitted baseline.

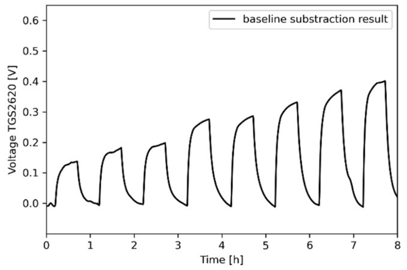


Figure 6. Data after baseline subtraction.

4.2. Features selection

To remove irrelevant or superfluous features and keep the most useful information, feature selection algorithms are applied [63, 93, 119], such as SVD, sequential forward selection, principal component analysis (PCA). Due to the high dimensionality of the data set that contains responses frequently from more than ten different gas sensors [57, 59, 93, 94, 119] and the increasing amount of features after applying feature extraction methods, algorithm is used to extract low dimensional features, because the PCA algorithm projects the higher dimensional data onto a lower-dimensional data subspace [57–59, 64, 118, 120]. The feature is also omitted when it has a high correlation with the other or alone provides low accuracy in the given task [60, 93, 119]. Another way to decrease the number of not significant features are statistical methods such as screening, selection, and multiple comparisons. Elimination of the impact of components with a significantly low signal-to-noise ratio and pollution present in Tedlar bags or measurement devices can be done by applying the manual prescreening method [67]. Reducing the number of redundant features can improve model accuracy and training speed by decreasing the computational complexity of algorithms [67, 68, 93] and prevent overfitting, which is common problem when analyzing data from multiple gas sensor matrices because the number of features may exceed the number of breath samples taken [24, 120, 121].

5. Algorithms

5.1. Principal component analysis

PCA currently is the most popular dimensionality reduction algorithm. It is an unsupervised learning model and the main aim of this method is to find the hyperplane closest to the data and then project the data onto it. The algorithm based on the PCA determines the axis that retains the highest value of the variance of the training data set. The number of determined mutually orthogonal axes is equal to the dimension of the features. A training set is transformed into a dot product of three matrices using SVD algorithm [122]. One of the resulting matrices contains the principal components arranged in the order of decreasing variance [70], which are then used to project the original training data set to a given number of dimensions (equation (2)).

$$X_d = X \cdot W_d, \qquad (2)$$

where X_d is the result of dimensionality reduction, d is the number of target dimensions, X is the original data set matrix, and W_d is a matrix of the d-first values from the principal component matrix calculated using SVD method.

Usually, instead of choosing randomly the number of target dimensions of the feature space, the number of dimensions allows to keep a given value of variance [122] and at the same time minimizing information loss [123]. On the other hand, preserving the highest variance by the PCA algorithm can also be a drawback in tasks such as regression [93]. The individual principal components are not correlated with each other [124].

Saidi *et al* used the PCA algorithm to differentiate between diabetes and health states. The authors have extracted the features such as dynamic slope of the conductance ($\mathrm{d}G/\mathrm{d}t$), AUC, and conductance change (ΔG) for each of sensors, but the result shows that only AUC and ΔG had the ability to classify health and diabetes samples and other sensors' responses were strongly correlated. The algorithm assigns the new sample to one of four classes: DM, chronic kidney disease, healthy subjects with high creatinine and healthy subjects with low creatinine, which have been separated in a three-dimensional graph using PCA methods. Moreover, the trained model correctly classified samples obtained after one month, which means that the model was stable and did not need retraining or recalibration. External validation carried out on samples from new people showed that the algorithm is 100% correct in diseases identification [60]. Eliminating features with low variance helps to decrease overlapping between two or more different concentration points [69]. Additionally, the PCA method can be used to calculate the Euclidean distance between samples. Results show that this metric has the ability to differentiate between healthy and diabetic patients. The Euclidean distance was calculated on the samples taken from exhaled breath of diabetes, non-diabetes as well as an reference gas acetone sample, where diabetes and reference, and non-diabetes and reference distances were calculated. The results have shown that the Euclidean distance between non-diabetics and reference acetone was higher than for diabetics and [68].

This unsupervised learning method is also commonly used to preprocess the data and reduce the dimensionality of complex data sets before applying supervised learning methods such as linear discriminant analysis (LDA) [67], *k*-nearest neighbours (KNN) classifier [57, 58], because results obtained independently by PCA algorithm may be not sufficient to properly generalize and classify breath samples [67].

Dimension reduction is also widely used for data visualization. To prepare clear and human-readable visualizations of multidimensional data is needed to limit the number of dimensions to two or three [67, 122].

5.2. Support vector machines

Support vector machines (SVM) are supervised learning models which can be used to solve classification and regression problems [61, 125]. The SVM classifier is one of the most frequently used classification algorithms for a small but high-dimensional data set [69], which most often characterizes medical data, especially VOCs measured with the use of several independent sensors or semiconductor gas sensor matrices [59, 60, 62–65, 71, 94, 111]. The principles of SVM are in detail presented and discussed in the literature, for example Burges [126] and Cherkassky *et al* [127]. Briefly, the *m*-dimensional input data set is separated by the algorithm into *l*-dimensional feature space using hyperplanes. Hyperplane is the decision surface obtained by solving the optimization problem to maximize the margin [125]. The number of determined hyperplanes is equal to $n - 1$, where n is the number of classes. The margin is calculated as the distance between the hyperplanes that are defined by moving the boundary hyperplane all the way to the first points of the classes. The points closest to the boundary hyperplane form the support vectors. The selection of the appropriate margin plays a key role in the design of the algorithm. If the margin is too small, a slight change in the decision boundary may even result in a change of the predicted class. On the other hand, a wider margin makes it possible to limit the phenomenon of overfitting and increases the generalization of the model. For these reasons, it is imperative to use margin maximization algorithms. SVM classification is robust to outliers [126–128].

For non-linearly separable datasets, the choice of a linear decision function can lead to underfitting the model to data. To avoid this problem, the kernel trick with the maximum-margin hyperplanes is widely used. In such case, every dot product is replaced by a nonlinear kernel function [125]. SVM classifiers commonly used in breath analysis systems were trained with different kernels, e.g. Gaussian kernel [59], a polynomial kernel function (3) [94], radial basis function as presented by the following equation (4) [111].

$$K(x, x_i) = (x' x_i + \gamma)^p, \tag{3}$$

where γ is a kernel parameter and p is polynomial degree [94].

$$K(x, x_i) = \exp\left(\frac{-\gamma x - {x_i}^2}{\sigma}\right), \tag{4}$$

where x is the class, x_i is the data set corresponding to that class, and σ is the variance of the testing data [111].

To avoid model overfitting, the leave-one-out cross-validation can be applied to SVM algorithm, as was presented by Saidi *et al* in [60].

One of the main disadvantages of the SVM algorithm is the choice of the proper kernel which will fit well input data and decrease time of training, especially for large datasets [125]. Results obtained by different researches show that SVM classifier model can obtain high sensitivity and specificity greater than 90% on the separated test set for the differentiate between health and diabetes VOCs samples [59–61, 94, 111]. Yan *et al* used the SVM algorithm to differentiate between the breath samples of healthy people and diabetics. The algorithm was trained on 140 randomly selected samples from sick people and the same number of samples for healthy people. The remaining samples (139 for each class) were used to validate the model. The authors trained the model 50 times, as a result, they obtained an average sensitivity of 91.51% and a specificity of 90.77% for diabetes screening [59].

5.3. K-nearest neighbour

KNN is a supervised learning algorithm that classifies a given unknown sample into a category based on the distance between k closest examples from the training data set. The distance between samples p and q in n-dimensional space is usually defined by Euclidean distance (5) [129–132] and less often by Manhattan, Minkowski [132] and Canberra metrics. Canberra distance is weighted Manhattan distance [65].

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \tag{5}$$

where p and q are two samples in n-dimensional space.

In the KNN classification algorithm new observations are assigned to the class in the following way: placing the new observation x in the data space, determining the KNN according to the selected distance metric, and then predicting the class based on the majority of classes that are represented by the selected KNNs. The class for the new observation can be selected not only on the basis of most classes, but also by a weighted method [133–135].

One of the difficulties in applying this method and obtaining satisfying results is to choose an accurate k number of the nearest neighbours because there are no strict or mathematical rule to determine the value of this parameter [129]. Too large k value results in a simplification problem and loose local information, on the other hand, if k is too small, the model becomes too sensitive to outliers [136]. Commonly, it is chosen after many experiments and comparing the accuracy, sensitivity, specificity metrics between different models [130]. Features in favor of using this classifier are its simplicity of implementation and possibility to adapt to local information, but on the other hand, this model has very high computational complexity associated with calculating the distance to each training example for each new sample [129].

In breath analysis, the KNN model is commonly prepared using features extracted from a data set using PCA method [57, 58]. Researches design the algorithm with small k values, e.g. $k = 3$ [58], $k = 5$ [57] and $k = 8$ [65]. Classification results using this classifier are promising, researches obtained the sensitivity of the diabetes diagnosis higher than 87%, the specificity higher than 86% [57, 58] and the accuracy 95% [65]. Yan *et al* used the KNN algorithm to distinguish between breath samples of healthy people and diabetics. The authors collected 294 healthy samples and 404 diabetes. 147 samples from each class were selected for training. The algorithm was tested on the same number of samples. The authors trained the model 50 times, as a result, they achieved an average sensitivity of 91.43% and a specificity of 89.86% for diabetes screening [58]. Guo *et al* trained the KNN algorithm to distinguish breath samples between healthy and sick people using data from 57 people with diabetes and 48 healthy people. Algorithm tests were performed on separate samples (60 for each class). The sensitivity of this diagnosis was 87.67% and the specificity was 86.87% [57].

Hariyanto *et al* showed that the KNN algorithm results in higher accuracy and was easier to implement than the SVM and neural networks (NN) models. Moreover, the designed KNN classifier performs the fastest classification in the web system designed by researches [65]. This method can be also applied to regression problems [129, 130].

5.4. Linear discriminant analysis

LDA is a supervised machine learning algorithm commonly used for binary classification problems [67, 132, 137]. The main goal of the LDA model is to maximize the variance between classes and minimize the variance within each group [132, 137, 138] by finding a projection vector that separates the input data in two categories [132, 138]. LDA assumes that the probability density functions in the classes k have D-dimensional normal distributions N with equal covariance matrices in the classes. In this algorithm, matrices characterizing the intergroup and intragroup variability are calculated, the ratio of determinants of these matrices is maximized. The main goal of the algorithm is its minimization of intra-group variability, and the maximization of inter-group variability [139]. The model is prone to overfitting on a training data set if it has a large number of variables [67]. LDA can be used as a linear classifier or feature selection method before applying other classification algorithms [132, 138].

Siegel *et al* attempted to use LDA model to detect hypoglycaemia in breath samples by distinguish hypo from non-hypo samples. To increase and test performance of the model leave-one-one-out cross-validation (LOOCV) method was used. The calculated metrics were 91% sensitivity and 86% specificity for the training dataset. These results prove that due to the observed differences in the composition of breathing during a hypoglycaemic episode, it is possible to create an automatic hypoglycaemia detection system [67].

5.5. Extreme gradient boosting

Extreme gradient boosting (XGBoost) is state-of-the-art algorithm which was developed by Chen and Guestrin in 2016 [140]. This algorithm can be successfully used in both regression and classification problems and often achieves higher performance than traditional machine learning algorithms [140–143]. It is an algorithm based on many decision trees. This type of learning is called ensemble learning. This algorithm makes a prediction based on the results obtained by individual trees, which are added during the training of the algorithm. In the case of using Gradient Boosting, the loss of the algorithm is minimized using the gradient descent method. The XGBoost algorithm shows high efficiency in dealing with missing data, and additionally, thanks to parallelization and hardware optimization through the use of out-of-core and cache-aware computing, it is one of the fastest operating algorithms, especially in the case of large data sets [140].

Paleczek *et al* used XGBoost to detect diabetes based on acetone concentration in a simulated artificial breath. The hyperparameters of the model were optimized using the grid search and cross-validation method to avoid overfitting. The algorithm was optimized to achieve high recall, which was, on the test data set, 100%, while the F1-score was 97.4%. The researches also compared the results achieved by traditional machine learning algorithms such as Decision Tree Classifier, Random Forest Classifier, SVM and KNN. The results of the comparison showed that the other decision tree-based algorithms achieved similarly high performance as XGBoost [40].

5.6. Neural networks

Not only traditional machine learning methods are used to detect diabetes in the exhaled air, but also different types of NNs such as ANN [71], deep NN (DNN) [64] and convolutional NN (CNN) [62, 63] were proposed.

Nowadays, ANNs are referred to as a computing architecture that consists of extremely high number of massively parallel simple neuron interconnections [144]. Each of the inputs is assigned a weight by which it is multiplied. The higher the weight of a given connection, the associated input has a greater effect on the output value of the neuron. If it is decided to use bias, its value is added to the multiplication result. Then the results of these multiplications and adding biases are summed up and the appropriate activation function is applied [144]. Widely used activation functions are unipolar and bipolar step functions, symmetric ramp function, logistic function, and hyperbolic tangent function. The choice of the right one depends on the problem faced by networks [145]. The main goal of ANN is to adjust the weight vector in such a way as to obtain the most accurate representation of the expected output values [144].

DNNs are in general ANN with additional hidden layers. Due to this modification, DNNs have a higher computational complexity. In deep learning, nonlinear data transformations are performed using many hidden layers. The lowest layers represent the basic features of the input data set, the next layers create more detailed features. In the case of using DNNs, it is not necessary to properly prepare the features, because it is implemented in individual deep layers of the network. One of the advantages of using DNN instead of ANN is that in contrast to ANN, it is possible to use a dropout layer in DNN [64]. This technique is the most commonly used technique for weight regularization. It consists of randomly switching off the weights of neurons with a given probability (most often 50%) [146].

The last of the mentioned NN structures, CNNs, are commonly used to deal with grid-like data topologies such as time series or images. The network performs a linear convolution between an input signal and a predefined kernel. The CNN output is called a feature map. In classification problems, the calculated features are transferred to a fully connected layer and then assigned to classes by the softmax layer according to the calculated highest probability of belonging to a class. There are several network regularization and optimization methods such as adding dropout layer, pooling, L^2 regularization, L^1 regularization, and batch normalization, which are commonly used to increase model generalization and avoid underfitting or overfitting the network to input data [146].

Prepared models can be optimized using different optimizers. The most popular are Adam, Adagrad, Adelta, Stochastic Gradient Decent [64]. The width of the model is determined by the dimensionality of the used layers [146].

Lekha *et al* proposed to modify convolutional layers and use a Gaussian based 1D kernel to filter the raw input signals and produce the feature maps forwarded to the next layer of the network. Researches attempted to compare the proposed 1D CNN classifier model with different machine learning models such as SVM, NN combined with preprocessing techniques such as PCA or SVD. During leave-one out cross validation algorithm achieved misclassification rate of 0.0714 and a mean square error of 0.1436. The obtained results showed that the calculated accuracy, sensitivity, and specificity metrics, respectively, almost 98%, 97.5%, and 97.5%, were higher for CNN than for SVM and NN, preceded by the feature extraction algorithms in the one-label multiclass classification problem (diabetes type 1, type 2 and healthy samples classification). Moreover, it was observed that the measured computational complexity of the 1D-CNN algorithm was significantly lower [63].

To obtain better performance, in another study, it was proposed to use CNN with Gaussian kernel and modify it by replacing the fully connected multilayer perceptron (MLP) classifier with the SVM algorithm. The proposed architecture integrates the concept of CNN feature extraction technique with the SVM classifier. The aforementioned modifications increase the classification accuracy, sensitivity, and specificity metrics caluculated on separated test set, respectively, to 98%, 99%, and 98%. The algorithm was trained on the signals from the examination of 15 breath samples, and six samples were used for testing. Moreover, replacing fully connected MLP with the SVM classifier significantly decreased the measured computational time. Researchers also compared the performance of other architectures such as traditional CNN, CNN-SVM with linear kernel, and CNN-SVM with polynomial kernel, but none of them showed a better ability to distinguish between the three classes which are healthy, type 1 diabetic and type 2 diabetic and the computational complexities were higher [62].

Not only the CNN architectures have been used, but the optimized DNN structure has been proposed by Sarno *et al* to separate the collected samples into three predefined categories such as healthy, prediabetes and diabetes. Before performing the classification, PCA algorithm was used to select the most significant averages of sensor responses. Research compared the results for different optimizers and the Adam optimizer showed the utmost accuracy that was 96%. The results obtained by the proposed optimized DNN were compared with commonly used machine learning methods such as KNN, SVM, Naive Bayes (NB), and LDA. These methods showed much lower accuracy than the proposed NN based model, where the highest was 83.33% for the SVM classifier [64].

6. Discussion

Metabolic variability between patients and other factors such as gender, age [67], as well as length of time living with type 1 diabetes [147] and course of treatment [35] hampers the development and testing of versatile e-nose system for the detection of diabetes in exhaled breath. Furthermore, as is shown in the literature, breathing acetone levels differ between type 1 diabetes and type 2 diabetes [83] and depend on whether the disease is being treated and controlled [35], therefore it is necessary to develop and test the e-nose system based on breath samples obtained from a large in a group of patients, varied in terms of metabolic conditions, medical history and type of disease, not just based on acetone concentration. One of the possibilities of customization the device for the patient is the development of the device calibration function [93], e.g. using a traditional blood sugar meter, and then using it to determine the individual relationship between blood glucose level and the exhaled acetone level obtained from the breath analyzer.

To prepare a versatile breath analysis system, it should be considered whether intrasubject variance factors such as diet [148] and insulin [149] should be handled by the algorithm of the system. In the papers presented in this review, researchers trained algorithms with input data obtained only from measurement devices. Additionally, the influence of metabolic factors on the relationship between acetone and blood glucose level is unknown. Due to the high and variable RH of human breathing, it is necessary to measure it during the breath analysis, as well as to take into account its influence on the sensor response in the algorithms [40, 63, 95, 114].

Based on the literature review it was observed that the studies are conducted on a small number of patients, therefore it is difficult to prepare and validate a system for breath analysis taking into account the variables affecting the acetone level in the exhaled air [63].

Another limitation of the use of breath analysis for the noninvasive diagnosis of patients is the very low concentration of VOCs in exhaled air [7]. The performance of breath analyzer can be increased by selecting sensors highly selective to acetone or by using other measurement methods and by using pre-concentrators at the input of the analyzers [35, 38, 39]. Ueta *et al* used In-needle preconcentration method whose the main advantages are repeatability without performance decrease and simple extraction/desorption process. The calculated recovery obtained using the needle extraction device was more than 99% for the standard acetone sample [22]. Rydosz *et al* proposed to use a pre-concentrator manufactured with the LTCC method. The results show that the concentration factor depends not only on measurement parameters such as adsorption time, gas flow and desorption temperatue, but also on absorbent volume, grain size and surface area. The obtained concentration factors, calculated for acetone, were up 16.35 at 30 min adsorption time. The main advantage of the LTCC micropreconcentrator is the possibility of integrating it with a matrix of gas sensors manufactured using this method [25, 26].

The breath analysis systems and diabetes detection algorithms presented in this paper have shown the importance of data preprocessing, as well as the appropriate features extraction and selection used to train the algorithm. In the case of systems based not only on MOS sensor array, but also on GS/MS and other systems presented in this paper, the researchers obtained a large number of features for the input data. In addition, they processed the raw data obtained from the measurement system in different ways, for example, by calculating the AUC [60] and/or the DWT [64], to obtain the highest possible efficiency of the system. Determining which of the many features obtained from the measuring device significantly increases the performance of the algorithm is possible, for example, by using tree algorithms that allow checking feature importance [40]. This information can be useful to determine a sufficient number of sensors and their type when custom designed sensor matrices are used.

PCA is a helpful algorithm to reduce the dimensionality of data, often used by researchers to reduce the number of input features to the algorithm, as well as for visualization and exploratory data analysis. In addition to using this method for feature selection, it is possible to use it as a classifier, but research has shown that better results are achieved by classifying the computed principal components using other machine learning algorithms such as LDA [67], KNN [57, 58].

An alternative to PCA, has been proposed by Lekha *et al*, The autors have used CNNs to extract features from sensor data [63]. The results showed that the use of CNN with the SVM showed higher performance in detecting diabetes than the use of CNN with the fully connected multilayer perceptron layer [62].

All algorithms presented in this paper show very high performance in diabetes detection tasks based on multidimensional data obtained from exhaled air analyzers. An important element in the design of algorithms is the selection of model hyperparameters, e.g. *k* number of the nearest neighbors, to ensure the best generalization and prevent overfitting [136]. Model overfitting can also be reduced

Table 5. Comparison of selected algorithms used to detect diabetes in exhaled breath from patient samples.

Algorithm	Features	Data samples	Accuracy	Sensitivity	Specificity	References
LDA	PCA	52 Type 1 DM (hypoglycaemia detection)		91%	84%	[67]
LDA	PCA, DWT	10 Healthy/ prediabetes/ diabetes	74.07%			[64]
KNN	PCA	108 Healthy 117 Diabetes		87.67%	86.87%	[57]
KNN	PCA	294 Healthy 117 Inpatient diabetes 287 Outpatient diabetes		91.43%	89.86%	[58]
KNN	PCA, DWT	10 Healthy/ prediabetes/ diabetes	81.47%			[64]
KNN	DWT	20 Healthy 20 Diabetes	95%			[65]
SVM	Calculated acetone concentrations	3 Healthy 4 Type 2 DM 3 Type 1 DM	100%			[61]
SVM	PCA, DWT	10 Healthy/ prediabetes/ diabetes	83.33%			[64]
SVM	PCA	295 Healthy 279 Diabetes		91.61%	90.77%	[59]
SVM	CNN with Gaussian kernel	12 Healthy 4 Type 2 DM 9 Type 1 DM	98%	99%	98%	[62]
SVM	CNN with polynomial kernel	12 Healthy 4 Type 2 DM 9 Type 1 DM	98%	97.5%	97.5%	[62]
SVM	CNN with linear kernel	12 Healthy 4 Type 2 DM 9 Type 1 DM	97.55%	97.25%	97.5%	[62]
SVM	dG/dt, AUC, ΔG	38 Healthy 6 Diabetes	100%			[60]
SVM	SVD	11 Healthy 9 Type 2 DM 5 Type 1 DM	97.4%	97.1%	97.4%	[63]
SVM	PCA	11 Healthy 9 Type 2 DM 5 Type 1 DM	96.1%	96.9%	94.9%	[63]
SVM	PCA	108 Healthy 90 Diabetes	92.66% (healthy), 93.52% (diabetes)			[94]
CNN	CNN	12 Healthy 4 Type 2 DM 9 Type 1 DM	97.25%	97%	96.5%	[62]
NN	SVD	11 Healthy 9 Type 2 DM 5 Type 1 DM	95.4%	95.9%	94.9%	[63]
NN	PCA	11 Healthy 9 Type 2 DM 5 Type 1 DM	96.1%	96.9%	94.9%	[63]
1D CNN	1D CNN	11 Healthy 9 Type 2 DM 5 Type 1 DM	97.9%	97.4%	97.4%	[63]
DNN	PCA, DWT	10 Healthy/ prediabetes/ diabetes	96.29%	91.61%	90.77%	[64]
NB	PCA, DWT	10 Healthy/ prediabetes/diabetes	74.07%			[64]

by using cross-validation methods such as LOOCV in the case of classic machine learning algorithms [63, 67], while using NNs, popular methods of network regularization are dropout, L^2 regularization, L^1 regularization, and batch normalization [146]. Lekha *et al* compared the use of NNs with SVM classifier. The results showed that CNN outperformed the other algorithms [63], and it is also possible to combine CNN with SVM to increace the performance of diabetes detection [62]. The results obtained by the researches with the use of selected algorithms presented in this review are summarized in table 5.

When analyzing breath samples, one of the challenges is choosing the right algorithm. The results showed that all algorithms achieved high performance. In the case of medical diagnostics, it is worth using explainable AI algorithms, such as decision trees or XGBoost, which allow you to accurately trace the action and decisions made by the algorithm. This is of great importance in the case of misdiagnoses, and the algorithm's conclusions can broaden the knowledge of doctors and indicate which correlations are most important for diagnosis. Another advantage of this type of algorithms is the ability to determine the features importance factors, which are valuable information when the system includes a sensor matrix, using sensors with semiconductor layers can use this knowledge to develop better parameters of these layers or compose a different set of sensors.

Medical diagnostics supported by machine learning and artificial intelligence methods, and especially non-invasive diagnostics based on exhaled air, has many advantages and limitations. Machine learning models enable the processing of a huge amount of multidimensional data, which is constantly increasing, which is not possible to be processed and understood by a human in a reasonable time. Thanks to the use of algorithms, it is possible to detect linear and non-linear relationships between data, detect biomarkers by analyzing the differences between samples from sick and healthy people, as well as select appropriate sensors in the matrices in order to minimize cross-sensivity. A noninvasive respiratory diagnostic device can reduce healthcare costs by performing more screening tests and diagnosing disease early. The main limitation in the use of a AI-assisted diagnostics is the individual variability of the parameters of each patient and the need to train models on a large number of people, with different medical histories, treatments and different livestyles in order to obtain the highest generalization. Continuous validation and calibration of medical devices is also necessary in order to quickly detect errors in the operation of the system. Not all algorithms are explainable, e.g. NNs, therefore in the case of a wrong diagnosis it is practically impossible to discover what influenced the decision of the model [150, 151]. In the case of breath testing, there is no specific protocol for collecting, storing samples and testing procedures, groups of scientists use different bags, as well as direct air supply to devices which mainly consist of different sensors, so it is difficult to compare the results obtained in the literature. It also limits access to data, as compared to other diagnostic methods such as x-ray or MRI, it is not possible to share data from different research centers and create huge databases.

7. Conclusions

The development of the breath analyzer is associated with many limitations such as small patient groups, metabolic variability, no determined correlation between acetone and blood glucose level, low concentrations, and a large number of different VOCs per breath. To prepare a versatile breath analysis system, it may be necessary to develop a calibration procedure for the patient. The use of a matrix of various sensors, selective for various compounds (especially acetone), and then the use of artificial intelligence can be helpful in determining the correlation of sensor responses with the patient's disease state. Before developing the classification algorithms, it is necessary to perform extraction and selection of features. It is important to select hyperparameters and apply regularization and validation in order to avoid underfitting and overfitting of the algorithm. In addition to traditional machine learning algorithms, the use of a novel XGBoost algorithm or CNNs to increase performance of breath analysis system for diabetes detection are worth to be considered.

Data availability statement

No new data were created or analysed in this study.

Funding

This work was supported by the National Centre of Research and Development, Grant 'Rzeczy są dla ludzi/0089/2020'.

ORCID iDs

Anna Paleczek https://orcid.org/0000-0002-1467-3017
Artur Rydosz https://orcid.org/0000-0002-9148-1094

References

[1] Phillips M 1992 Breath tests in medicine *Sci. Am.* **267** 74–79

[2] Tortora G J and Derrickson B H 2018 *Principles of Anatomy and Physiology* (New York: Wiley)

[3] Smolinska A, Klaassen E M M, Dallinga J W, van de Kant K D G, Jobsis Q, Moonen E J C, van Schayck O C P, Dompeling E and van Schooten F J 2014 Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children *PLoS One* **9** e95668

[4] Phillips M, Herrera J, Krishnan S, Zain M, Greenberg J and Cataneo R N 1999 Variation in volatile organic compounds in the breath of normal humans *J. Chromatogr.* B **729** 75–88

[5] Barash O, Zhang W, Halpern J M, Hua Q-L, Pan Y-Y, Kayal H, Khoury K, Liu H, Davies M P A and Haick H 2015 Differentiation between genetic mutations of breast cancer by breath volatolomics *Oncotarget* **6** 44864

[6] Popov T A 2011 Human exhaled breath analysis *Ann. Allergy Asthma Immunol.* **106** 451–6; quiz 457

[7] Davis C, Frank M, Mizaikoff B and Oser H 2010 The future of sensors and instrumentation for human breath analysis *IEEE Sens. J.* **10** 3–6

[8] Selvaraj R, Vasa N J, Nagendra S M S and Mizaikoff B 2020 Advances in mid-infrared spectroscopy-based sensing techniques for exhaled breath diagnostics *Molecules* **25** 9

[9] Ma W, Liu X and Pawliszyn J 2006 Analysis of human breath with micro extraction techniques and continuous monitoring of carbon dioxide concentration *Anal. Bioanal. Chem.* **385** 1398–408

[10] Capone S, Tufariello M, Forleo A, Longo V, Giampetruzzi L, Radogna A V, Casino F and Siciliano P 2018 Chromatographic analysis of VOC patterns in exhaled breath from smokers and nonsmokers *Biomed. Chromatogr.* **32** e4132

[11] Longo V, Forleo A, Ferramosca A, Notari T, Pappalardo S, Siciliano P, Capone S and Montano L 2021 Blood, urine and semen volatile organic compound (VOC) pattern analysis for assessing health environmental impact in highly polluted areas in Italy *Environ. Pollut.* **286** 117410

[12] Gaude E, Nakhleh M K, Patassini S, Boschmans J, Allsworth M, Boyle B and van der Schee M P 2019 Targeted breath analysis: exogenous volatile organic compounds (EVOC) as metabolic pathway-specific probes *J. Breath Res.* **13** 032001

[13] Binson V A, Subramoniam M and Mathew L 2021 Noninvasive detection of COPD and lung cancer through breath analysis using MOS sensor array based e-nose *Expert Rev. Mol. Diagn.* **21** 1223–33

[14] Kharitonov S A and Barnes P J 2006 Exhaled biomarkers *Chest* **130** 1541–6

[15] Harkins M S, Fiato K-L and Iwamoto G K 2004 Exhaled nitric oxide predicts asthma exacerbation *J. Asthma* **41** 471–6

[16] Ratiu I A, Ligor T, Bocos-Bintintan V, Mayhew C A and Buszewski B 2021 Volatile organic compounds in exhaled breath as fingerprints of lung cancer, asthma and COPD *J. Clin. Med.* **10** 1

[17] Tenero L, Sandri M, Piazza M, Paiola G, Zaffanello M and Piacentini G 2020 Electronic nose in discrimination of children with uncontrolled asthma *J. Breath Res.* **14** 046003

[18] Sagita N *et al* 2021 Detection of asthma and chronic obstructive pulmonary disease (COPD) with an electronic nose (E-Nose) instrumentation system *2nd Int. Conf. on Science, Technology, and Modern Society (ICSTMS 2020)* pp 127–31

[19] Sakumura Y, Koyama Y, Tokutake H, Hida T, Sato K, Itoh T, Akamatsu T and Shin W 2017 Diagnosis by volatile organic compounds in exhaled breath from lung cancer patients using support vector machine algorithm *Sensors* **17** 2

[20] Dent A G, Sutedja T G and Zimmerman P V 2013 Exhaled breath analysis for lung cancer *J. Thorac. Dis.* **5** S540–50 (available at: https://jtd.amegroups.com/article/view/1560)

[21] Li J, Peng Y and Duan Y 2013 Diagnosis of breast cancer based on breath analysis: an emerging method *Crit. Rev. Oncol. Hematol.* **87** 28–40

[22] Herman-Saffar O, Boger Z, Libson S, Lieberman D, Gonen R and Zeiri Y 2018 Early non-invasive detection of breast cancer using exhaled breath and urine analysis *Comput. Biol. Med.* **96** 227–32

[23] Yang H-Y, Wang Y-C, Peng H-Y and Huang C-H 2021 Breath biopsy of breast cancer using sensor array signals and machine learning analysis *Sci. Rep.* **11** 103

[24] Binson V A, Subramoniam M and Mathew L 2021 Discrimination of COPD and lung cancer from controls through breath analysis using a self-developed e-nose *J. Breath Res.* **15** 046003

[25] Bouchikhi B, Zaim O, El Bari N, Lagdali N, Benelbarhdadi I and Ajana F Z 2021 Diagnosing lung and gastric cancers through exhaled breath analysis by using electronic nose technology combined with pattern recognition methods *IEEE Sens.* 1–4

[26] Liu L, Li W, He Z, Chen W, Liu H, Chen K and Pi X 2021 Detection of lung cancer with electronic nose using a novel ensemble learning framework *J. Breath Res.* **15** 026014

[27] Christiansen A, Davidsen J R, Titlestad I, Vestbo J and Baumbach J 2016 A systematic review of breath analysis and detection of volatile organic compounds in COPD *J. Breath Res.* **10** 034002

[28] Bregy L, Nussbaumer-Ochsner Y, Martinez-Lozano Sinues P, García-Gómez D, Suter Y, Gaisl T, Stebler N, Gaugg M T, Kohler M and Zenobi R 2018 Real-time mass spectrometric identification of metabolites characteristic of chronic obstructive pulmonary disease in exhaled breath *Clin. Mass Spectrom.* **7** 29–35

[29] Wang Z and Wang C 2013 Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements *J. Breath Res.* **7** 037109

[30] Minh T D C, Blake D R and Galassetti P R 2012 The clinical potential of exhaled breath analysis for diabetes mellitus *Diabetes Res. Clin. Pract.* **97** 195–205

[31] Deng C, Zhang J, Yu X, Zhang W and Zhang X 2004 Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization *J. Chromatogr.* B **810** 269–75

[32] Nelson N, Lagesson V, Nosratabadi A R, Ludvigsson J and Tagesson C 1998 Exhaled isoprene and acetone in newborn infants and in children with diabetes mellitus *Pediatr. Res.* **44** 363–7

[33] Španěl P, Dryahina K and Smith D 2007 Acetone, ammonia and hydrogen cyanide in exhaled breath of several volunteers aged 4–83 years *J. Breath Res.* **1** 011001

[34] Ghimenti S, Tabucchi S, Lomonaco T, Di Francesco F, Fuoco R, Onor M, Lenzi S and Trivella M G 2013 Monitoring breath during oral glucose tolerance tests *J. Breath Res.* **7** 017115

[35] Ueta I, Saito Y, Hosoe M, Okamoto M, Ohkita H, Shirai S, Tamura H and Jinno K 2009 Breath acetone analysis with miniaturized sample preparation device: in-needle preconcentration and subsequent determination by gas chromatography–mass spectroscopy *J. Chromatogr.* B **877** 2551–6

[36] Rydosz A 2018 Sensors for enhanced detection of acetone as a potential tool for noninvasive diabetes monitoring *Sensors* **18** E2298

[37] Sun M, Chen Z, Gong Z, Zhao X, Jiang C, Yuan Y, Wang Z, Li Y and Wang C 2015 Determination of breath acetone in 149 type 2 diabetic patients using a ringdown breath-acetone analyzer *Anal. Bioanal. Chem.* **407** 1641–50

[38] Rydosz A, Maziarz W, Pisarkiewicz T, de Torres H B and Mueller J 2013 A micropreconcentrator design using low temperature cofired ceramics technology for acetone detection applications *IEEE Sens. J.* **13** 1889–96

[39] Rydosz A 2014 Micropreconcentrator in LTCC technology with mass spectrometry for the detection of acetone in healthy and type-1 diabetes mellitus patient breath *Metabolites* **4** 921–31

[40] Paleczek A, Grochala D and Rydosz A 2021 Artificial breath classification using XGBoost algorithm for diabetes detection *Sensors* **21** 12

[41] Buszewski B, Kesy M, Ligor T and Amann A 2007 Human exhaled air analytics: biomarkers of diseases *Biomed. Chromatogr.* **21** 553–66

[42] Neupane S, Peverall R, Richmond G, Blaikie T P J, Taylor D, Hancock G and Evans M L 2016 Exhaled breath isoprene rises during hypoglycemia in type 1 diabetes *Diabetes Care* **39** e97–e98

[43] Melo R E, Popov T A and Solé D 2010 Exhaled breath temperature, a new biomarker in asthma control: a pilot study *J. Bras. Pneumol.* **36** 693–9

[44] Smith D, Španěl P, Gilchrist F J and Lenney W 2013 Hydrogen cyanide, a volatile biomarker of *Pseudomonas aeruginosa* infection *J. Breath Res.* **7** 044001

[45] Gilchrist F J, Razavi C, Webb A K, Jones A M, Španěl P, Smith D and Lenney W 2012 An investigation of suitable bag materials for the collection and storage of breath samples containing hydrogen cyanide *J. Breath Res.* **6** 036004

[46] Grabowska-Polanowska B, Faber J, Skowron M, Miarka P, Pietrzycka A, Śliwka I and Amann A 2013 Detection of potential chronic kidney disease markers in breath using gas chromatography with mass-spectral detection coupled with thermal desorption method *J. Chromatogr.* A **1301** 179–89

[47] Haines A, Metz G, Dilawari J, Blendis L and Wiggins H 1977 Breath-methane in patients with cancer of the large bowel *Lancet* **2** 481–3

[48] Sivertsen S M, Bjørneklett A, Gullestad H P and Nygaard K 1992 Breath methane and colorectal cancer *Scand. J. Gastroenterol.* **27** 25–28

[49] Weitz Z, Birnbaum A, Skosey J, Sobotka P and Zarling E 1991 High breath pentane concentrations during acute myocardial infarction *Lancet* **337** 933–5

[50] Mendis S, Sobotka P A and Euler D E 1995 Expired hydrocarbons in patients with acute myocardial infarction *Free Radic. Res.* **23** 117–22

[51] Olopade C O, Christon J A, Zakkar M, Swedler W I, Rubinstein I, Hua C-W and Scheff P A 1997 Exhaled pentane and nitric oxide levels in patients with obstructive sleep apnea *Chest* **111** 1500–4

[52] Davies S, Spanel P and Smith D 1997 Quantitative analysis of ammonia on the breath of patients in end-stage renal failure *Kidney Int.* **52** 223–8

[53] Popa C, Dutu D C A, Cernat R, Matei C, Bratu A M, Banita S and Dumitras D C 2011 Ethylene and ammonia traces measurements from the patients' breath with renal failure via LPAS method *Appl. Phys.* B **105** 669–74

[54] World Health Organization 2019 *Global Report on Diabetes* (available at: www.who.int/westernpacific/health-topics/diabetes) (Accessed 22 October 2021)

[55] American Diabetes Association 2002 Screening for diabetes *Diabetes Care* **25** s21–s24

[56] International Diabetes Federation 2021 IDF news (available at: www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html) (Accessed 14 November 2021)

[57] Guo D, Zhang D, Li N, Zhang L and Yang J 2010 A novel breath analysis system based on electronic olfaction *IEEE Trans. Biomed. Eng.* **57** 2753–63

[58] Yan K and Zhang D 2012 A novel breath analysis system for diabetes diagnosis *2012 Int. Conf. on Computerized Healthcare (ICCH)* pp 166–70

[59] Yan K, Zhang D, Wu D, Wei H and Lu G 2014 Design of a breath analysis system for diabetes screening and blood glucose level prediction *IEEE Trans. Biomed. Eng.* **61** 2787–95

[60] Saidi T, Zaim O, Moufid M, El Bari N, Ionescu R and Bouchikhi B 2018 Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects *Sens. Actuators* B **257** 178–88

[61] Lekha S and Suchetha M 2015 Non-invasive diabetes detection and classification using breath analysis *2015 Int. Conf. on Communications and Signal Processing (ICCSP)* pp 0955–8

[62] Lekha S and Suchetha M 2017 A novel 1-D convolution neural network with SVM architecture for real-time detection applications *IEEE Sens. J.* **18** 724–31

[63] Lekha S and Suchetha M 2017 Real-time non-invasive detection and classification of diabetes using modified convolution neural network *IEEE J. Biomed. Health Inform.* **22** 1630–6

[64] Sarno R, Sabilla S I and Wijaya D R 2020 Electronic nose for detecting multilevel diabetes using optimized deep neural network *Eng. Lett.* **28** 31–42

[65] Sarno R *et al* 2017 Detection of diabetes from gas analysis of human breath using e-nose *2017 11th Int. Conf. on Information & Communication Technology and System (ICTS)* pp 241–6

[66] Kalidoss R, Umapathy S, Kothalam R and Sakthivelu U 2020 Adsorption kinetics feature extraction from breathprint obtained by graphene based sensors for diabetes diagnosis *J. Breath Res.* **15** 016005

[67] Siegel A P, Daneshkhah A, Hardin D S, Shrestha S, Varahramyan K and Agarwal M 2017 Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: towards developing an alternative to diabetes alert dogs *J. Breath Res.* **11** 026007

[68] Yu J-B, Byun H-G, So M-S and Huh J-S 2005 Analysis of diabetic patient's breath with conducting polymer sensor array *Sens. Actuators* B **108** 305–8

[69] Jalal A H, Umasankar Y, Chowdhury M and Bhansali S 2017 A fuel cell sensing platform for selective detection of acetone in hyperglycemic patients *Meet. Abstr.* **MA2017-02** 2130

[70] Gupta A, Singh T S and Yadava R D S 2018 MEMS sensor array-based electronic nose for breath analysis-a simulation study *J. Breath Res.* **13** 016003

[71] Yadav L and Manjhi J 2014 Non-Invasive biosensor for diabetes monitoring *Asian J. Pharm. Clin. Res.* **7** 206–11

[72] American Diabetes Association 2021 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2021 *Diabetes Care* **44** S15–S33

[73] Huang Y, Liu F, Chen A-M, Yang P-F, Peng Y, Gong J-P, Li Z and Zhong G-C 2021 Type 2 diabetes prevention diet and the risk of pancreatic cancer: a large prospective multicenter study *Clin. Nutr.* **40** 5595–604

[74] Ley S H, Hamdy O, Mohan V and Hu F B 2014 Prevention and management of type 2 diabetes: dietary components and nutritional strategies *Lancet* **383** 1999–2007

[75] Sperling M A, Wolfsdorf J I, Menon R K, Tamborlane W V, Maahs D, Battelino T and Phillip M 2021 Diabetes mellitus *Sperl. Pediatric Endocrinol.* 814–83

[76] Barr A J 2018 The biochemical basis of disease *Essays Biochem.* **62** 619–42

[77] Carrasco-Sánchez F J, Fernández-Rodríguez J M, Ena J, Gómez-Huelgas R and Carretero-Gómez J 2021 Medical treatment of type 2 diabetes mellitus: recommendations of the diabetes, obesity and nutrition group of the spanish society of internal medicine *Revista Clínica Española (English Edition)* **221** 101–8

[78] Forouhi N G and Wareham N J 2010 Epidemiology of diabetes *Medicine* **38** 602–6

[79] Kreider K E 2019 The diagnosis and management of atypical types of diabetes *J. Nurse Pract.* **15** 171–6

[80] Yan Y, Wang Q, Li W, Zhao Z, Yuan X, Huang Y and Duan Y 2014 Discovery of potential biomarkers in exhaled breath for diagnosis of type 2 diabetes mellitus based on GC-MS with metabolomics *RSC Adv.* **4** 25430–9

[81] Trefz P, Obermeier J, Lehbrink R, Schubert J K, Miekisch W and Fischer D-C 2019 Exhaled volatile substances in children suffering from type 1 diabetes mellitus: results from a cross-sectional study *Sci. Rep.* **9** 1–9

[82] King J, Kupferthaler A, Unterkofler K, Koc H, Teschl S, Teschl G, Miekisch W, Schubert J, Hinterhuber H and Amann A 2009 Isoprene and acetone concentration profiles during exercise on an ergometer *J. Breath Res.* **3** 027006

[83] Jiang C, Sun M, Wang Z, Chen Z, Zhao X, Yuan Y, Li Y and Wang C 2016 A portable real-time ringdown breath acetone analyzer: toward potential diabetic screening and management *Sensors* **16** 1199

[84] Mathew T L, Pownraj P, Abdulla S and Pullithadathil B 2015 Technologies for clinical diagnosis using expired human breath analysis *Diagnostics* **5** 27–60

[85] Rooth G and Ostenson S 1966 Acetone in alveolar air, and the control of diabetes *Lancet* **2** 1102–5

[86] Miekisch W, Schubert J K and Noeldge-Schomburg G F 2004 Diagnostic potential of breath analysis—focus on volatile organic compounds *Clin. Chim. Acta* **347** 25–39

[87] Lebovitz H E 1995 Diabetic ketoacidosis *Lancet* **345** 767–72

[88] Ruzsányi V and Kalapos M P 2017 Breath acetone as a potential marker in clinical practice *J. Breath Res.* **11** 024002

[89] Saasa V, Beukes M, Lemmer Y and Mwakikunga B 2019 Blood ketone bodies and breath acetone analysis and their correlations in type 2 diabetes mellitus *Diagnostics* **9** 224

[90] Wang C, Mbi A and Shepherd M 2009 A study on breath acetone in diabetic patients using a cavity ringdown breath analyzer: exploring correlations of breath acetone with blood glucose and glycohemoglobin A1C *IEEE Sens. J.* **10** 54–63

[91] Schwarz K *et al* 2009 Breath acetone—aspects of normal physiology related to age and gender as determined in a PTR-MS study *J. Breath Res.* **3** 027003

[92] Teshima N, Li J, Toda K and Dasgupta P K 2005 Determination of acetone in breath *Anal. Chim. Acta* **535** 189–99

[93] Yan K and Zhang D 2014 Blood glucose prediction by breath analysis system with feature selection and model fusion *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2014** 6406–9

[94] Guo D, Zhang D, Li N, Zhang L and Yang J 2010 Diabetes identification and classification by means of a breath analysis system *Int. Conf. on Medical Biometrics* pp 52–63

[95] Beauchamp J, Herbig J, Gutmann R and Hansel A 2008 On the use of Tedlar® bags for breath-gas sampling and analysis *J. Breath Res.* **2** 046001

[96] Pleil J D and Lindstrom A B 1995 Collection of a single alveolar exhaled breath for volatile organic compounds analysis *Am. J. Ind. Med.* **28** 109–21

[97] Miekisch W and Schubert J K 2006 From highly sophisticated analytical techniques to life-saving diagnostics: technical developments in breath analysis *TrAC Trends Anal. Chem.* **25** 665–73

[98] Righettoni M, Schmid A, Amann A and Pratsinis S E 2013 Correlations between blood glucose and breath components from portable gas sensors and PTR-TOF-MS *J. Breath Res.* **7** 037110

[99] Mansour E, Vishinkin R, Rihet S, Saliba W, Fish F, Sarfati P and Haick H 2020 Measurement of temperature and relative humidity in exhaled breath *Sens. Actuators* B **304** 127371

[100] Ferrus L, Guenard H, Vardon G and Varene P 1980 Respiratory water loss *Respir. Physiol.* **39** 367–81

[101] Tyagi H, Daulton E, Bannaga A S, Arasaradnam R P and Covington J A 2021 Electronic nose for bladder cancer detection *Chem. Proc.* **5** 1

[102] Wang S, He Y, Zhang J, Chen S and He B 2020 Comparison of taste and odour characteristics of three mass-produced aquaculture clams in China *Aquac. Res.* **51** 664–73

[103] Thomas J N, Roopkumar J and Patel T 2021 Machine learning analysis of volatolomic profiles in breath can identify non-invasive biomarkers of liver disease: a pilot study *PLoS One* **16** e0260098

[104] Akturk H K, Snell-Bergeon J, Pyle L, Fivekiller E, Garg S and Cobry E 2021 Accuracy of a breath ketone analyzer to detect ketosis in adults and children with type 1 diabetes *J. Diabetes Complicat.* **35** 108030

[105] Saslow L R, Moskowitz J T, Mason A E, Daubenmier J, Liestenfeltz B, Missel A L, Bayandorian H, Aikens J E, Kim S and Hecht F M 2020 Intervention enhancement strategies among adults with type 2 diabetes in a very low–carbohydrate web-based program: evaluating the impact with a randomized trial *JMIR Diabetes* **5** e15835

[106] Iii D J S, Ratto T V, Ratto M and McCarter J P 2020 Characterization of a high-resolution breath acetone meter for ketosis monitoring *Peer J.* **8** e9969

[107] Dixit K, Fardindoost S, Ravishankara A, Tasnim N and Hoorfar M 2021 Exhaled breath analysis for diabetes diagnosis and monitoring: relevance, challenges and possibilities *Biosensors* **11** 12

[108] Hosfield B D, Drucker N A, Pecoraro A R, Shelley W C, Li H, Baxter N T, Hawkins T B and Markel T A 2021 The assessment of microbiome changes and fecal volatile organic compounds during experimental necrotizing enterocolitis *J. Pediatr. Surg.* **56** 1220–5

[109] de León-martínez L D, Rodríguez-Aguilar M, Gorocica-Rosete P, Domínguez-Reyes C A, Martínez-Bustos V, Tenorio-Torres J A, Ornelas-Rebolledo O, Cruz-Ramos J A, Balderas-Segura B and Flores-Ramírez R 2020 Identification of profiles of volatile organic compounds in exhaled breath by means of an electronic nose as a proposal for a screening method for breast cancer: a case-control study *J. Breath Res.* **14** 046009

[110] Falkenhain K, Locke S R, Lowe D A, Reitsma N J, Lee T, Singer J, Weiss E J and Little J P 2021 Keyto app and device versus WW app on weight loss and metabolic risk in adults with overweight or obesity: a randomized trial *Obesity* **29** 1606–14

[111] Boubin M and Shrestha S 2019 Microcontroller implementation of support vector machine for detecting blood glucose levels using breath volatile organic compounds *Sensors* **19** 10

[112] Rydosz A 2020 Chapter 28—Nanosensors for exhaled breath monitoring as a possible tool for noninvasive diabetes detection *Nanosensors for Smart Cities* ed B Han, V K Tomer, T A Nguyen, A Farmani and P Kumar Singh (Amsterdam: Elsevier) pp 467–81

[113] Tiele A, Wicaksono A, Ayyala S K and Covington J A 2020 Development of a compact, IoT-enabled electronic nose for breath analysis *Electronics* **9** 1

[114] Tricoli A, Righettoni M and Pratsinis S E 2009 Minimal cross-sensitivity to humidity during ethanol detection by SnO_2–TiO_2 solid solutions *Nanotechnology* **20** 315502

[115] Tellez-Cruz M M, Escorihuela J, Solorza-Feria O and Compañ V 2021 Proton exchange membrane fuel cells (PEMFCs): advances and challenges *Polymers* **13** 3064

[116] Guruprasad B and Shwetha M S 2020 Design and fabrication of cantilever MEMS sensor model for electro-chemical gas sensor *Int. J. Eng. Res. Technol.* **9** 704–15

[117] Hübschmann H-J 2015 *Handbook of GC-MS: Fundamentals and Applications* (New York: Wiley)

[118] Zhang D, Guo D and Yan K 2017 Breath signal analysis for diabetics *Breath Analysis for Medical Applications* (Berlin: Springer) pp 241–58

[119] Paulsson N, Larsson E and Winquist F 2000 Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose *Sens. Actuators* A **84** 187–97

[120] Marzorati D, Mainardi L, Sedda G, Gasparri R, Spaggiari L and Cerveri P 2021 MOS sensors array for the discrimination of lung cancer and at-risk subjects with exhaled breath analysis *Chemosensors* **9** 8

[121] Weber P, Pauling J K, List M and Baumbach J 2020 BALSAM—an interactive online platform for breath analysis, visualization and classification *Metabolites* **10** 10

[122] Géron A 2019 *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (United States, California, Sebastopol: O'Reilly Media)

[123] Jolliffe I T and Cadima J 2016 Principal component analysis: a review and recent developments *Phil. Trans. R. Soc.* A **374** 20150202

[124] Lever J, Krzywinski M and Altman N 2017 Points of significance: principal component analysis *Nat. Methods* **14** 641–3

[125] Maimon O and Rokach L 2005 *Data Mining and Knowledge Discovery Handbook* (Boston, MA: Springer)

[126] Burges C J 1998 A tutorial on support vector machines for pattern recognition *Data Min. Knowl. Discov.* **2** 121–67

[127] Cherkassky V and Ma Y 2004 Practical selection of SVM parameters and noise estimation for SVM regression *Neural Netw.* **17** 113–26

[128] Soumaya Z, Taoufiq B D, Benayad N, Yunus K and Abdelkrim A 2021 The detection of Parkinson disease using the genetic algorithm and SVM classifier *Appl. Acoust.* **171** 107528

[129] NirmalaDevi M, Alias Balamurugan S A and Swathi U 2013 An amalgam KNN to predict diabetes mellitus *2013 IEEE Int. Conf. on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)* pp 691–5

[130] Patikar S, Saha P, Neogy S and Chowdhury C 2020 An approach towards prediction of diabetes using modified Fuzzy *K* nearest neighbor *2020 IEEE Int. Conf. on Computing, Power and Communication Technologies (GUCON)* pp 73–76

[131] Guo G, Wang H, Bell D, Bi Y and Greer K 2003 *KNN Model-Based Approach in Classification* vol 2888 (Berlin, Heidelberg: Springer) p 996

[132] Shafi S and Ansari G 2021 Early prediction of diabetes disease & classification of algorithms using machine learning approach *SSRN Electron. J.* 1–10

[133] Gou J, Ma H, Ou W, Zeng S, Rao Y and Yang H 2019 A generalized mean distance-based *k*-nearest neighbor classifier *Expert Syst. Appl.* **115** 356–72

[134] Gou J, Qiu W, Yi Z, Shen X, Zhan Y and Ou W 2019 Locality constrained representation-based *K*-nearest neighbor classification *Knowl. Based Syst.* **167** 38–52

[135] Arowolo M O, Adebiyi M O, Adebiyi A A and Olugbara O 2021 Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier *J. Big Data* **8** 29

[136] Blatt R, Bonarini A and Matteucci M 2010 Pattern classification techniques for lung cancer diagnosis by an electronic nose *Computational Intelligence in Healthcare 4* (Berlin: Springer) pp 397–423

[137] Wang Y, Wu D and Yuan X 2020 LDA-based deep transfer learning for fault diagnosis in industrial chemical processes *Comput. Chem. Eng.* **140** 106964

[138] Huang H, Lin D, Chen W, Yu Y, Xu J, Liang Z, Lin X, Dong Z and Shi H 2014 Nondestructive discrimination between normal and hematological malignancy cell lines using near-infrared Raman spectroscopy and multivariate analysis *Laser Phys. Lett.* **11** 085601

[139] Bishop C M 2006 Pattern recognition *Mach. Learn.* **128** 179–220

[140] Chen T and Guestrin C XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (*New York, NY, USA*) pp 785–94

[141] Torlay L, Perrone-Bertolotti M, Thomas E and Baciu M 2017 Machine learning–XGBoost analysis of language networks to classify patients with epilepsy *Brain Inform.* **4** 159–69

[142] Ogunleye A and Wang Q-G 2020 XGBoost model for chronic kidney disease diagnosis *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17** 2131–40

[143] Ogunleye A and Wang Q-G 2018 Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease *2018 IEEE 14th Int. Conf. on Control and Automation (ICCA)* pp 805–10

[144] Gupta N *et al* 2013 Artificial neural network *Netw. Complex Syst.* **3** 24–28

[145] Nunes I and Da Silva H S 2018 *Artificial Neural Networks: A Practical Course* (Berlin: Springer) (https://doi.org/10.1007/978-3-319-43162-8)

[146] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)

[147] Phillips M, Cataneo R N, Cheema T and Greenberg J 2004 Increased breath biomarkers of oxidative stress in diabetes mellitus *Clin. Chim. Acta* **344** 189–94

[148] Španěl P, Dryahina K, Rejšková A, Chippendale T W and Smith D 2011 Breath acetone concentration; biological variability and the influence of diet *Physiol. Meas.* **32** N23

[149] Turner C 2011 Potential of breath and skin analysis for monitoring blood glucose concentration in diabetes *Expert Rev. Mol. Diagn.* **11** 497–503

[150] Ahamed F and Farid F 2018 Applying internet of things and machine-learning for personalized healthcare: issues and challenges *2018 Int. Conf. on Machine Learning and Data Engineering (iCMLDE)* pp 19–21

[151] Shailaja K, Seetharamulu B and Jabbar M A 2018 Machine learning in healthcare: a review *2018 s Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA)* pp 910–4

3. The experimental results obtained during the laboratory studies

This chapter presents the results of laboratory studies using an electronic nose system developed by the Author and the Biomarkers Analysis LAB AGH Research group to detect biomarkers of metabolic diseases in exhaled air. The first phase of the study was focused on artificially prepared gas mixtures, which provided a controlled and repeatable experimental environment. The composition of the mixtures was based on a literature review to closely mimic human exhaled air. This approach enabled the evaluation of the measurement system's performance under laboratory conditions, free from variables resulting from environmental influences, dietary intake, or individual variability. Particular attention was paid to acetone (C_3H_6O, CAS No. 67-64-1), considered one of the best-documented biomarkers present in exhaled air associated with carbohydrate metabolism disorders. Its concentration in exhaled air increases in diabetes and prediabetes, as well as in conditions of increased ketogenesis. Due to its direct relationship with glucose metabolism, acetone is considered a key diagnostic indicator, useful for non-invasive diagnostics and monitoring the progression of diabetes.

Analysis of data from the e-nose system requires the use of machine learning algorithms because the responses of gas sensors are multidimensional and nonlinear signals, which are difficult to interpret using classical methods. Depending on the experimental goal, regression algorithms were used to predict the concentrations of individual biomarkers, while classification algorithms were employed to distinguish between samples representing different health states. The results obtained in the laboratory studies formed the basis for further clinical studies on exhaled air samples, described in the next chapter.

This chapter summarises three research papers documenting the successive stages of laboratory research. Each chapter focuses on a different aspect of the e-nose system's use in detecting metabolic disorders.

The first research paper [AP1] examined the potential of using the developed e-nose system to detect diabetes based on the concentration of acetone in prepared gas mixtures. The experiments were conducted using samples with different concentrations of acetone (C_3H_6O, CAS 67-64-1), ethanol (C_2H_6O, CAS 64-17-5), propane (C_3H_8, CAS 74-98-6), ethylbenzene (C_8H_{10}, CAS 100-41-4), carbon dioxide (CO_2, CAS 124-38-9) and relative humidity (RH) corresponding to levels observed in the breath of healthy individuals and diabetic patients. The study tested several machine learning algorithms for classifying samples. The results showed that the e-nose system, supported by XGBoost, is capable of classifying mixtures simulating the exhaled air of healthy individuals and those with diabetes. The main challenge was to differentiate acetone levels in the mixtures into two classes: concentrations less than 1.5 ppm and concentrations greater than or equal to 1.5 ppm. This threshold was selected based on literature research, and the mixtures also contained other components found in exhaled air. The effect of gas mixture humidity on sensor response was also examined, as this is a crucial factor given the high humidity of exhaled air. As part of the work, for the XGBoost algorithm, an accuracy of 99%, a recall of 100% and a specificity of 97.9% were obtained for the classification of gas mixtures based on acetone concentration. This work was the first test of the developed e-nose and examined whether and which machine learning algorithms could be used to classify gas mixtures based on acetone levels.

The second research paper [AP2] focused on the problem of interference resulting from the presence of interfering compounds, particularly ethanol. Ethanol is a common component of human breath - it can come from both endogenous and external sources (e.g., alcohol consumption, use of oral hygiene fluids or eating certain meals) - and can significantly affect the accuracy of acetone measurements. Typical acetone concentrations in healthy individuals range from 0.3 to 1.5 ppm, and in diabetics, they can reach several ppm or higher. In contrast, ethanol concentrations can exceed hundreds of ppm or even higher. Gas sensors, especially MOS sensors, are often poorly selective, and ethanol has similar adsorption and reactive properties on the sensor surface as acetone, making it difficult to detect very low acetone concentrations when ethanol is also present in the mixtures. In this study, gas mixtures containing various proportions of acetone and ethanol were prepared, and then the possibility of predicting acetone concentration was analysed using regression algorithms. The results showed that appropriately selected machine learning methods can effectively compensate the effect of ethanol presence in the mixtures, maintaining high accuracy in predicting acetone concentrations. The study achieved a mean absolute error (MAE) of 0.245 ppm for diabetic breath acetone levels using four sensors and XGBoost. In mixtures with high ethanol content and 0-8.62 ppm of acetone, CatBoost performed best with an error of 0.568 ppm of acetone concentration prediction. This research paper had high practical significance, as it confirmed that the e-nose system can also be helpful in more complex and realistic conditions, where the presence of interfering compounds is unavoidable and should be considered when designing e-nose systems.

The third research conference paper [AP3] expanded the scope of the analyses, focusing on the classification of gas samples into three categories corresponding to different health states: healthy individuals, pre-diabetics, and diabetics. For this purpose, gas mixtures were prepared to reflect the typical biomarker profiles characteristic of each group, and the composition of the mixtures was selected based on literature research. Classification algorithms in Python were employed for the analysis, enabling effective differentiation between the studied classes with high accuracy. The highest health prediction accuracy scores were obtained using CatBoost and were 95%, 79% and 88% for healthy, prediabetes, diabetes classes, respectively. The results confirmed that the e-nose system has the potential not only to detect the disease itself but also to identify intermediate states, including those differing by small acetone concentrations, which is particularly important for prevention and early diagnosis. The results were presented by the Author at the *"8th International Conference on Bio-Sensing Technology"* – 12-15 May 2024, Seville, Spain.

The presented research papers [AP1, AP2, AP3] document a coherent sequence of laboratory studies on the e-nose system developed by the Author and the LAB Research group, which can be generalised to other e-nose devices for the diagnosis of metabolic diseases. The research began by detecting a single biomarker, then proceeded to investigate how different interferences affected the e-nose and how well machine learning could handle them. Finally, the system was tested in multiclass classification. Altogether, the findings suggest that the system could be developed into a useful diagnostic tool. The obtained results enabled the refinement of both measurement methods and machine learning algorithms, which were subsequently applied in clinical trials. These provide the foundation for further real human samples analysis presented in the next chapter.

3.1. Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection

sensors

MDPI

Article

Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection

Anna Paleczek *, Dominik Grochala and Artur Rydosz

Institute of Electronics, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland; grochala@agh.edu.pl (D.G.); rydosz@agh.edu.pl (A.R.)

* Correspondence: paleczek@student.agh.edu.pl

Abstract: Exhaled breath analysis has become more and more popular as a supplementary tool for medical diagnosis. However, the number of variables that have to be taken into account forces researchers to develop novel algorithms for proper data interpretation. This paper presents a system for analyzing exhaled air with the use of various sensors. Breath simulations with acetone as a diabetes biomarker were performed using the proposed e-nose system. The XGBoost algorithm for diabetes detection based on artificial breath analysis is presented. The results have shown that the designed system based on the XGBoost algorithm is highly selective for acetone, even at low concentrations. Moreover, in comparison with other commonly used algorithms, it was shown that XGBoost exhibits the highest performance and recall.

Keywords: breath acetone; diabetes; XGBoost; VOCs; machine learning; algorithms; e-nose

check for updates

Citation: Paleczek, A.; Grochala, D.; Rydosz, A. Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection. *Sensors* **2021**, *21*, 4187. https://doi.org/10.3390/s21124187

Academic Editor: Simon X. Yang

Received: 6 June 2021
Accepted: 17 June 2021
Published: 18 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Nowadays, groups of researchers are focused on non-invasive methods for diagnosing various diseases. One of the promising tools is exhaled breath analysis. Its potential in medical diagnosis has been known since the time of Hippocrates when he used the smell of the breath to diagnose liver disease and uncontrolled diabetes [1].

The air inhaled and exhaled by humans consists mainly of nitrogen, oxygen and carbon dioxide (Figure 1). Exhaled air contains more carbon dioxide and less oxygen than inhaled air because oxygen is used to generate energy during respiration, while carbon dioxide is produced as a by-product of the energy production process. Among the major components, exhaled breath consists of over 3500 Volatile Organic Compounds (VOCs) and a single breath consists of around 500 various VOCs, which are typically in the part per million (ppm), part per billion (ppb) or part per trillion (ppt) range [2]. Some of them are named biomarkers since their presence, as well as various concentration levels, may indicate several diseases. Biomarkers are compounds present in the body that can be used as indicators of physiology and diseases present. These types of VOCs are called endogenous VOCs and are produced by the metabolism of cells. On the other hand, the second type of VOCs are exogenous VOCs used to assess the effects of substances such as drugs, diet, cigarettes, toxic or noxious vapors and environmental pollution on the body. Exogenous VOCs are present in, for example, breath or blood as a result of circulation and/or internal metabolism [3–5]. Clear separation of biomarkers into these two groups is not possible because the same VOCs can be induced physiologically in the body as a result of disease, and also under the influence of external factors [4,5]. A general approach to determining biomarkers for a given pathological condition is to compare the VOC composition of a group of healthy and sick people [3]. There are several types of biomarkers: monitoring, predictive, prognostic, safety and susceptibility/risk biomarkers [6]. Systemic biomarkers are used to determine the functioning of the whole organism, while lung biomarkers are used to determine the processes and changes taking place in the respiratory system [7]. Currently, research is focused on biomarkers of various diseases, for example asthma [8,9], various types of cancers [10–13], chronic obstructive pulmonary disease [14,15] and, recently, metabolic disorders, such as diabetes [7,16–24], which will allow non-invasive detection and monitoring of these diseases using exhaled air. However, diet and pathological changes may affect the exhaled breath compositions; therefore, every person has their own unique molecular breath signature [7,25]. Similarly to a fingerprint, the exhaled profile is called the "breath-fingerprint" or "personal breath profile". Common biomarkers of several diseases are listed in Table 1.

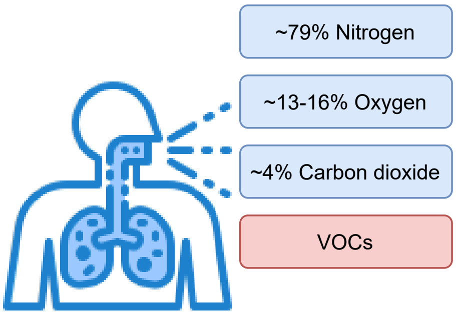


Figure 1. General composition of humans' exhaled breath.

Table 1. Potential disease biomarkers in the breath.

Disease	Biomarkers	References
Diabetes	Acetone	[2,7,16–24,26]
Asthma	Nitric Oxide	[2,8,9]
Cystic fibrosis	Hydrogen cyanide	[27,28]
Lung cancer	VOC pattern	[10,11,26]
Chronic kidney disease	Trimethylamine	[29]
Colorectal cancer	Methane	[30,31]
Myocardial infarction	Pentane	[32,33]
Obstructive sleep apnea	Pentane and Nitric Oxide	[34]
Renal failure	Ammonia	[35,36]

Usually, the biomarker concentrations are too low to be detected without the utilization of advanced analytical systems such as GC/MS (Gas Chromatograph coupled to a Mass Spectrometer) [37,38], SIFT–MS (Selected Ion Flow Tube–Mass Spectrometry) [39,40], PTR–MS (Proton Transfer Reaction–Mass Spectrometry) [41]. One of the promising techniques to increase the volume of biomarkers is the utilization of preconcentrators, including micropreconcentrators [22,42,43].

One disease prevalent in civilization that requires constant monitoring is diabetes. Briefly, there are two main types of diabetes: type 1 (T1DM) and type 2 (T2DM); T2DM is the most common (90% of all cases). According to data provided by the World Health Organization (WHO), approximately 500 million people worldwide have diabetes, and this number is constantly growing. The vast majority of them live in low- and middle-income countries. The WHO also reports 1.6 million deaths annually from diabetes [44]. Diabetes over time damages the nervous system, blood vessels and heart, as well as the eyes and kidneys, leading to an increased risk of premature death [45]. Due to the ever-increasing number of people with diabetes and deaths from it, the WHO reports that there is a globally agreed goal to halt the development of diabetes and obesity by 2025 [44]. At present, there are no known methods of preventing type 1 diabetes. Its treatment consists of continuous monitoring of blood glucose level (BGL) and the patient's insulin intake. However, in the case of type 2 diabetes, it is possible to reduce its incidence by adhering to a proper diet, increasing physical activity, and reducing smoking. In addition to diet and exercise, early diagnosis plays an important role in the treatment of diabetes, so it is important to develop an easily accessible and non-invasive device that can be used for screening [44–46]. In terms of exhaled breath analysis, acetone was identified as a biomarker of diabetes [7,16–24,47]. Results presented in Table 2 show that breath acetone concentrations for healthy peoples were lower than for diabetes patients.

Table 2. Acetone concentration in health and diabetes samples.

Diabetic Stage	Measured Acetone Concentration	References
T2DM	1.76–3.73 ppm	[18]
Healthy	0.22–0.80 ppm	
Controlled diabetic	0.19–0.66 ppmv	[22]
Untreated T2DM	0.92–1.20 ppmv	
Diabetes	1.25–2.5 ppm (or up to 25 ppm)	[23]
Healthy	0.2–1.8 ppm	
T1DM	4.9 ± 16 ppm	[47]
T2DM	1.5 ± 1.3 ppm	
Healthy	1.1 ± 0.5 ppm	
Diabetes	>1.8 ppmv	[48]
Healthy	<0.8 ppmv	
T1DM	2.19 ppmv (mean)	[49]
Healthy	0.48 ppmv (mean)	
Healthy	0.177–2.441 ppm	[50]
Healthy	0.176–0.518 ppm	[51]

Experimental results have shown that relative humidity (RH) and temperature of exhaled human breath vary between subjects. Mansour et al. examined Parisian and Halifa participants. The measured values were 31.4–35.4 °C and 65.0–88.6% for Halifa participants and 31.4–34.8 °C and 41.9–91.0% for Parisian participants [52]. Ferrus et al. showed that the RH in exhaled air from humans varies between 89 and 97% [53]. Due to the high relative humidity of the breath and its influence on the sensitivity of the measurement systems (especially metal oxide semiconductor sensors) [54–56], it is necessary to use moisture absorbers to properly store the breath samples and to take into account the influence of humidity on the measurements in designed algorithms.

The researchers present the results of using various supervised machine learning and deep learning algorithms to classify breath samples and detect diabetes. The most popular are K Nearest Neighbours (KNN) [57–60], Support Vector Machines (SVM) [37,59,61–63], Naive Bayes (NB) [59,64], Deep Neural Network (DNN) [59] and also Convolutional Neural Networks (CNN) [65]. The extraction and selection of features was most often performed using Principal Component Analysis [57,59,61,66]. The main limitation of the conducted research is the lack of an adequate number of patient samples. Only a small fraction of the research has been carried out on sample numbers above a hundred [57,58,61].

In this paper, the experimental results on the e-nose system for discrimination between healthy and diabetic patients based on the exhaled breath analysis are presented. Within this study, an artificial breath profile was developed to simulate real conditions and enable testing without involving real samples.

2. Materials and Methods

The scheme of the system proposed in this paper is presented in Figure 2.

All algorithms were developed using scikit-learn Machine Learning in Python [67,68] and XGBoost, an open-source software library that provides a gradient boosting framework for C++, Java, Python, R, Julia, Perl, and Scala [69].

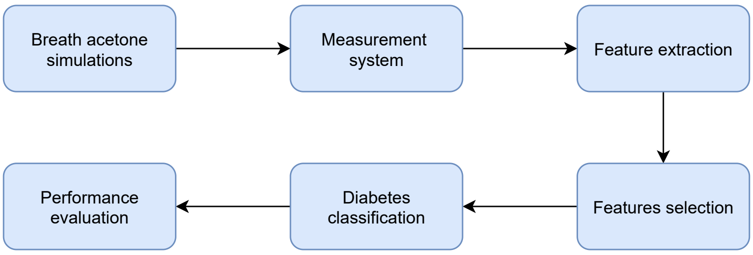


Figure 2. Block scheme of the proposed system.

2.1. Equipment

Selected gas sensors (listed in Table 3) were placed in a measurement chamber with a 180 mL capacity and supplied with appropriate voltages in accordance with their data sheets. Due to the relative humidity influence on sensors' sensitivity, in addition to gas sensors, temperature, relative humidity and pressure sensors were also used. The BME280 (Bosch Sensortec, Reutlingen, Germany) and SHT85 (Sensirion, Staefa ZH, Switzerland) sensors were placed inside the measurement chamber, while the second SHT85 sensor was placed before the gases entered the measurement chamber. All used sensors, except SGP30 and SHT85, responded to the dosed gases as voltage. For SGP30, the sensor returned Total Volatile Organic Compounds (TVOCs) and an equivalent carbon dioxide reading (eCO2) over the I2C communication bus. TGS1820 (Figaro Engineering Inc, Mino, Osaka, Japan), TGS2620 (Figaro Engineering Inc, Mino, Osaka, Japan), TGS8100 (Figaro Engineering Inc, Mino, Osaka, Japan), MQ3 (Waveshare, Shenzhen, China) and MICS5524 (Amphenol SGX Sensortech, Corcelles-Cormondreche, Switzerland) sensors' responses were measured using Keithley 617 (Tektronix, Beaverton, OR, USA), Keithley 6514 (Tektronix, Beaverton, United States) and multimeter Keysight 34450A electrometers (Keysight, Santa Rosa, CA, USA). If the sensor sent the measured values using the Serial Peripheral Interface (SPI) or Inter-Integrated Circuit (I2C) communication bus, the ESP32 dev board (Espressif Systems, Shanghai, China) was used to read these values and send them to the measurement application written in the Python programming language. Figure 3 shows a scheme of the proposed e-nose measurement system. The glass flask shown in Figure 3 was used to simulate the humidity.

2.2. Exhaled Breath Simulations

The gas mixtures composed of synthetic air, acetone, ethanol, propane and ethylbenzene were dosed with a variable relative humidity to simulate exhaled air using the GF40 series (Brooks, Hatfield, United States) mass flow controllers with a Brooks 0254 controller. Due to the high humidity of the exhaled air, the measurements simulated humidity ranging from 0 to 70%. However, the relative humidity measured inside the chamber was 0 to 40% due to the increased temperature in the measurement chamber. Taking into account the number of all possible combinations of gas mixtures, the total duration of measurements was estimated to be more than 700 days. Thanks to the use of an artificial exhaled breath mixture, the experiments could be conducted constantly (24 h/7 d) without involving the diabetic patients. Since acetone is the key biomarker of diabetes, it was decided to measure the response to various concentrations of acetone contaminated with other gases in the concentration ranges that have been previously confirmed by the utilization of analytical techniques such as GC/MS [37,38]. Based on the obtained results presented in Table 2, the simulations assumed that the concentration of acetone in the exhaled air for a healthy person is <1.5 ppm and for a diabetic patient is $\geq$1.5 ppm.

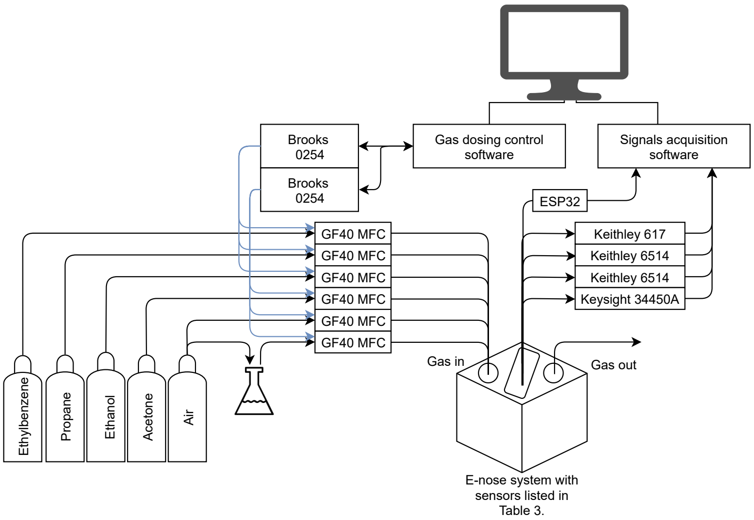


Figure 3. Scheme of the proposed measurement system.

Table 3. Sensors used in measurements.

Sensor	Target Gases	Typical Detection Range
TGS1820	$(CH_3)_2CO$	1–20 ppm $(CH_3)_2CO$
TGS2620	C_2H_5OH, Solvent apors	50–5000 ppm C_2H_5OH
TGS8100	Air contaminants (H_2, C_2H_5OH etc.)	1–30 ppm H_2
MICS5524	CO, VOCs	1–1000 ppm CO 10–500 ppm C_2H_5OH 1–1000 ppm H_2 1–500 ppm NH_3 >1000 ppm CH_4
MQ3	C_2H_5OH, CH_4, Benzine, Hexane, LPG, CO	0.04–4 mg/L C_2H_5OH
SGP30	CO_2, VOCs	0–1000 ppm H_2 0–1000 ppm C_2H_5OH 0–60,000 ppb eq tVOCs 400–60,000 ppm eq CO_2

2.3. Preprocessing

In order to obtain input data for the algorithms, preprocessing and features extraction were carried out. The use of baseline subtraction is important due to baseline drift. The result of the long-term stability test is given in Figure 4.

The baseline was fitted to the raw data obtained from the sensors and then subtracted (Figure 5).

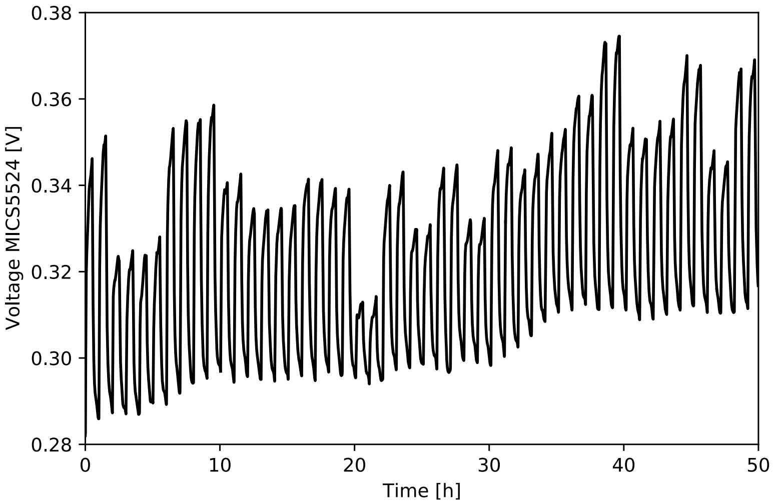


Figure 4. Result of the long-term test for different gas mixtures—MICS5524.

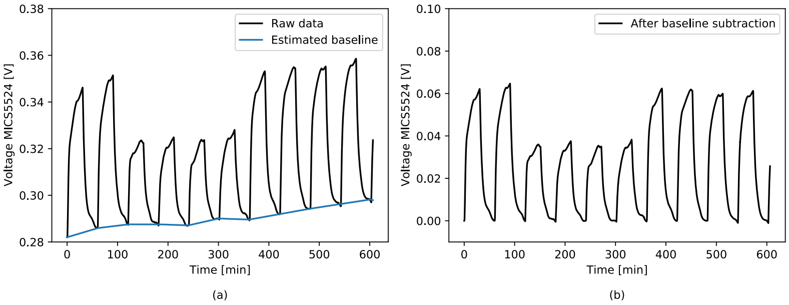


Figure 5. Baseline subtraction. (**a**) Sensor raw response with fitted baseline; (**b**) result of the baseline subtraction.

The following features have been extracted from each gas sensor:

- The sensor response (S) defined by Equation (1):

$$S = \frac{R_S}{R_0} \tag{1}$$

- The sensor response change (ΔS) defined by the Equation (2):

$$\Delta S = R_S - R_0 \tag{2}$$

where:
R_S—sensor exposed to target gas, e.g., acetone;
R_0—sensor exposed to pure synthetic air;

- Area under sensor's response curve (AUC) calculated when the sensor is exposed to gas. Result approximated by the trapezoidal numerical integration.

The prepared dataset from the simulation of acetone in the breath was divided into two separate sets—the training set and the test set. In order to simulate the real case, where samples from healthy subjects are overwhelmingly obtained [37,38,70,71], the simulations were conducted with an unbalanced number of samples. Moreover, not every algorithm, i.e., Support Vector Machines, K Nearest Neighbours [72–74], works well with an unbalanced dataset; therefore, such experiments are crucial. Due to the unbalanced number of samples belonging to the "healthy" and "diabetes" classes, the data were divided in such a way that the same percentage of samples from each class was included in both the test and training sets. Distribution of samples in the dataset are given in Figure 6.

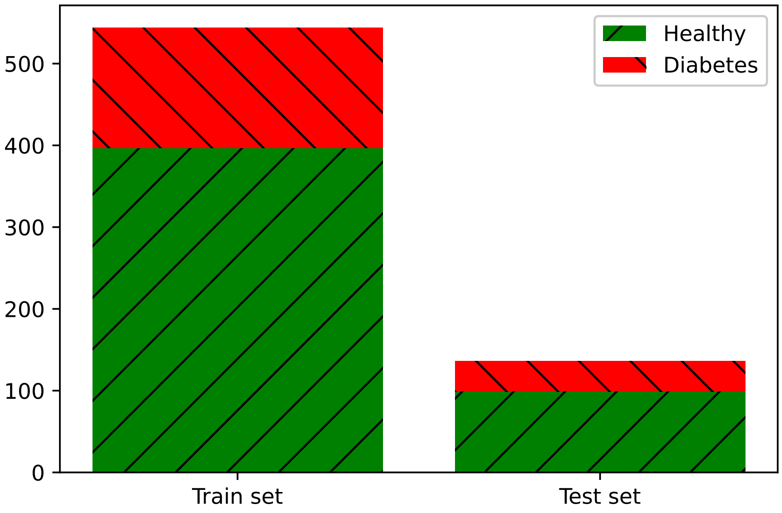


Figure 6. Dataset abundance and distribution.

2.4. Features Selection

Due to the correlation between the features extracted from the raw data from each sensor, we decided to use the calculated *S* results and the values read from the temperature and humidity sensors as an input to the algorithms. As detailed in Section 3.3, the gas sensors, except SGP30, used the *S* value that slightly changes with the change in humidity, which is important when measuring exhaled air, characterized by high humidity.

2.5. XGBoost Classifier

Recently, extreme gradient boosting (XGBoost) state-of-the-art algorithms are becoming more and more popular not only for classification, but also for regression problems, due to their high performance [69,75–77]. The XGBoost alghorithm is a scalable tree boosting system which was developed by Chen and Guestrin in 2016. Parallel, distributed, out-of-core and cache-aware computing makes the algorithm more than ten times faster than popular models used in machine learning (ML) and deep learning (DL). Another advantage of this algorithm is that it is well optimized and scalable. Due to this innovation, it can be successfully used to process billions of examples in distributed or memory-limited settings. This cutting-edge application of gradient boosting machines was designed to handle real-world problems where the input data sparsity is a common issue. The algorithm is aware of the presence of missing values, too frequent zero values in the dataset and results of applied feature engineering techniques. The ensemble technique is the recursive addition of new models until further addition no longer noticeably enhances the performance of existing models. The loss of the model is minimized by the gradient descent algorithm [69].

2.6. Hyperparameter Optimization

To determine the best performance, the model's hyperparemeters were optimized by a grid search algorithm. Model evaluation was performed using the stratified k-fold cross-validation method. It is commonly used to evaluate models with limited datasets. We decided to use a stratified version of this algorithm due to the unbalanced dataset; it splits the dataset, keeping the equal proportions of each output class in each fold. The use of this method enables the selection of optimal model hyperparameters and reduces overfitting of the data. The training set was divided into k sets, then the model was trained with the use of $k-1$ datasets, and the remaining set was used to validate the model using the selected metrics. The final value of a metric is the average of the k iteration [78,79].

2.7. Classifiers' Performance Evaluation Metrics

In this paper, we mainly focused on obtaining the highest possible sensitivity value (recall score) defined by Equation (3):

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

where:
TPR—true positive rate (recall, sensitivity);
TP—true positive;
FN—false negative [80].

This metric is especially important in medical applications, when the dataset is unbalanced, and we strive to minimize the type II error. For example, in the case of screening tests, it is important to mark all potentially sick patients and possibly, in further, more accurate, as well as invasive and more expensive tests, confirm or rule out diabetes.

3. Results and Discussion

3.1. Sensors' Sensitivity to Gases Used in Simulations

Figure 7 shows the responses of each sensor to different acetone concentrations. Each concentration was repeated at least twice in order to check the stability of the sensors and the repeatability of the response to individual gas concentrations. The results show that each of the sensors is sensitive to changes in acetone concentration, and in the case of the same concentration being used several times, the sensors are stable and the responses are repeatable.

3.2. Sensors' Selectivity to Acetone

The results of measurements of the sensor response to various gas mixtures with a constant concentration of acetone—1.5 ppm in each mixture, given in Figure 8—show that none of the sensors included in the designed e-nose system is fully acetone selective. Therefore, it is important to use a sensor array where each sensor is selective for different gases/gas mixtures.

3.3. Relative Humidity Dependency

Due to the high humidity of the breath, measurements were made at different simulated humidities. For each of the sensors used, the characteristics of the relative dependence of the sensor's response to humidity were determined and the dependence of the sensitivity to 1 ppm of acetone on the ambient humidity was also calculated. Results are given in Figure 9.

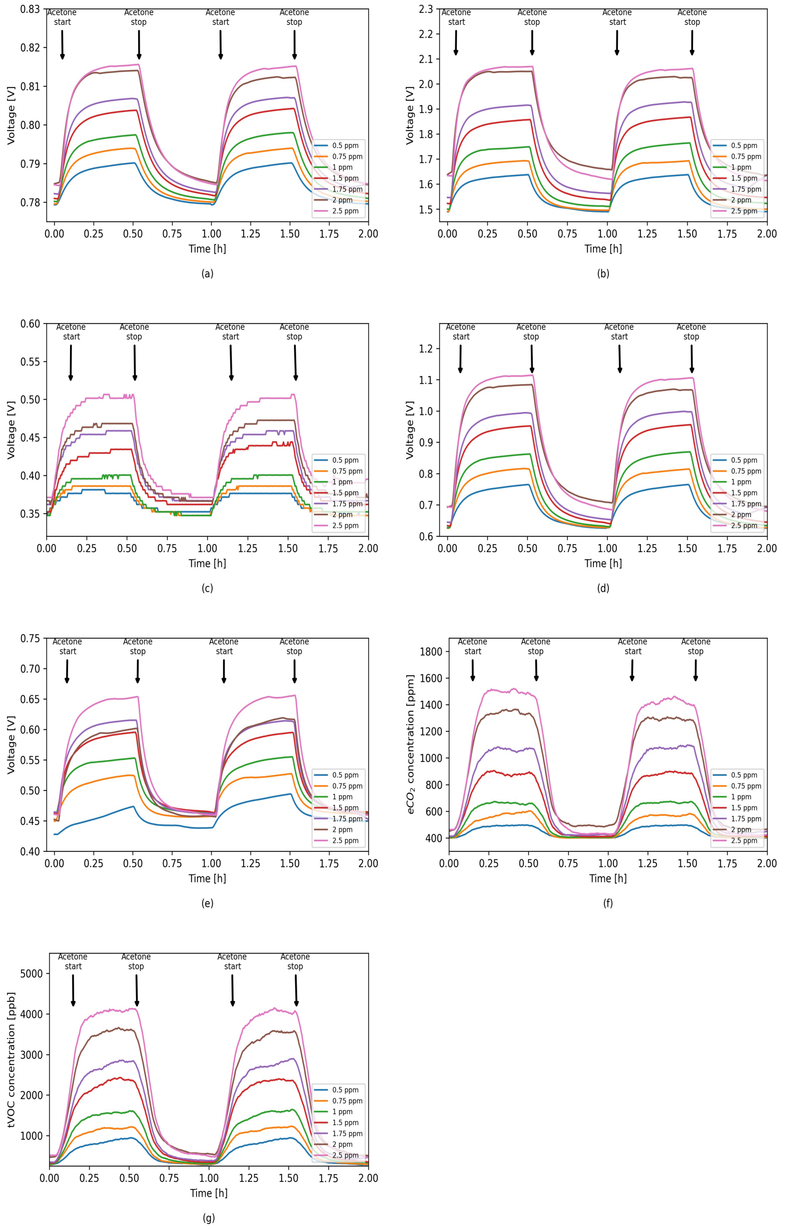


Figure 7. Sensors' responses to different acetone concentrations in 0% RH. (**a**) TGS1820; (**b**) TGS2620; (**c**) TGS8100; (**d**) MQ3; (**e**) MICS5524; (**f**) SGP30 eCO2; (**g**) SGP30 tVOC.

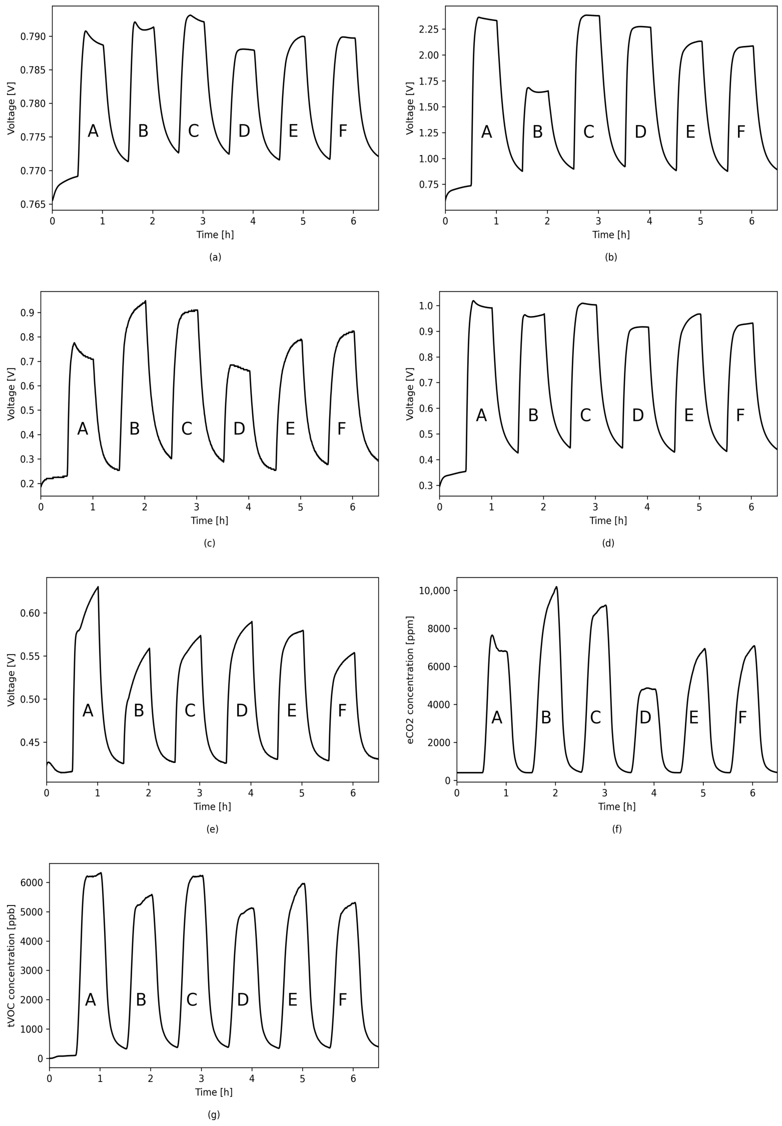


Figure 8. Sensors' responses to different simulated mixtures in 0% RH. A. 1.5 ppm acetone, 2.5 ppm ethanol, 1 ppm propane; B. 1.5 ppm acetone, 1 ppm ethanol, 2.5 ppm ethylbenzene; C. 1.5 ppm acetone, 1.5 ppm ethanol, 1 ppm ethylbenzene, 1 ppm propane; D. 1.5 ppm acetone, 1.5 ppm ethanol, 1 ppm propane; E. 1.5 ppm acetone, 1.5 ppm ethanol, 0.5 ppm ethylbenzene, 0.5 ppm propane; F. 1.5 ppm acetone, 1 ppm ethanol, 1 ppm ethylbenzene, 0.5 ppm propane; (**a**) TGS1820; (**b**) TGS2620; (**c**) TGS8100; (**d**) MQ3; (**e**) MICS5524; (**f**) SGP30 eCO2; (**g**) SGP30 tVOC.

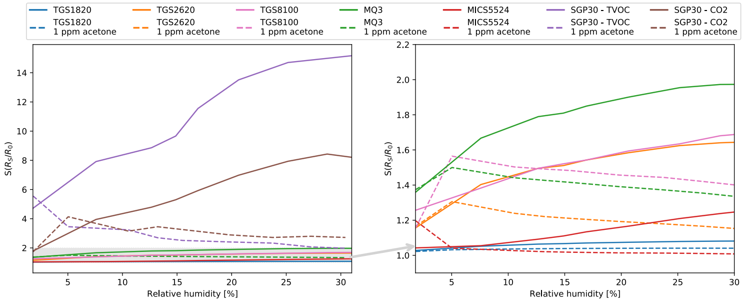


Figure 9. Sensors' sensitivity in different relative humidities in chamber.

3.4. Classification

The optimal model hyperparameters were determined using the grid search algorithm. In order to assess whether the model is underfitted or overfitted, validation was used with the use of a separate validation set. Learning curves showing the dependence of the classification error on the number of training epochs are shown in Figure 10.

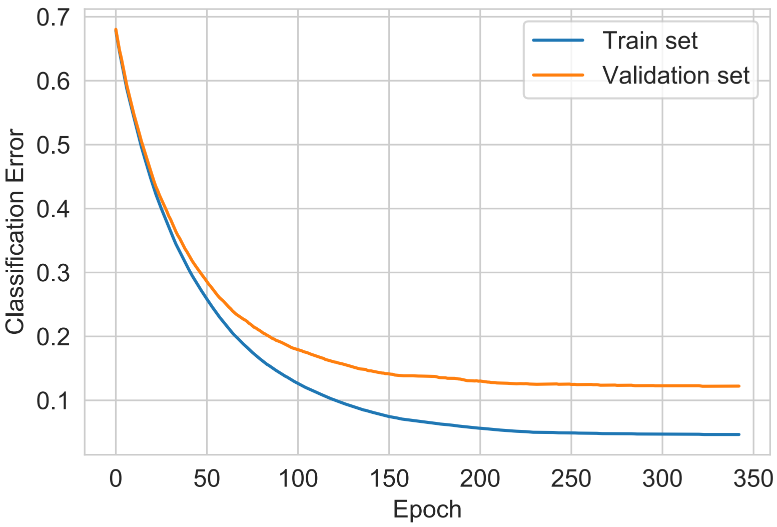


Figure 10. XGBoost learning curves.

3.5. Feature Importance

The results of the algorithm showed that the three most important features for the classification were measurements from the MQ3, TGS1820, SGP30 and SHT85 sensors placed inside the chamber. Feature importance values for the most significant sensors are given in Figure 11.

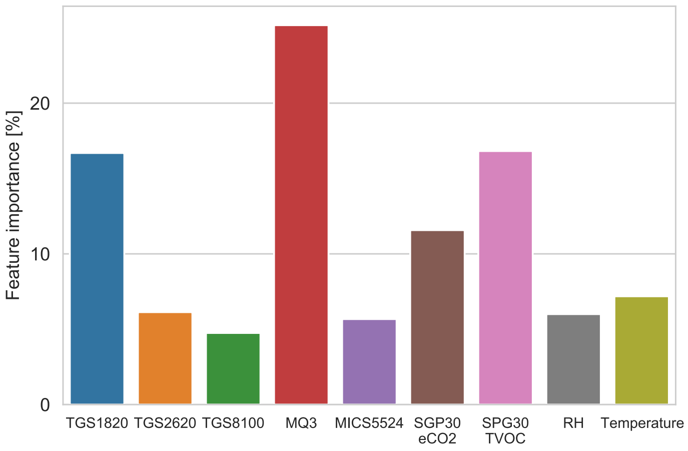


Figure 11. XGBoost features' importance.

3.6. Performance Evaluation

In the case of using the algorithm based on the gradient of boosted trees, the recall equals 1, which means that all the sick patients were correctly marked as sick and the type II error was minimized. The other calculated performance evaluation metrics are summarized in Table 4. As we assumed, the algorithm's hyperparameters were selected in such a way that it achieved the highest recall value.

Table 4. Classifier performance evaluation results.

Metric	Result
Accuracy	99%
Recall	100%
Specificity	97.9%
Area under ROC curve	97.9%
F1-score	97.4%

Confusion Matrix

The algorithm's confusion matrix is shown in Figure 12. It shows that the healthy diabetes samples were classified properly. The confusion matrix allows one to accurately quantify the true positive, true negative, false positive and false negative test samples. Based on these values, the remaining metrics are calculated. In the case of the proposed XGBoost Classifier algorithm, two cases of simulated diabetes patients were incorrectly classified. This is a type I statistical error.

3.7. Comparison with Classic Machine Learning Algorithms

In this paper, we also compared the classification performance achieved using the XGBoost algorithm with the results of classic classifiers such as Support Vector Machines (SVM), K Nearest Neightbour (KNN), Decision Tree Classifier (DT) and Random Forest Classifier (RF), commonly used in previous research. For these algorithms, the hyperparameters were also determined using the grid search method and the K-Fold validation was performed. The classification was carried out using the same train and test sets as for XGBoost.

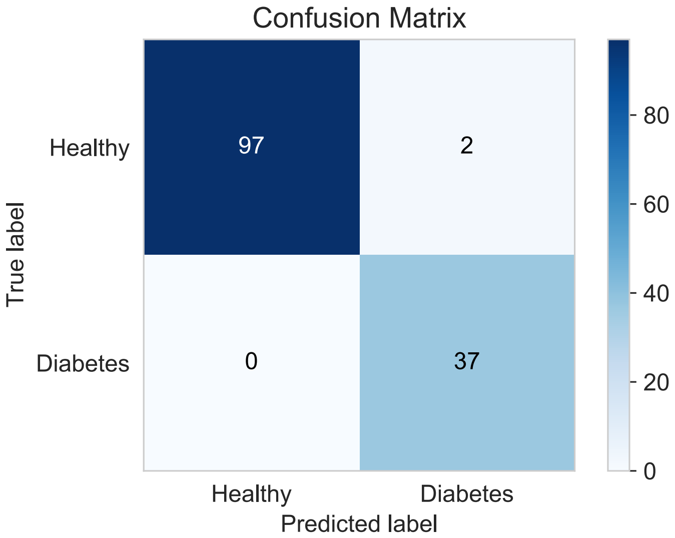


Figure 12. XGBoost Classifier confusion matrix.

Figure 13 shows a comparison of the achieved recall of the algorithms.

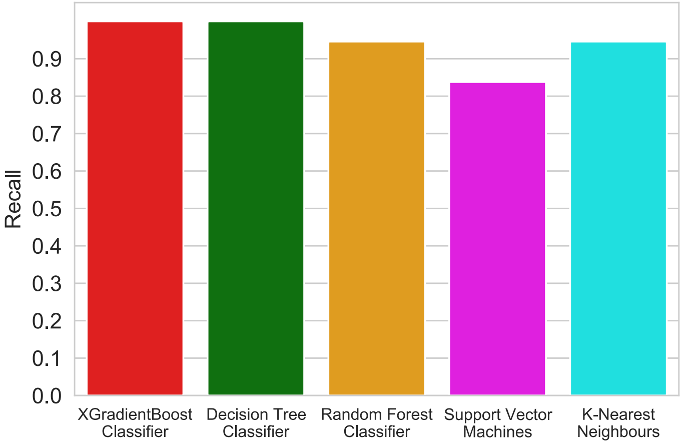


Figure 13. Recall comparison of different algorithms.

The receiver operating characteristics (ROC) curve shows the dependence between recall and 1-specificity. It is commonly used in machine learning tasks for medical applications. The closer the curve for a given model is to the point (0,1), the better the classifier. The most common problem in designing models for medical data is that the data contain more healthy cases than disease ones [81]. Figure 14 shows the ROC comparison for each of the algorithms used in this research.

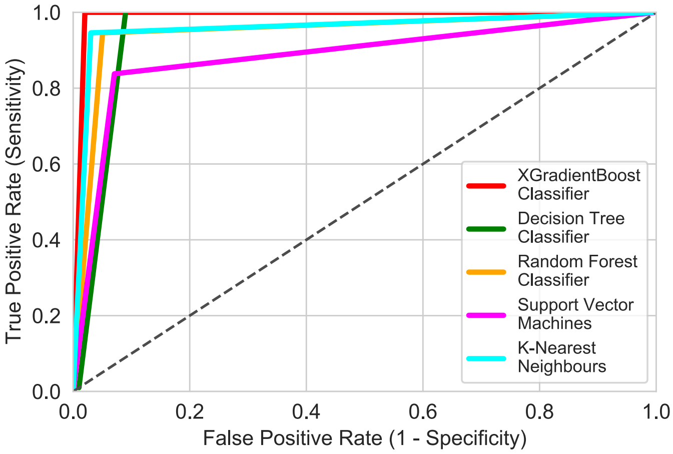


Figure 14. ROC comparison of different algorithms.

All of the used algorithms exhibited good performances. Each of these algorithms obtained recall and false positive rates of over 80%. By analyzing the determined metrics, it can be seen that the XGBoost Classifier has the highest accuracy and recall equal to 99 and 100%, respectively. Decision Tree Classifier obtained a recall identical to the XGBoost Classiffier, but the results differ in the amount of false positives. It is true that in screening tests, the most important detection is as many true positives as possible, but reducing the number of false positives, i.e., healthy ones classified as sick, reduces the cost of further diagnosis.

3.8. Discussion

Due to the individual variability shown in the literature, depending on, inter alia, sex, age, diet, duration of diabetes life, the course of treatment and its type, it is necessary to conduct tests on breath samples. It may also be necessary to develop a method for calibrating the device tailored to an individual patient. The results presented in this paper show that the designed system is highly selective for acetone, even at low concentrations. In order to confirm the selectivity of the system towards all breath components, it is necessary to carry out measurements on samples of exhaled air taken from healthy people and diabetics. The graphs of dependence of the sensor's response and sensitivity on the ambient humidity in the measurement chamber showed that the all sensors used, except SGP30, are slightly sensitive to humidity. Measurements of humidity in the chamber and taking these results into account in the input data to the algorithms made it possible to compensate for its influence. In the case of the presented sensors' system and the algorithm used, the classification of diabetics was independent of the relative humidity inside the measuring chamber. Comparison with other commonly used algorithms showed that XGBoost showed the highest performance and recall. One of the disadvantages of the system is the long response and retention time of each of the sensors used; therefore, in order to use such a system for medical applications, it is necessary to use a different sensor matrix, a preconcentrator, increase the total air flows in the chamber or reduce the volume of the measurement chamber.

4. Conclusions

Exhaled breath analysis consists of several steps including sample collection, compound detection, data analysis, and data interpretation. Each stage could be realized in various manners. So far, the researchers have made efforts to develop the compound detection units, for example, by the utilization of electronic noses, which offer cheap, fast, and reliable results. However, due to the number of compounds present in exhaled human breath as well as high humidity concentration, the detection unit has to be supported by an artificial intelligence element to deliver reliable results. In this paper, the XGBoost algorithm for diabetes detection based on the exhaled breath analysis is presented. The results have shown that the designed system based on the XGBoost algorithm was highly selective for acetone, even at low concentrations. Moreover, in comparison with other commonly used algorithms, it was shown that XGBoost exhibits the highest performance and recall, which makes it a first choice for data analysis in terms of diabetes detection.

Author Contributions: Conceptualization, A.P.; methodology, A.P.; software, A.P.; validation, A.P and D.G.; formal analysis, A.P., A.R. and D.G.; investigation, A.P. and D.G.; data curation, A.P.; writing—original draft preparation, A.P.; writing—review and editing, A.R.; visualization, A.P.; supervision, A.R.; project administration, A.R.; funding acquisition, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Centre, NCN Poland 2017/26/D/ST7/00355.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Phillips, M. Breath tests in medicine. *Sci. Am.* **1992**, *267*, 74–79. [CrossRef]
2. Selvaraj, R.; Vasa, N.J.; Nagendra, S.M.S.; Mizaikoff, B. Advances in Mid-Infrared Spectroscopy-Based Sensing Techniques for Exhaled Breath Diagnostics. *Molecules* **2020**, *25*, 2227. [CrossRef] [PubMed]
3. Gaude, E.; Nakhleh, M.K.; Patassini, S.; Boschmans, J.; Allsworth, M.; Boyle, B.; van der Schee, M.P. Targeted breath analysis: Exogenous volatile organic compounds (EVOC) as metabolic pathway-specific probes. *J. Breath Res.* **2019**, *13*, 032001. [CrossRef] [PubMed]
4. Longo, V.; Forleo, A.; Ferramosca, A.; Notari, T.; Pappalardo, S.; Siciliano, P.; Capone, S.; Montano, L. Blood, urine and semen Volatile Organic Compound (VOC) pattern analysis for assessing health environmental impact in highly polluted areas in Italy. *Environ. Pollut.* **2021**, 117410. [CrossRef] [PubMed]
5. Capone, S.; Tufariello, M.; Forleo, A.; Longo, V.; Giampetruzzi, L.; Radogna, A.V.; Casino, F.; Siciliano, P. Chromatographic analysis of VOC patterns in exhaled breath from smokers and nonsmokers. *Biomed. Chromatogr.* **2018**, *32*, e4132. [CrossRef]
6. Califf, R.M. Biomarker definitions and their applications. *Exp. Biol. Med.* **2018**, *243*, 213–221. [CrossRef] [PubMed]
7. Popov, T.A. Human exhaled breath analysis. *Ann. Allergy Asthma Immunol.* **2011**, *106*, 451–456. [CrossRef]
8. Melo, R.E.; Popov, T.A.; Solé, D. Exhaled breath temperature, a new biomarker in asthma control: A pilot study. *J. Bras. Pneumol.* **2010**, *36*, 693–699. [CrossRef]
9. Harkins, M.S.; Fiato, K.L.; Iwamoto, G.K. Exhaled nitric oxide predicts asthma exacerbation. *J. Asthma* **2004**, *41*, 471–476. [CrossRef]
10. Sakumura, Y.; Koyama, Y.; Tokutake, H.; Hida, T.; Sato, K.; Itoh, T.; Akamatsu, T.; Shin, W. Diagnosis by volatile organic compounds in exhaled breath from lung cancer patients using support vector machine algorithm. *Sensors* **2017**, *17*, 287. [CrossRef]
11. Dent, A.G.; Sutedja, T.G.; Zimmerman, P.V. Exhaled breath analysis for lung cancer. *J. Thorac. Dis.* **2013**, *5*, S540.
12. Herman-Saffar, O.; Boger, Z.; Libson, S.; Lieberman, D.; Gonen, R.; Zeiri, Y. Early non-invasive detection of breast cancer using exhaled breath and urine analysis. *Comput. Biol. Med.* **2018**, *96*, 227–232. [CrossRef]
13. Li, J.; Peng, Y.; Duan, Y. Diagnosis of breast cancer based on breath analysis: An emerging method. *Crit. Rev. Oncol.* **2013**, *87*, 28–40. [CrossRef]
14. Christiansen, A.; Davidsen, J.R.; Titlestad, I.; Vestbo, J.; Baumbach, J. A systematic review of breath analysis and detection of volatile organic compounds in COPD. *J. Breath Res.* **2016**, *10*, 034002. [CrossRef]
15. Bregy, L.; Nussbaumer-Ochsner, Y.; Sinues, P.M.L.; García-Gómez, D.; Suter, Y.; Gaisl, T.; Stebler, N.; Gaugg, M.T.; Kohler, M.; Zenobi, R. Real-time mass spectrometric identification of metabolites characteristic of chronic obstructive pulmonary disease in exhaled breath. *Clin. Mass Spectrom.* **2018**, *7*, 29–35. [CrossRef]
16. Wang, Z.; Wang, C. Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements. *J. Breath Res.* **2013**, *7*, 037109. [CrossRef]
17. Minh, T.D.C.; Blake, D.R.; Galassetti, P.R. The clinical potential of exhaled breath analysis for diabetes mellitus. *Diabetes Res. Clin. Pract.* **2012**, *97*, 195–205. [CrossRef]

18. Deng, C.; Zhang, J.; Yu, X.; Zhang, W.; Zhang, X. Determination of acetone in human breath by gas chromatography–mass spectrometry and solid-phase microextraction with on-fiber derivatization. *J. Chromatogr. B* **2004**, *810*, 269–275. [CrossRef]
19. Nelson, N.; Lagesson, V.; Nosratabadi, A.R.; Ludvigsson, J.; Tagesson, C. Exhaled isoprene and acetone in newborn infants and in children with diabetes mellitus. *Pediatr. Res.* **1998**, *44*, 363–367. [CrossRef]
20. Španěl, P.; Dryahina, K.; Smith, D. Acetone, ammonia and hydrogen cyanide in exhaled breath of several volunteers aged 4–83 years. *J. Breath Res.* **2007**, *1*, 011001. [CrossRef] [PubMed]
21. Ghimenti, S.; Tabucchi, S.; Lomonaco, T.; Francesco, F.D.; Fuoco, R.; Onor, M.; Lenzi, S.; Trivella, M.G. Monitoring breath during oral glucose tolerance tests. *J. Breath Res.* **2013**, *7*, 017115. [CrossRef]
22. Ueta, I.; Saito, Y.; Hosoe, M.; Okamoto, M.; Ohkita, H.; Shirai, S.; Tamura, H.; Jinno, K. Breath acetone analysis with miniaturized sample preparation device: In-needle preconcentration and subsequent determination by gas chromatography–mass spectroscopy. *J. Chromatogr. B* **2009**, *877*, 2551–2556. [CrossRef]
23. Rydosz, A. Sensors for enhanced detection of acetone as a potential tool for noninvasive diabetes monitoring. *Sensors* **2018**, *18*, 2298. [CrossRef]
24. Sun, M.; Chen, Z.; Gong, Z.; Zhao, X.; Jiang, C.; Yuan, Y.; Wang, Z.; Li, Y.; Wang, C. Determination of breath acetone in 149 Type 2 diabetic patients using a ringdown breath-acetone analyzer. *Anal. Bioanal. Chem.* **2015**, *407*, 1641–1650. [CrossRef] [PubMed]
25. Davis, C.E.; Frank, M.; Mizaikoff, B.; Oser, H. The future of sensors and instrumentation for human breath analysis. *IEEE Sens. J.* **2010**, *10*, 3–6. [CrossRef]
26. Buszewski, B.; Kęsy, M.; Ligor, T.; Amann, A. Human exhaled air analytics: Biomarkers of diseases. *Biomed. Chromatogr.* **2007**, *21*, 553–566. [CrossRef]
27. Smith, D.; Španěl, P.; Gilchrist, F.J.; Lenney, W. Hydrogen cyanide, a volatile biomarker of Pseudomonas aeruginosa infection. *J. Breath Res.* **2013**, *7*, 044001. [CrossRef] [PubMed]
28. Gilchrist, F.J.; Razavi, C.; Webb, A.K.; Jones, A.M.; Španěl, P.; Smith, D.; Lenney, W. An investigation of suitable bag materials for the collection and storage of breath samples containing hydrogen cyanide. *J. Breath Res.* **2012**, *6*, 036004. [CrossRef] [PubMed]
29. Grabowska-Polanowska, B.; Faber, J.; Skowron, M.; Miarka, P.; Pietrzycka, A.; Śliwka, I.; Amann, A. Detection of potential chronic kidney disease markers in breath using gas chromatography with mass-spectral detection coupled with thermal desorption method. *J. Chromatogr. A* **2013**, *1301*, 179–189. [CrossRef]
30. Haines, A.; Dilawari, J.; Metz, G.; Blendis, L.; Wiggins, H. Breath-methane in patients with cancer of the large bowel. *Lancet* **1977**, *310*, 481–483. [CrossRef]
31. Sivertsen, S.; Bjørneklett, A.; Gullestad, H.; Nygaard, K. Breath methane and colorectal cancer. *Scand. J. Gastroenterol.* **1992**, *27*, 25–28. [CrossRef]
32. Weitz, Z.; Birnbaum, A.; Skosey, J.; Sobotka, P.; Zarling, E. High breath pentane concentrations during acute myocardial infarction. *Lancet* **1991**, *337*, 933–935. [CrossRef]
33. Mendis, S.; Sobotka, P.A.; Euler, D.E. Expired hydrocarbons in patients with acute myocardial infarction. *Free Radic. Res.* **1995**, *23*, 117–122. [CrossRef] [PubMed]
34. Olopade, C.O.; Christon, J.A.; Zakkar, M.; Swedler, W.I.; Rubinstein, I.; Hua, C.w.; Scheff, P.A. Exhaled pentane and nitric oxide levels in patients with obstructive sleep apnea. *Chest* **1997**, *111*, 1500–1504. [CrossRef] [PubMed]
35. Davies, S.; Spanel, P.; Smith, D. Quantitative analysis of ammonia on the breath of patients in end-stage renal failure. *Kidney Int.* **1997**, *52*, 223–228. [CrossRef] [PubMed]
36. Popa, C.; Dutu, D.; Cernat, R.; Matei, C.; Bratu, A.; Banita, S.; Dumitras, D.C. Ethylene and ammonia traces measurements from the patients' breath with renal failure via LPAS method. *Appl. Phys. B* **2011**, *105*, 669–674. [CrossRef]
37. Saidi, T.; Zaim, O.; Moufid, M.; El Bari, N.; Ionescu, R.; Bouchikhi, B. Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects. *Sens. Actuators B Chem.* **2018**, *257*, 178–188. [CrossRef]
38. Siegel, A.P.; Daneshkhah, A.; Hardin, D.S.; Shrestha, S.; Varahramyan, K.; Agarwal, M. Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: Towards developing an alternative to diabetes alert dogs. *J. Breath Res.* **2017**, *11*, 026007. [CrossRef]
39. Storer, M.; Dummer, J.; Lunt, H.; Scotter, J.; McCartin, F.; Cook, J.; Swanney, M.; Kendall, D.; Logan, F.; Epton, M. Measurement of breath acetone concentrations by selected ion flow tube mass spectrometry in type 2 diabetes. *J. Breath Res.* **2011**, *5*, 046011. [CrossRef] [PubMed]
40. Dummer, J.F.; Storer, M.K.; Hu, W.P.; Swanney, M.P.; Milne, G.J.; Frampton, C.M.; Scotter, J.M.; Prisk, G.K.; Epton, M.J. Accurate, reproducible measurement of acetone concentration in breath using selected ion flow tube-mass spectrometry. *J. Breath Res.* **2010**, *4*, 046001. [CrossRef]
41. Thekedar, B.; Szymczak, W.; Höllriegl, V.; Hoeschen, C.; Oeh, U. Investigations on the variability of breath gas sampling using PTR-MS. *J. Breath Res.* **2009**, *3*, 027007. [CrossRef]
42. Rydosz, A.; Maziarz, W.; Pisarkiewicz, T.; de Torres, H.B.; Mueller, J. A micropreconcentrator design using low temperature cofired ceramics technology for acetone detection applications. *IEEE Sens. J.* **2013**, *13*, 1889–1896. [CrossRef]
43. Rydosz, A. Micropreconcentrator in LTCC technology with mass spectrometry for the detection of acetone in healthy and type-1 diabetes mellitus patient breath. *Metabolites* **2014**, *4*, 921–931. [CrossRef] [PubMed]
44. WHO. *Global Report on Diabetes (2019)*; WHO: Geneva, Switzerland, 2019.

45. Ley, S.H.; Hamdy, O.; Mohan, V.; Hu, F.B. Prevention and management of type 2 diabetes: Dietary components and nutritional strategies. *Lancet* **2014**, *383*, 1999–2007. [CrossRef]
46. Hegde, H.; Shimpi, N.; Panny, A.; Glurich, I.; Christie, P.; Acharya, A. Development of non-invasive diabetes risk prediction models as decision support tools designed for application in the dental clinical environment. *Inform. Med. Unlocked* **2019**, *17*, 100254. [CrossRef] [PubMed]
47. Jiang, C.; Sun, M.; Wang, Z.; Chen, Z.; Zhao, X.; Yuan, Y.; Li, Y.; Wang, C. A portable real-time ringdown breath acetone analyzer: Toward potential diabetic screening and management. *Sensors* **2016**, *16*, 1199. [CrossRef] [PubMed]
48. Saasa, V.; Beukes, M.; Lemmer, Y.; Mwakikunga, B. Blood ketone bodies and breath acetone analysis and their correlations in type 2 diabetes mellitus. *Diagnostics* **2019**, *9*, 224. [CrossRef]
49. Wang, C.; Mbi, A.; Shepherd, M. A study on breath acetone in diabetic patients using a cavity ringdown breath analyzer: Exploring correlations of breath acetone with blood glucose and glycohemoglobin A1C. *IEEE Sens. J.* **2009**, *10*, 54–63. [CrossRef]
50. Schwarz, K.; Pizzini, A.; Arendacka, B.; Zerlauth, K.; Filipiak, W.; Schmid, A.; Dzien, A.; Neuner, S.; Lechleitner, M.; Scholl-Bürgi, S.; et al. Breath acetone—aspects of normal physiology related to age and gender as determined in a PTR-MS study. *J. Breath Res.* **2009**, *3*, 027003. [CrossRef]
51. Teshima, N.; Li, J.; Toda, K.; Dasgupta, P.K. Determination of acetone in breath. *Anal. Chim. Acta* **2005**, *535*, 189–199. [CrossRef]
52. Mansour, E.; Vishinkin, R.; Rihet, S.; Saliba, W.; Fish, F.; Sarfati, P.; Haick, H. Measurement of temperature and relative humidity in exhaled breath. *Sens. Actuators B Chem.* **2020**, *304*, 127371. [CrossRef]
53. Ferrus, L.; Guenard, H.; Vardon, G.; Varene, P. Respiratory water loss. *Respir. Physiol.* **1980**, *39*, 367–381. [CrossRef]
54. Beauchamp, J.; Herbig, J.; Gutmann, R.; Hansel, A. On the use of Tedlar® bags for breath-gas sampling and analysis. *J. Breath Res.* **2008**, *2*, 046001. [CrossRef] [PubMed]
55. Tricoli, A.; Righettoni, M.; Pratsinis, S.E. Minimal cross-sensitivity to humidity during ethanol detection by SnO2–TiO2 solid solutions. *Nanotechnology* **2009**, *20*, 315502. [CrossRef] [PubMed]
56. Lekha, S.; Suchetha, M. Real-time non-invasive detection and classification of diabetes using modified convolution neural network. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 1630–1636. [CrossRef] [PubMed]
57. Guo, D.; Zhang, D.; Li, N.; Zhang, L.; Yang, J. A novel breath analysis system based on electronic olfaction. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 2753–2763.
58. Yan, K.; Zhang, D. A novel breath analysis system for diabetes diagnosis. In Proceedings of the 2012 International Conference on Computerized Healthcare (ICCH), Hong Kong, China, 17–18 December 2012; pp. 166–170.
59. Sarno, R.; Sabilla, S.I.; Wijaya, D.R. Electronic Nose for Detecting Multilevel Diabetes using Optimized Deep Neural Network. *Eng. Lett.* **2020**, *28*.
60. Hariyanto; Sarno, R.; Wijaya, D.R. Detection of diabetes from gas analysis of human breath using e-Nose. In Proceedings of the 2017 11th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 31 October 2017; pp. 241–246.
61. Yan, K.; Zhang, D.; Wu, D.; Wei, H.; Lu, G. Design of a breath analysis system for diabetes screening and blood glucose level prediction. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 2787–2795. [CrossRef]
62. Guo, D.; Zhang, D.; Li, N.; Zhang, L.; Yang, J. Diabetes identification and classification by means of a breath analysis system. In *Proceedings of the International Conference on Medical Biometrics, Hong Kong, China, 28–30 June 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 52–63.
63. Lekha, S.; Suchetha, M. Non-invasive diabetes detection and classification using breath analysis. In Proceedings of the 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, India, 2–4 April 2015; pp. 0955–0958.
64. Kalidoss, R.; Umapathy, S.; Kothalam, R.; Sakthivelu, U. Adsorption kinetics feature extraction from breathprint obtained by graphene based sensors for diabetes diagnosis. *J. Breath Res.* **2020**, *15*, 016005. [CrossRef]
65. Lekha, S.; Suchetha, M. A novel 1-D convolution neural network with SVM architecture for real-time detection applications. *IEEE Sens. J.* **2017**, *18*, 724–731. [CrossRef]
66. Zhang, D.; Guo, D.; Yan, K. Breath Signal Analysis for Diabetics. In *Breath Analysis for Medical Applications*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 241–258.
67. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*; Springer: Prague, Czech Republic, 2013; pp. 108–122.
68. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
69. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; KDD ’16; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
70. Fan, G.T.; Yang, C.L.; Lin, C.H.; Chen, C.C.; Shih, C.H. Applications of Hadamard transform-gas chromatography/mass spectrometry to the detection of acetone in healthy human and diabetes mellitus patient breath. *Talanta* **2014**, *120*, 386–390. [CrossRef]

71. Bajtarevic, A.; Ager, C.; Pienz, M.; Klieber, M.; Schwarz, K.; Ligor, M.; Ligor, T.; Filipiak, W.; Denz, H.; Fiegl, M.; et al. Noninvasive detection of lung cancer by analysis of exhaled breath. *BMC Cancer* **2009**, *9*, 1–16. [CrossRef]
72. Cateni, S.; Colla, V.; Vannucci, M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing* **2014**, *135*, 32–41. [CrossRef]
73. Liu, W.; Chawla, S. Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 345–356.
74. Akbani, R.; Kwek, S.; Japkowicz, N. Applying support vector machines to imbalanced datasets. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 39–50.
75. Torlay, L.; Perrone-Bertolotti, M.; Thomas, E.; Baciu, M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform.* **2017**, *4*, 159–169. [CrossRef]
76. Ogunleye, A.; Wang, Q.G. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 2131–2140. [CrossRef] [PubMed]
77. Ogunleye, A.; Wang, Q.G. Enhanced XGBoost-based automatic diagnosis system for chronic kidney disease. In Proceedings of the 2018 IEEE 14th International Conference on Control and Automation (ICCA), Anchorage, AK, USA, 12–15 June 2018; pp. 805–810.
78. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112, p. 181.
79. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 17 June 2021).
80. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media: Newton, MA, USA, 2019.
81. Fawcett, T. An introduction to ROC analysis. ROC Analysis in Pattern Recognition. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]

3.2. The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies

Sensors & Actuators: B. Chemical 408 (2024) 135550

Contents lists available at ScienceDirect

Sensors and Actuators: B. Chemical

journal homepage: www.elsevier.com/locate/snb

The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies

Anna Paleczek [a,*], Artur Rydosz [a,b]

[a] Institute of Electronics, Biomarkers Analysis LAB, AGH University of Krakow, Al. Mickiewicza 30, Krakow 30-059, Poland

[b] Advanced Diagnostic Equipment sp. z o.o., Krakow, Poland

ARTICLE INFO

Keywords:
Acetone
Blood glucose level
CatBoost
Gas concentration estimation
Machine learning
Sensor array

ABSTRACT

The growing number of people with diabetes has paved the way for the creation of non-invasive devices to measure blood glucose levels and enable the detection of diabetes non-invasively, for example, through exhaled air. This paper presents the development of a breath detection system using multiple sensors (e-nose system) for the non-invasive estimation of blood glucose levels. The system employs commercially available ethanol, CO_2, and acetone sensors, with synthetic breath mixtures used for testing, including four scenarios with high ethanol concentrations (0–570 ppm) as an influence factor. Results showed a mean absolute error of 0.245 ppm when analysing commonly observed acetone concentrations in diabetic breath using four sensors and the XGBoost Regressor. In mixtures with high ethanol concentrations and varying acetone concentrations (0–8.62 ppm), the CatBoost Regressor outperformed other machine learning algorithms with a mean absolute error of 0.568 ppm. This study emphasises the significant impact of ethanol on acetone detection and the need to consider ethanol levels in the developing of non-invasive devices for blood glucose prediction based on the exhaled acetone measurements. The research shows that a set of three gas sensors is optimal for estimating acetone concentrations in gas mixtures. The presented results constitute a preliminary step towards developing a non-invasive device for estimating blood glucose levels based on breath analysis, with the novelty of considering alcohol intake as a potential influencing factor.

1. Introduction

According to the World Health Organisation (WHO), the number of people with diabetes is growing rapidly, especially in low- and middle-income countries. The most common types are type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus (GDM). T2DM accounts for about 90% of patients. Diabetes causes many diseases of the heart and the kidneys as well as blindness.

Monitoring of blood glucose and early prevention play the most important role for all types of diabetes [1]. Currently, invasive devices such as self-blood glucose monitors (SBGM) and minimally invasive continuous glucose monitoring (CGM) systems are used to monitor blood glucose levels (BGL) [2–5]. However, there remains a need to develop a non-invasive system, for example, based on breath measurements. The non-invasive devices will be very helpful for screening tests and for people with needle fears and phobia, especially for children with diabetes. The number of people with type 2 diabetes and obesity that require regular glucose blood tests is increasing exponentially, and some studies have shown that this group does not follow clinical recommendations mostly due to needle fears [6].

Human breath consists mainly of nitrogen (78–79%), oxygen (13–16%), carbon dioxide (4%) and volatile organic compounds (VOCs) [7]. According to the literature, there are more than 3000 different VOCs in human exhaled air, which occur at very low concentrations of a few parts per million (ppm) or even a few parts per trillion (ppt) [8,9]. They can be of exogenous origin, for example, air pollution, drugs and cigarette smoke, as well as of endogenous origin [10,11]. The latter group is used most often as an indicator of the metabolic state of the body and includes biomarkers of diseases such as diabetes [12–14], asthma [15–17] and cancer [18–22].

A well-known biomarker of diabetes is exhaled acetone, which is produced in the body as a result of metabolic processes of glucose from blood, such as the oxidation of free fatty acids [23–25]. Many studies report acetone concentrations in exhaled air in the range of 0.2–5 ppm, where lower concentrations are observed in healthy people and higher in people with diabetes [26–31].

[*] Corresponding author.
E-mail address: paleczek@agh.edu.pl (A. Paleczek).

https://doi.org/10.1016/j.snb.2024.135550
Received 11 November 2023; Received in revised form 25 January 2024; Accepted 24 February 2024
Available online 27 February 2024

The expected correlation between acetone in exhaled air and blood glucose level is a positive correlation, where as BGL increases, the level of exhaled acetone increases. Such relationships have been observed previously in the literature [32,33]. However, some articles show that this correlation may also be negative and as BGL increases, the concentration of acetone in the exhaled air may decrease, and this relationship is individual for each person. These were the results shown by Rydosz. Researcher analysed acetone samples from several people with type 1 diabetes. Based on the research, he showed a negative correlation between acetone in exhaled air and blood glucose level [34]. Similar research was also conducted by Prabhakar *et al.* in which they also observed a negative correlation between acetone and blood glucose level examined in patients after fasting [35]. A negative correlation was also observed by Sha et al. [36]. This is an important observation and means that large-scale studies should be conducted, involving people with all types of diabetes and various treatments, as well as healthy people [37]. By creating a device based on machine learning algorithms, it will be possible to train or calibrate them for each patient.

Taking into account the low concentrations of VOCs, devices such as the gas chromatograph (GC) coupled to a mass spectrometer (MS) system [38] and proton transfer reaction time of flight mass spectrometry (PTR-TOF-MS) [14] are used to measure VOCs in the breath, but their disadvantage is their high price and large size. For this reason, scientists are working on various measurement techniques and devices that enable detection but with definitely lower prices, such as metal oxide semiconductor sensors (MOS). Due to the complex composition of breath and the lack of selectivity and poor sensitivity at low concentrations of gas sensors [39], it is necessary to use an array of sensors (e-nose) in order to identify the components of exhaled air. One of the methods to increase the 3 S properties of the gas sensor (known as sensitivity, selectivity and specificity) is to apply artificial intelligence algorithms, for example, deep learning models, for disease detection (classification) and gas concentration estimation (regression) [40].

In general, estimating gas concentration in mixtures using data from sensor matrices is a complex problem and this includes exhaled breath analysis. The most commonly used solutions use real-time analysis of the entire sensor response signal, taking into account the dynamics of its response [39,41–43]. Another approach is based on extracting information from one time point for each mix. Values such as sensitivity are then calculated and used as a characteristic vector to train machine learning and/or deep learning algorithms [44–46].

With regard to detecting diabetes in exhaled human air, the most common approach is the classification of people as either healthy or with diabetes [23,40,45,47–53]. Solutions based on the collection of breath samples from patients both online [45,54] and offline [55,56] are proposed as well as algorithm tests based on synthetic breath [47,57,58]. The most used classification algorithms are K-Nearest Neighbours [45,56], Support Vector Machines [45] and XGBoost [47,59]. However, due to research that indicates the relationship between blood glucose concentration and exhaled acetone concentration, it is necessary to develop a regression algorithm that enables the estimation of acetone concentration in gas mixtures. Such algorithms have been proposed by other researchers; for example, Li *et al.* developed a system to recognise gas concentrations in acetone and ethanol gas mixtures based on thin-line fabricated SnO_2 gas sensor and the 1D-CNN model. The authors used three concentrations of acetone (5, 10, 15 ppm) and three concentrations of ethanol (0.5, 1, 1.5 ppm) which results in the testing of nine types of mixtures. They compared different sensor data pre-processing methods, such as normalisation and trained 1D-CNN and modified 1D-CNN. The results showed that the minimum mean absolute error (MAE) to estimate acetone concentration was 0.47 ppm which was achieved by a modified 1D-CNN model and a SnO_2 sensor, the sensing area of which was optimised by the nanometric thin lines structure to increase its surface-to-volume ratio and the modified 1D-CNN algorithm [58].

Wang *et al.* used the Tafel curve for the quantitative detection of acetone and ethanol in gas mixtures at concentrations in the range of 0–200 ppm. As compound concentration estimators, the authors tested XGBoost, Gradient Boosting Decision Tree (GBDT) and Random Forest (RF) algorithms. The gradient-boosted algorithm showed a lower mean absolute percentage error (MAPE) of 11.3% and 23% for acetone and ethanol, respectively [59]. Zhu *et al.* recently proposed collecting data of acetone and ethanol and those mixtures using a CGS-8 intelligent gas sensing analysis system (Beijing Elite Technology Company Ltd.) and a seven-MOS sensor array. The authors proposed Kernel Principal Component Analysis (KPCA) as feature extraction and then hybrid machine learning algorithms such as AdaBoost, XGBoost (XGB), MVRVM, and SVM with or without optimisation (Grid Search, Particle Swarm Optimisation algorithm) for classification and gas concentration prediction. In the prediction of acetone concentration, they achieved absolute errors in the range from −0.072 to 0.094 ppm, RMSE 0.027 and R^2 0.9999. For ethanol, the results were −0.103 to 0.102 ppm, 0.030 and 0.9999, respectively. The presented results were for acetone and ethanol concentrations in the range of 1–12 ppm and for unknown relative humidity in the chamber [60].

The review of the literature suggests that the concentration of acetone in exhaled air can be correlated with the glucose concentration in the blood with a high level of precision. Additionally, a large proportion of the papers and patents in which machine learning algorithms are used are based on measurements of acetone in high concentrations (mostly above the acetone concentrations present in the exhaled human breath), e.g. Thakur *et al.* used partial least squares (PLS) and multiple linear regression (MLR) to predict acetone concentration in the range of 0–80 ppm [57], Chen *et al.* proposed a backpropagation neural network (BPNN) to predict the concentration of gases such as ethanol, toluene, acetone and formaldehyde in the range of 0–100 ppm [44] which, however, is far from practical application in breath analysis, where, as mentioned above, the concentration of acetone does not exceed 5 ppm.

Algorithms that enable the prediction of low concentrations of acetone in gas mixtures are necessary, especially when the real breath sampling is considered with portable devices such as e-nose systems. Moreover, to be considered as clinically accepted, the proposed algorithms should fulfil the requirements described by the ISO standards. Due to the lack of the ISO standard for exhaled breath analysis, the currently used standard for SBGM devices [61] can be used as a reference. However, it is expected that an ISO standard dedicated for exhaled analysis will be available in the coming years. Furthermore, it has to be underlined that the consumption of ethanol-based drinks will affect the glucose concentration in the blood even hours after consumption, thus the effect of high ethanol concentration on e-nose response for diabetes detection in exhaled breath has to be investigated and taken into account when non-invasive devices are considered. According to the American Diabetes Association (ADA), the drinking of alcohol by people with diabetes is not prohibited but requires caution for patient safety. The patient should monitor blood sugar levels and seek guidance from their healthcare team before consuming alcohol [62].

In this paper, we focus on the prediction of acetone concentration in simulated exhaled air, including after possible alcohol intake, and the selection of a set of commercially available sensors that will be sufficient to determine the acetone content. Additionally, it has to be understood that prior to the clinical trials, laboratory studies have to be conducted in order to limit the intake of alcohol by the patients taking part in the clinical trials. Thus, the presented results are the first step for including the influence of high ethanol concentration to the response of the e-nose system designed for acetone measurements of exhaled breath.

2. Experimental Section

2.1. Measurement system

The proposed gas measurement system is presented in Fig. 1. The system contained TGS1820, TGS2620 and TGS2600 sensors (Figaro

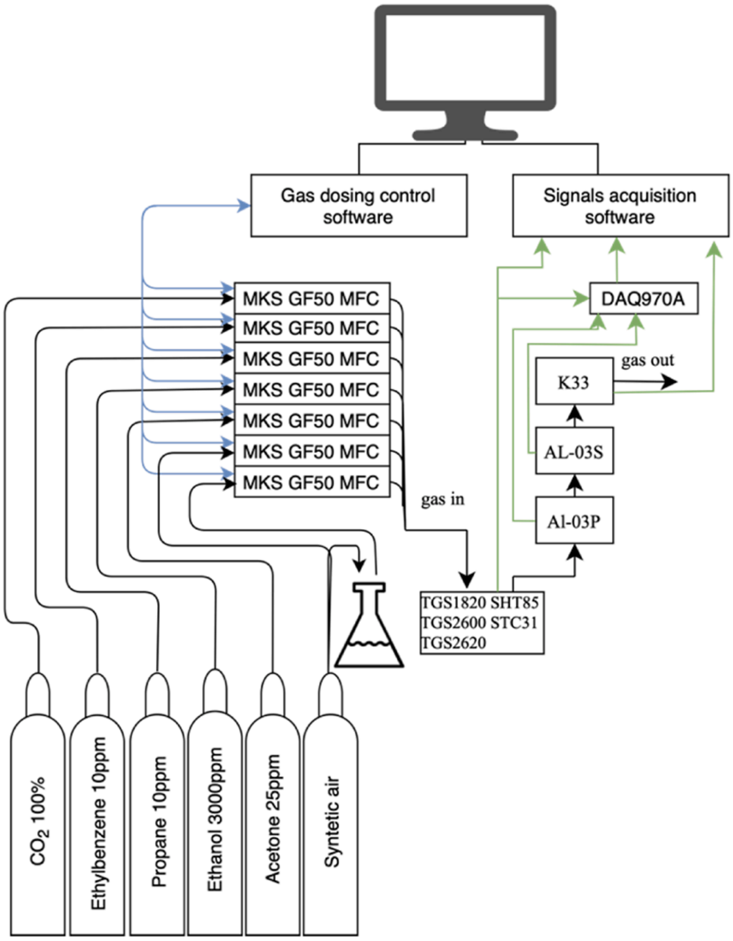


Fig. 1. Gas measurement system.

Engineering Inc, Mino, Osaka, Japan) and AL-03 P and AL-03S sensors (MGK SENSOR Co., Ltd., Saitama, Japan) and their voltage responses were measured using DAQ970A (Keysight Technologies, Santa Rosa, CA, USA). Figaro sensors were used with dedicated DevBoards that enable voltage measurement as a sensor response, while signal amplifiers were designed for MGK SENSOR sensors in accordance with the documentation provided by the sensor manufacturer. The SHT85 and STC31 sensors (Sensirion, Staefa ZH, Switzerland) were measured using the SEK-SensorBridge (Sensirion, Staefa ZH, Switzerland). The typical detection range and target gas of each of the sensors used are presented in Table 1. The communication between the PC and the K33 sensor (Senseair, Delsbo, Sweden) was implemented in the Python programming language using the UART and Modbus communication protocol, according to the manufacturer's suggestions. The control and communication of the gas dosing system, Sensirion sensors and DAQ970A were implemented in the Python programming language using libraries such as pyModbusTCP, pySerial, and pyVisa.

Table 1
Details of the sensors used in the measurements.

Manufacturer	Sensor Type	Target gas and typical detection range	Signal
Figaro Engineering Inc.	TGS1820	Acetone (1–20 ppm)	Voltage
	TGS2620	Ethanol (50–5000 ppm)	Voltage
	TGS2600	Air contaminants (1 ~ 30 ppm of H_2)	Voltage
MGK SENSOR Co.	AL-03S	Ethanol (0 ~ 2.0 mg/L)	Voltage
	AL-03 P	Ethanol (0 ~ 2.0 mg/L)	Voltage
Senseair	K33	Carbon dioxide (0–10%)	Carbon dioxide level as digital sensors' response
Sensirion	STC31	Carbon dioxide (0–100%)	Carbon dioxide level as digital sensors' response
	SHT85	Relative humidity (0–100%)	RH level as digital sensors' response

2.2. Gas mixtures

The gases were dosed using GF50 MKS (MKS, Andover, Massachusetts, USA) mass flow controllers (MFC). Each of the MFC had a full N_2 range of 10 standard cubic centimetres per minute (sccm). The total flow during the measurement was set to 10 sccm. For each mixture, the gas dosing time was 30 min and the purging time with synthetic air was 45 min. The specified durations were necessary to allow sufficient time for gas mixing in the gas configuration depicted in Fig. 1. It is important to note that both the gas dosing and purging times can be notably decreased if the patient exhales directly into the chamber equipped with gas sensors.

The prepared gas mixtures assumed concentrations of acetone in the range of 0–8 ppm (focussing most on the range of 0.5–2 ppm, which is most often mentioned in the literature as the range found with diabetic patients) [26–31] and high concentrations of ethanol in the range of 0–600 ppm. The water bubbler was used to manipulate the RH of the gas mixtures. The relative humidity of the simulation of the mixtures was in the range of 50–95% to mimic the humidity level in human exhaled air [63]. Detailed measurement conditions are presented in Table 2. The dataset is available online [64].

Fig. 2 shows the concentration of acetone in gas mixtures used during measurements and the marked range of acetone concentration commonly reported in literature on diabetes breath.

To simulate the possible concentration of ethanol in human breath, we assumed the maximum to be 2‰ BAC which is equal to 0.95 mg/l of BrAC - Blood Breath Ratio (BBR) assumed as 2100:1 (BBR range from 2000:1–2400:1 empirically determined for each country; BBR 2100:1 is used, for example, in Germany, the USA and Sweden) [65–67]. We then convert mg/l to ppm using Eq. 1 [68]. Assuming a temperature of 36°C and pressure during measurements of 1013 hPa, the 0.95 mg/l BrAC equals 522 ppm. Therefore, the 0–570 ppm range of ethanol was used since it covers the assumed conditions well. In practice, when BAC extend 2‰ the measurement based on exhaled breath analysis should not been performed and the user should be advised to use another method, preferably SBGM.

$$ppm = \frac{mg}{l} \bullet \frac{22,4}{M} \bullet \frac{(273+T)}{273} \bullet \frac{1}{10} \bullet \frac{1013}{P} \bullet 10000 \qquad (1)$$

where:

- *ppm* - BrAC in ppm;
- mg/l – BrAC in mg/l;
- 22.4 (L) - the volume of 1 mol at 1 atmospheric pressure at 0°C;
- 273 (K) - corresponds to 0°C, needed for conversion;
- 1013 (hPa) – atmospheric pressure;
- *P* – atmospheric pressure during measurements;
- *M* – molecular weight.

2.3. Data preprocessing

Sensor responses were calculated as given in Eq. 2. The data set was split into train and test sets in the 80:20 ratio.

$$S = R_g - R_0 \qquad (2)$$

Table 2
Composition of gas mixtures.

Gas	Concentration range (median) [unit]
Acetone	0–8.62 (1.4) [ppm]
Ethanol	0–570 (72) [ppm]
Propane	0–4.8 (0.35) [ppm]
CO_2	2.5–6.8 (4.4) [%]
Ethylbenzene	0–4.2 (0.3) [ppm]
RH	20–90 (70) [%]

Table 3
Measurement scenarios proposed in the paper.

Scenario No.	Acetone concentration [ppm]	Ethanol concentration [ppm]
1.	0–2.5	0
2.	0–2.5	0–570
3.	0–8.62	0–570
4.	0–8.62	0

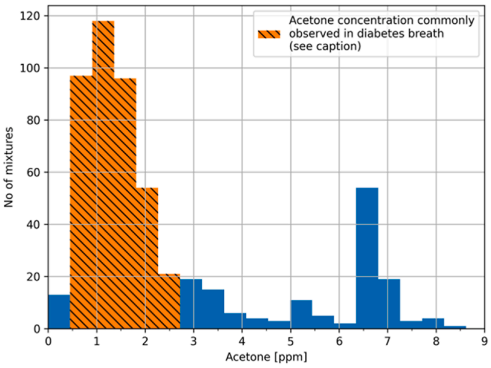


Fig. 2. Acetone concentration in gas mixtures. The highlighted range includes the concentration of acetone commonly observed in diabetes breath based on the literature review [26–31].

where:

- *S* – sensor response;
- R_g – sensor response under exposure to the gas mixture;
- R_0 – sensor response under the synthetic air and RH exposure.

2.4. Algorithms

As a first step, we decided to compare mean absolute error for different sets of sensors and algorithm such as Linear Regression (LR), Random Forest Regressor, Decision Tree Regressor (DT), XGBoost Regressor, Light GBM Regressor (LGBM) and CatBoost Regressor (CB). Algorithms were implemented using the Python programming language and scikit-learn Machine Learning in Python [69,70] and XGB [71], LGBM [72,73], CB [74,75] software libraries. Sets were constructed using Table 4. For each test, the sensor response and algorithm were chosen from each column. The test also included the value of 'None', which indicates the absence of a specific sensor.

The best set with regard to the mean absolute error and the mean square error was chosen. The sensor choice procedure was performed to estimate the acetone concentration in four different scenarios (Table 3).

Table 4
Sensors and algorithms used to create test sets. The responses of each sensor were given as R_G and *S*. *S* means the sensor response given by Equation 2 and R_G means the sensor response under exposure to the gas mixture.

Acetone sensor	Ethanol sensor	CO_2 sensor	RH sensor	Algorithm
None, TGS1820	None, TGS2620, TGS2600, AL-03 P, AL-03S	None, K33, STC31	None, SHT85	LR, RF, DT, XGB, CB, LGBM

Scenarios 1 and 2 were created to predict the acetone concentration commonly observed in diabetes breath (see Fig. 2), whereas Scenario 3 and 4 were created to predict acetone concentration across a wider range also reported in the literature.

3. Results

3.1. Sensors' characteristics

The presence of high concentrations of ethanol affects the sensitivity of many metal oxide-based sensors, including sensors used in this study (see Table 1). The TGS1820 sensor was designed to work as a standalone acetone sensor [76]; however, in the case of predicting acetone concentration, it was not sufficient to use only the TGS1820 sensor (as the experimental results have shown). The sensor response under exposure to the various concentration of acetone in the mixtures is presented in Fig. 3 and Fig. 4. Analysis of the obtained results shows that the TGS1820 sensor is imperfectly selective for acetone in the presence of high concentrations of ethanol; however, by applying data processing algorithms, this issue can be overcome.

The results showed that it is important to analyse both R_G and S. In the absence of ethanol in the mixture, we noticed the linear S of the TGS1820 sensor. However, in the case of the R_G, the sensor response to acetone concentrations is less linear. However, the S analysis showed that the sensor has an offset in response which is dependent upon the concentration of ethanol present in the mixture. These observations were obtained for several sensors used during the experiments. For this reason, it was necessary to use a set of sensors and machine learning algorithms and consider developing a system with two possible cases: in the first case, if any alcohol is detected, the device will select an algorithm specifically designed for individuals who have consumed alcohol. Alternatively, in the second case, upon detecting alcohol during the breath test, the device will inform the patient that it cannot provide a measurement due to the presence of alcohol.

The TGS1820 response R_G to a different concentration of pure acetone at different levels of relative humidity was shown in Fig. 3a, whereas Fig. 3b presents the TGS1820 response R_G which is dependent upon the concentration of acetone in all the mixtures used during measurements, i.e. acetone and ethanol in the range of 0–8.62 ppm and 0–570 ppm, respectively. Ethanol levels are explained in the figure legend. Similarly, the results for S are presented in Figs. 4a and 4b. The results show that TGS1820 R_G and S have a linear dependence (R^2 score in range 0.69–0.99 – detailed results are shown in Figs. 4a and 4b) in pure acetone and RH mixtures; however, in the presence of ethanol and other gases in the mixture, the sensor response is no longer linear in either case. This proves that in the case of the presence of ethanol in the sample, simple algorithms are not valid and more sophisticated approaches have to be implemented. Another approach is to use a highly selective ethanol sensor that can operate at high RH concentrations, i.e. AL-03S.

As shown in Fig. 5, the AL-03S ethanol sensor has high selectivity and linear response (R^2 score 0.97) for ethanol concentration at different levels of RH, so it can be used as a reference sensor to determine alcohol intake before measurements. Therefore, we recommend to use it a first step to detect alcohol consumption and as a reference sensor for algorithms.

The primary distinction between acetone and ethanol is their chemical categorisation: acetone is a ketone, while ethanol is an alcohol. While these organic compounds consist of carbon, hydrogen, and oxygen, their divergent chemical and physical properties classify them into distinct groups. Acetone, formulated as $(CH_3)_2CO$, notably varies from ethanol, which has the chemical formula C_2H_5OH. Despite their separate groups and differing properties, both compounds, due to their shared carbon, hydrogen, and oxygen content, can interfere with each other in the context of sensors relying on metal oxides. Semiconductor sensors show a similar mechanism of sensitivity to compounds such as acetone and ethanol, and ethanol after alcohol consumption reaches concentrations 100 times higher than the concentrations of acetone present in human breath, even in diabetic patients for whom increased concentrations are observed. The ethanol sensors used in this work show high selectivity and sensitivity to the acetone contained in the mixtures and do not show the influence of other gases contained in the mixture on their response. The tests were performed on gas mixtures with different RH contents, but the results showed that the sensors did not show a significant difference in response at different relative humidity values, which is an important advantage because of the high RH content in exhaled air.

The choice of whether to consider the raw sensor response R_G or the relative response S in the case of TGS1820 depends on the presence of ethanol. If ethanol is absent from the mix, the S response to acetone is more linear than the response of R_G. However, in the presence of ethanol, the characteristic S is not linear, and a better choice is the R_G response, which shows a linear dependence (offset) of the sensor response depending on the concentration of ethanol in the mixture (as presented in Fig. 3 and Fig. 4). Interestingly, in the case of the ethanol sensor, better results were obtained using the AL-03S sensor when there

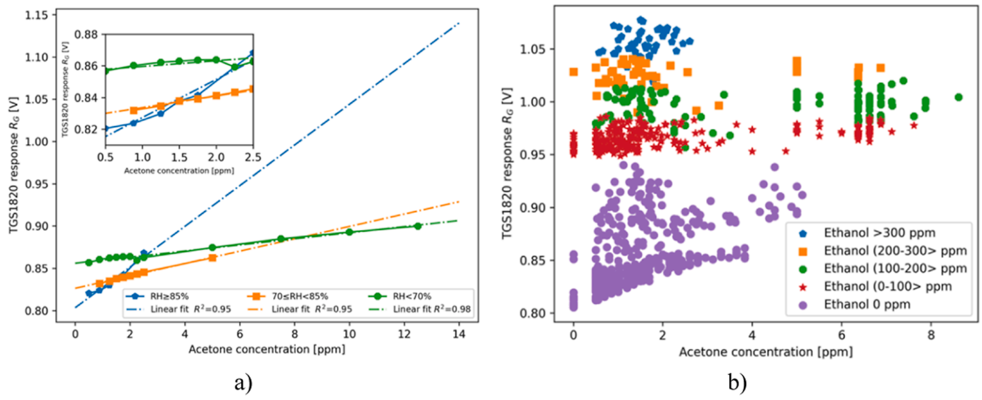


Fig. 3. Dependence of TGS1820 R_G on the concentration of acetone in: a) pure acetone in different levels of relative humidity; b) all gas mixtures used in measurements in different levels of RH (ethanol concentration is shown in the figure legend).

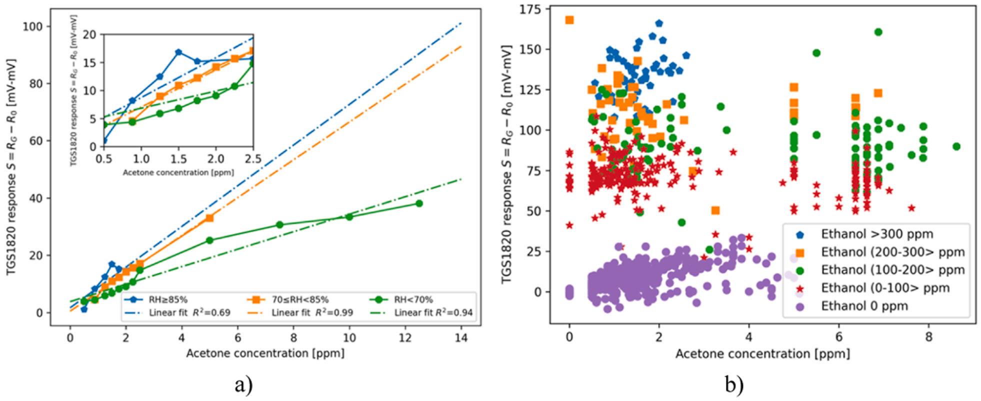


Fig. 4. Dependence of the sensitivity *S* TGS1820 on acetone concentration in: a) pure acetone at different levels of relative humidity; b) all gas mixtures used in measurements at different levels of RH (ethanol concentration is shown in the figure legend).

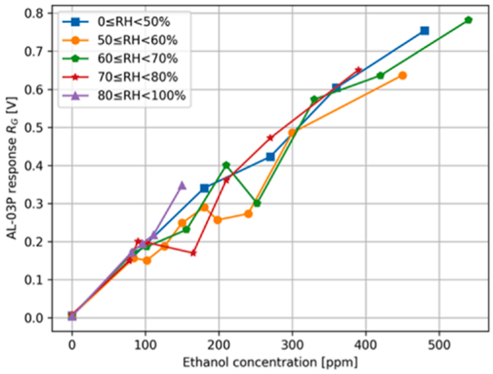


Fig. 5. Dependence of AL-03S R_G on ethanol concentration in gas mixtures.

was no ethanol in the mixture and the TGS2620 when the mixture contained ethanol. For this reason, when designing the target device, it is necessary to make assumptions regarding its operating conditions and prepare the final ensemble model.

Table 5
Results of experiments in the estimation of the acetone concentration. A-acetone concentration, E-ethanol concentration, and FR – full range of concentrations. The results are given in the following order: MAE, min error, max error, RMSE, mean MAE of 5-fold cross-validation.

No.	Scenario	Results	Best set
1.	A < 2.5 E = 0	0.245, -0.691, 0.600, 0.091, 0.297	TGS1820_S, AL-03S_RG, STC31_RG, SHT85, XGB
2.	A < 2.5 E FR	0.358, -1.182, 0.844, 0.215, 0.392	TGS1820_RG, TGS2620_RG, SHT85, CB
3.	A FR E FR	0.567, -2.677, 2.621, 0.639, 0.541	TGS1820_RG, TGS2620_RG, SHT85, STC_31_RG, CB
4.	A FR E = 0	0.430, -1.488, 1.396, 0.336, 0.418	TGS1820_S, AL03-S_S, STC_31_S, SHT85, CB

3.2. Acetone concentration prediction

The results have shown that the CatBoost Regressor outperforms other algorithms. For the first scenario (see Table 3), the best results were obtained using the XGBoost Regressor, where the feature importance analysis shows that the most important for acetone concentration prediction were TGS1820_S (42%) and STC31_RG (37%). In the second scenario, data from TGS1820 had the highest feature importance (55%) and the TGS2620_RG ethanol sensor had 37% importance. In the case of 0–8.12 ppm acetone concentrations in mixtures (Scenario 3) the importance of TGS1820 decreased slightly and TGS2620 increased. In the last scenario presented in Table 5, the most important was, as in all cases, the TGS1820 response (46%), but of almost similar importance was the ethanol sensor (44%), so it is very important to take the into account and to also use an ethanol sensor in addition to the acetone sensor. The analysis shows that in the case of the presence of ethanol in gas mixtures, ethanol sensors are crucial and significantly improve the prediction. The TGS1820 acetone sensor was of greatest importance in all scenarios, but it is not sufficient as a standalone sensor to predict acetone concentrations in gas mixtures due to the cross-selectivity of the observed sensor. The tests performed according to the four proposed scenarios, which describe situations that may occur in real conditions quite well, have shown that the TGS1820 sensor is completely ineffective in situations when even a trace of ethanol appears in the sample, despite its response having the highest feature importance. It has to be underlined that alcohol cannot be completely excluded from the diet of a person with diabetes; therefore, it is necessary to take this factor into account when it comes to the practical use of test results and the development of a device for predicting blood glucose levels based on exhaled air. As discussed before, the results in Table 5 show that without the presence of ethanol, the most important factor is the R_G response, but with the presence of ethanol, the most important factor is the sensor *S* sensitivity, so it will be important to develop a multi step algorithm which first analyses ethanol concentration and then applies an appropriate model for acetone concentration prediction.

When analysing the results, the use of the CO_2 sensor slightly decreased the prediction error. The best results were obtained in a set with the STC31 sensor. Furthermore, the use of the SHT85 sensor to monitor RH increased the accuracy of the prediction. However, to limit the number of sensors (and device costs) in the target breath acetone monitoring device, it is possible to monitor RH using the slightly less accurate SHT31 sensor incorporated into the STC31 sensor used for CO_2 detection. Thus, three parameters can be measured by a single module, namely temperature, RH and CO_2.

The predictions of acetone concentration tested under four measurement scenarios that mimic real-case situations are presented in Fig. 6. These are the results of the algorithm running on the test data. When comparing Fig. 6a and Fig. 6b (acetone $<$ 2.5 ppm, without and with ethanol presence), there is a noticeable overestimation of the acetone prediction results in the case where ethanol is present in the mixture, the same when comparing Fig. 6c and Fig. 6b. The 'offset' effect was also observed in Fig. 3 and Fig. 4, but it was handled by the proposed algorithms. In each scenario, the algorithm is prone to underestimate the predicted acetone relative to the acetone present in the mixture, especially in Scenario 4 (acetone full range without ethanol). In the target device for estimating the blood glucose concentration based on exhaled air, it will also be necessary to calibrate the device based on data collected from patients.

4. Discussion

There is a lack of research that combines acetone detection (including diabetes prediction) in exhaled air in mixtures with high concentrations of ethanol that simulate alcohol intake and even other influencing factors such as food, drinks oral hygiene products and smoking. Such analyses are important considerations in the research and development of non-invasive devices, for example, those designed to monitor blood glucose levels based on exhaled air. Taking into account the results presented in this study, it is necessary to measure alcohol in breath during acetone measurement to ensure that BGL estimation will be reliable and not influenced by ethanol. It is important to note that this presence may not necessarily be attributed to direct alcohol consumption but could also originate from certain drugs and various food items such as sweets and salsas. Additionally, when real-case diabetes monitoring is considered, it has to be understood that blood glucose is measured before and after food intake, thus ethanol may appear in exhaled breath and the e-nose also has to have scenarios to work in this case. Moreover, there is a need to identify other factors such as eating different types of food, smoking before measurement using a breath-based device, and test those influence to sensors response. Then, for example, as shown in this study, a different algorithm based on the preliminary detection of the influencing factor is applied and the user is warned about the possibility of a lack of reliability of the device.

Review of the literature shows that there is also a lack of acetone concentration prediction in the low range commonly observed in

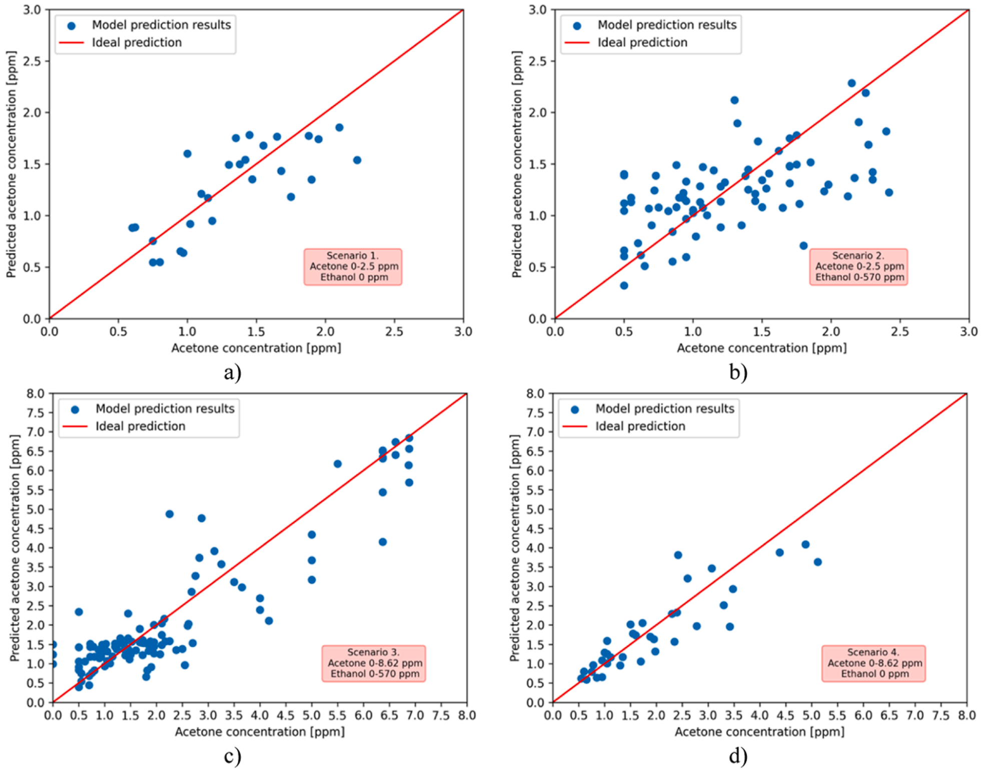


Fig. 6. Algorithms results on a test set: a) Scenario 1, b) Scenario 2, c) Scenario 3, d) Scenario 4.

diabetic breath using machine learning algorithms. One such example is the work of Chen *et al.* who proposed the prediction of acetone in the range of 0–100 ppm [44], Wang *et al.* predicted acetone in the range of 0–200 ppm [59].

We have developed algorithms using gas mixtures with acetone in the range of 0–8.12 ppm and a median value of 1.42 ppm, which are the lowest values from similar studies included in the literature review, and it is more demanding and challenging to correctly predict small differences between concentrations such as 0.2 ppm than 20 ppm. Sensitivity and detection limits are directly related to the sensors and their physical capabilities, and the use of machine learning algorithms may slightly improve selectivity but not sensitivity. Li *et al.* tested only nine mixtures with acetone and ethanol (in comparison with the 560 different compositions used within this study) at low concentrations and achieved MAE 0.42 ppm in the prediction of the acetone concentration using 1D-CNN [58]. This result is similar to ours (named Scenario 4), but the experiments were performed on limited gas concentrations. In the study performed by Zhu *et al.,* the authors achieved a lower RMSE range of −0.072 to 0.094 ppm (our RMSE results are 0.091–0.639 ppm) in predicting the concentration of acetone gas. Experiments were conducted using sixty-two gas samples containing acetone, ethanol (and those mixtures) but were limited only to 0–12 ppm for each gas [60]. Considering the use of a system proposed by the authors in the detection of diabetes, more contaminants (factors of influence) are added to the gas mixtures to best mimic breath.

Our study was also the first one to analyse the influence of sensors on the prediction of acetone concentration and proposed an optimal set of sensors to develop a non-invasive device to estimate BGL. Reducing the number of sensors in the device is important to reduce production costs, device size and power consumption. In this paper, we have also presented one of the first uses of CatBoost for gas sensor data, which are typically quantitative variables (not categorical) and contain noise, especially for the prediction of acetone concentration.

As presented above, it is confirmed that the TGS1820, the acetone sensor, is not fully selective for acetone in the presence of a high concentration of ethanol; therefore, it is necessary to develop several algorithms to predict the concentration of acetone in exhaled air and the initial detection of ethanol, and to then use the appropriate algorithm to predict acetone and warn the patient that the results after drinking alcohol can be unreliable. The set of sensors and the choice of the ethanol sensor depend on whether the target device is adapted to detect acetone after consuming ethanol. The results show that the use of three sensors is sufficient to predict acetone in exhaled air.

5. Conclusions

In conclusion, this research has demonstrated the effectiveness of employing ethanol, CO_2 and acetone sensors in estimating acetone concentrations in breath samples, which is crucial for measuring diabetic blood glucose. The study revealed that while a combination of four sensors and the XGBoost Regressor yielded a mean absolute error of 0.245 ppm for typical acetone concentrations, the CatBoost Regressor performed better (0.568 ppm) in scenarios with high levels of ethanol. These findings underscore the necessity of accounting for the influence of ethanol on acetone detection when designing non-invasive glucose measurement devices. The study suggests that a set of three gas sensors offers the optimal estimation of acetone concentrations and marks a promising initial step towards developing non-invasive glucose measurement tools based on breath analysis, pioneering the consideration of alcohol intake as a potential influential factor. The next step of the research work will be a study using the breath of healthy people and people with diabetes in order to find the correlation between blood glucose level and exhaled acetone also after alcohol intake during the measurement and verification of the proposed system. Additionally, the developed regression algorithm will be adjusted to fulfil the requirements described in EN ISO 15197:2015 [61].

Funding

The work was partially supported by the IDUB AGH 4122 grant and the statutory activity at the Institute of Electronics AGH and the NCN OPUS 2021/41/B/ST7/0027612 grant.

CRediT authorship contribution statement

Anna Paleczek: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Artur Rydosz:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

We cited dataset in the article: https://doi.org/10.58032/AGH/S0CH7M.

References

[1] Diabetes. [cited 2023 Jan 22]. Available from: 〈https://www.who.int/health-topics/diabetes#tab=tab_1〉.

[2] R.M. Bergenstal, D.M. Mullen, E. Strock, M.L. Johnson, M.X. Xi, Randomized comparison of self-monitored blood glucose (BGM) versus continuous glucose monitoring (CGM) data to optimize glucose control in type 2 diabetes, J. Diabetes Complicat. 36 (3) (2022) 108106.

[3] M. Montagnana, M. Caputo, D. Giavarina, G. Lippi, Overview on self-monitoring of blood glucose, Clin. Chim. Acta 402 (1–2) (2009) 7–13.

[4] J.R. Petrie, A.L. Peters, R.M. Bergenstal, R.W. Holl, G.A. Fleming, L. Heinemann, Improving the clinical value and utility of CGM systems: issues and recommendationsa joint statement of the European Association for the Study of Diabetes and the American Diabetes Association Diabetes Technology Working Group, Diabetes Care 40 (12) (2017) 1614–1621. 〈https://diabetesjournals.org/care/article/40/12/1614/36887/Improving-the-Clinical-Value-and-Utility-of-CGM〉 [cited 2023 Mar 21].

[5] N. Poolsup, N. Suksomboon, A.M. Kyaw, Systematic review and meta-analysis of the effectiveness of continuous glucose monitoring (CGM) on glucose control in diabetes, Diabetol. Metab. Syndr. 5 (1) (2013) 1–14. 〈https://link.springer.com/articles/10.1186/1758-5996-5-39〉. cited 2023 Mar 21].

[6] A. Rydosz, 2022, Diabetes Without Needles: Non-invasive Diagnostics and Health Management, Elsevier1–302, (Available from), http://www.sciencedirect.com:5070/book/9780323998871/diabetes-without-needles..

[7] B. Buszewski, M. Kesy, T. Ligor, A. Amann, Human exhaled air analytics: biomarkers of diseases, Biomed. Chromatogr. 21 (6) (2007) 553–566.

[8] M. Phillips, J. Herrera, S. Krishnan, M. Zain, J. Greenberg, R.N. Cataneo, Variation in volatile organic compounds in the breath of normal humans, J. Chromatogr. B Biomed. Sci. Appl. 729 (1–2) (1999) 75–88.

[9] A. Smolinska, E.M.M. Klaassen, J.W. Dallinga, K.D.G. van de Kant, Q. Jobsis, E.J. C. Moonen, et al., Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children, PLoS One 9 (4) (2014). Available from: 〈https://pubmed.ncbi.nlm.nih.gov/24752575/〉.

[10] M. Westhoff, M. Friedrich, J.I. Baumbach, Simultaneous measurement of inhaled air and exhaled breath by double multicapillary column ion-mobility spectrometry, a new method for breath analysis: results of a feasibility study, ERJ Open Res. 8 (1) (2022). 〈https://openres.ersjournals.com/content/8/1/00493-2021〉 (Available from).

[11] W. Filipiak, V. Ruzsanyi, P. Mochalski, A. Filipiak, A. Bajtarevic, C. Ager, et al., Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants, J. Breath. Res. 6 (3) (2012) 036008. Available from: 〈https://iopscience.iop.org/article/10.1088/1752-7155/6/3/036008〉.

[12] J. Guo, D. Zhang, T. Li, J. Zhang, L. Yu, Green light-driven acetone gas sensor based on electrospinned CdS nanospheres/Co3O4 nanofibers hybrid for the detection of exhaled diabetes biomarker, J. Colloid Interface Sci. 606 (2022) 261–271.

[13] I. Fufurin, P. Berezhanskiy, I. Golyak, D. Anfimov, E. Kareva, A. Scherbakova, et al., Deep learning for type 1 diabetes mellitus diagnosis using infrared quantum cascade laser spectroscopy, Materials 15 (9) (2022) 2984. Available from: 〈https://www.mdpi.com/1996-1944/15/9/2984/htm〉.

[14] S.K. Das, K.K. Nayak, P.R. Krishnaswamy, Q. Wang, S. Ricote, Y. Wang, et al., Indole as a new tentative marker in exhaled breath for non-invasive blood glucose

monitoring of diabetic subjects, J. Breath. Res. 16 (2) (2022) 026001. Available from: ⟨https://iopscience.iop.org/article/10.1088/1752-7163/ac4610⟩.

[15] J.M. Escamilla-Gil, M. Fernandez-Nieto, N. Acevedo, Understanding the cellular sources of the fractional exhaled nitric oxide (FeNO) and its role as a biomarker of type 2 inflammation in asthma, Biomed. Res. Int. 2022 (2022).

[16] P. Xepapadaki, Y. Adachi, C.F. Pozo Beltrán, Z.A. El-Sayed, R.M. Gómez, E. Hossny, et al., Utility of biomarkers in the diagnosis and monitoring of asthmatic children, World Allergy Org. J. 16 (1) (2023) 100727.

[17] G. Guida, V. Carriero, F. Bertolini, S. Pizzimenti, E. Heffler, G. Paoletti, et al., Exhaled nitric oxide in asthma: from diagnosis to management, Curr. Opin. Allergy Clin. Immunol. 23 (1) (2023) 29–35.

[18] H. Tyagi, E. Daulton, A.S. Bannaga, R.P. Arasaradnam, J.A. Covington, Electronic nose for bladder cancer detection, Chem. Proc. 5 (2021) 1.

[19] M.H.M.C. Scheepers, Z. Al-Difaie, L. Brandts, A. Peeters, B. van Grinsven, N. D. Bouvy, Diagnostic performance of electronic noses in cancer diagnoses using exhaled breath: a systematic review and meta-analysis, JAMA Netw. Open [Internet] 5 (6) (2022) e2219372–e2219372. Available from: ⟨https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2793773⟩.

[20] S.R. Sutaria, S.S. Gori, J.D. Morris, Z. Xie, X.A. Fu, M.H. Nantz, Lipid peroxidation produces a diverse mixture of saturated and unsaturated aldehydes in exhaled breath that can serve as biomarkers of lung cancer- a review, Metabolites 2022 12 (6) (2022) 561. ⟨https://www.mdpi.com/2218-1989/12/6/561/htm⟩ (Available from).

[21] S. Kazeminasab, R. Ghanbari, B. Emamalizadeh, V. Jouyban-Gharamaleki, A. Taghizadieh, A. Jouyban, et al., Exhaled breath condensate efficacy to identify mutations in patients with lung cancer: a pilot study, Nucleosides Nucl. Nucl. Acids 41 (4) (2022) 370–383. ⟨https://www.tandfonline.com/doi/abs/10.1080/15257770.2022.2046278⟩ (Available from).

[22] R. Anzivino, P.I. Sciancalepore, S. Dragonieri, V.N. Quaranta, P. Petrone, D. Petrone, et al., The role of a polymer-based e-nose in the detection of head and neck cancer from exhaled breath, Sensors 2022 22 (17) (2022) 6485. Available from: ⟨https://www.mdpi.com/1424-8220/22/17/6485/htm⟩.

[23] A.T. Güntner, I.C. Weber, S. Schon, S.E. Pratsinis, P.A. Gerber, Monitoring rapid metabolic changes in health and type-1 diabetes with breath acetone sensors, Sens. Actuators B Chem. 367 (2022) 132182.

[24] W. Li, Y. Liu, X. Lu, Y. Huang, Y. Liu, S. Cheng, et al., A cross-sectional study of breath acetone based on diabetic metabolic disorders, J. Breath. Res. 9 (1) (2015) 016005. Available from: ⟨https://iopscience.iop.org/article/10.1088/1752-7155/9/1/016005⟩.

[25] P.R. Galassetti, B. Novak, D. Nemet, C. Rose-Gottron, D.M. Cooper, S. Meinardi, et al., Breath ethanol and acetone as indicators of serum glucose levels: an initial report, Diab. Technol. Therap. 7 (1) (2005) 115–123. ⟨https://www.liebertpub.com/doi/10.1089/dia.2005.7.115⟩. Available from: ⟨https://www.liebertpub.com/doi/10.1089/dia.2005.7.115⟩.

[26] A. Rydosz, Sensors for enhanced detection of acetone as a potential tool for noninvasive diabetes monitoring, Sensors 18 (7) (2018) E2298.

[27] C. Deng, J. Zhang, X. Yu, W. Zhang, X. Zhang, Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization, J. Chromatogr. B 810 (2) (2004) 269–275.

[28] I. Ueta, Y. Saito, M. Hosoe, M. Okamoto, H. Ohkita, S. Shirai, et al., Breath acetone analysis with miniaturized sample preparation device: in-needle preconcentration and subsequent determination by gas chromatography–mass spectroscopy, J. Chromatogr. B 877 (24) (2009) 2551–2556.

[29] K. Schwarz, A. Pizzini, B. Arendacká, K. Zerlauth, W. Filipiak, A. Schmid, et al., Breath acetone—aspects of normal physiology related to age and gender as determined in a PTR-MS study, J. Breath. Res. 3 (2) (2009) 027003.

[30] V. Saasa, M. Beukes, Y. Lemmer, B. Mwakikunga, Blood ketone bodies and breath acetone analysis and their correlations in type 2 diabetes mellitus, Diagnostics 9 (4) (2019).

[31] H. Dong, L. Qian, Y. Cui, X. Zheng, C. Cheng, Q. Cao, et al., Online accurate detection of breath acetone using metal oxide semiconductor gas sensor and diffusive gas separation, Front Bioeng. Biotechnol. 10 (2022) 296.

[32] C. Wang, A. Mbi, M. Shepherd, A study on breath acetone in diabetic patients using a cavity ringdown breath analyzer: exploring correlations of breath acetone with blood glucose and glycohemoglobin A1C, IEEE Sens. J. 10 (1) (2010) 54–63.

[33] M. Sun, Z. Wang, Y. Yuan, Z. Chen, X. Zhao, Y. Li, et al., Continuous monitoring of breath acetone, blood glucose and blood ketone in 20 type 1 diabetic outpatients over 30 days, J. Anal. Bioanal. Tech. 8 (5) (2017) 1–8. ⟨https://www.omicsonline.org/open-access/continuous-monitoring-of-breath-acetone-blood-glucose-and-blood-ketone-in-20-type-1-diabetic-outpatients-over-30-days-2155-9872-1000386-95035.html⟩ (Available from).

[34] A. Rydosz, A negative correlation between blood glucose and acetone measured in healthy and type 1 diabetes mellitus patient breath, J. Diabetes Sci. Technol. 9 (4) (2015) 881 [cited 2024 Jan 18].

[35] A. Prabhakar, A. Quach, D. Wang, H. Zhang, M. Terrera, D. Jackemeyer, et al., Breath acetone as biomarker for lipid oxidation and early ketone detection, Glob. J. Obes. Diabetes Metab. Syndr. 1 (1) (2014) 012–019. Available from: ⟨http://www.peertechzpublications.org/Obesity-Diabetes-Metabolic-Syndrome/GJODMS-1-103.php⟩.

[36] M.S. Sha, M.R. Maurya, S. Shafath, J.J. Cabibihan, A. Al-Ali, R.A. Malik, et al., Breath analysis for the in vivo detection of diabetic ketoacidosis, ACS Omega [Internet] 7 (5) (2022) 4257–4266. ⟨https://pubs.acs.org/doi/full/10.1021/acsomega.1c05948⟩ (Available from:).

[37] C. Turner, C. Walton, S. Hoashi, M. Evans, Breath acetone concentration decreases with blood glucose concentration in type I diabetes mellitus patients during hypoglycaemic clamps, J. Breath. Res. 3 (4) (2009) 046004. Available from: ⟨https://iopscience.iop.org/article/10.1088/1752-7155/3/4/046004⟩.

[38] F. Monedeiro, M. Monedeiro-Milanowski, I.A. Ratiu, B. Brożek, T. Ligor, B. Buszewski, Needle trap device-GC-MS for characterization of lung diseases based on breath VOC profiles, Molecules 2021 26 (6) (2021) 1789. Available from: ⟨https://www.mdpi.com/1420-3049/26/6/1789/htm⟩.

[39] S. De Vito, A. Castaldo, F. Loffredo, E. Massera, T. Polichetti, I. Nasti, et al., Gas concentration estimation in ternary mixtures with room temperature operating sensor array using tapped delay architectures, Sens Actuators B Chem. 124 (2) (2007) 309–316. Available from: ⟨https://www.researchgate.net/publication/235641849_Gas_concentration_estimation_in_ternary_mixtures_with_room_temperature_operating_sensor_array_using_tapped_delay_architectures⟩.

[40] A. Paleczek, A. Rydosz, Review of the algorithms used in exhaled breath analysis for the detection of diabetes, J. Breath. Res. 16 (2) (2022) 026003. Available from: ⟨https://iopscience.iop.org/article/10.1088/1752-7163/ac4916⟩.

[41] V. Pareek, S. Chaudhury, S. Singh, Hybrid 3DCNN-RBM Network for Gas Mixture Concentration Estimation with Sensor Array, IEEE Sens J. (2021).

[42] M. Li, J. He, R. Zhou, L. Ning, Y. Liang, Research on prediction model of mixed gas concentration based on CNN-LSTM network, ACM Int. Conf. Proc. Ser. (2021). ⟨https://dl.acm.org/doi/10.1145/3503047.3503110⟩ [cited 2023 Jan 22]; Available from:.

[43] M. Kang, I. Cho, J. Park, J. Jeong, K. Lee, B. Lee, et al., High accuracy real-time multi-gas identification by a batch-uniform gas sensor array and deep learning algorithm, ACS Sens. [Internet] 7 (2) (2022) 430–440. ⟨https://pubs.acs.org/doi/full/10.1021/acssensors.1c01204⟩ (Available from:).

[44] Z. Chen, Y. Zheng, K. Chen, H. Li, J. Jian, Concentration estimator of mixed VOC gases using sensor array with neural networks and decision tree learning, IEEE Sens. J. 17 (6) (2017 Mar 15) 1884–1892.

[45] Hariyanto, R. Sarno, D.R. Wijaya, Detection of diabetes from gas analysis of human breath using e-Nose. Proceedings of the 11th International Conference on Information and Communication Technology and System, ICTS, 2017, pp. 241–246, 2018 Jan 19;2018-January.

[46] Y. Xu, X. Zhao, Y. Chen, W. Zhao, Research on a mixed gas recognition and concentration detection algorithm based on a metal oxide semiconductor olfactory system sensor array, Sensors 2018 18 (10) (2018) 3264. Available from: ⟨https://www.mdpi.com/1424-8220/18/10/3264/htm⟩.

[47] A. Paleczek, D. Grochala, A. Rydosz, Artificial breath classification using xgboost algorithm for diabetes detection, Sensors 21 (12) (2021).

[48] Rydosz A., Marszałek K., Putynkowski G. A Novel Approach for Device Dedicated to Non-Invasive Diabetes Control. J Diabetes Treat. 2020;

[49] S. Lekha, M. Suchetha, Non-invasive diabetes detection and classification using breath analysis. 2015 International Conference on Communication and Signal Processing, ICCSP 2015, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 0955–0958.

[50] R. Sarno, S.I. Sabilla, D.R. Wijaya, Electronic nose for detecting multilevel diabetes using optimized deep neural network, Eng. Lett. 28 (1) (2020).

[51] D. Guo, D. Zhang, N. Li, L. Zhang, J. Yang, Diabetes identification and classification by means of a breath analysis system, Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.) (2010) 52–63. ⟨https://link.springer.com/chapter/10.1007/978-3-642-13923-9_6⟩ [cited 2023 Mar 15];6165 LNCS.

[52] H.K. Akturk, J. Snell-Bergeon, L. Pyle, E. Fivekiller, S. Garg, E. Cobry, Accuracy of a breath ketone analyzer to detect ketosis in adults and children with type 1 diabetes, J. Diabetes Complicat. 35 (2021) 1056–8727, https://doi.org/10.1016/j.jdiacomp.2021.108030.

[53] K. Yan, D. Zhang, D. Wu, H. Wei, G. Lu, Design of a breath analysis system for diabetes screening and blood glucose level prediction, IEEE Trans. Biomed. Eng. 61 (11) (2014) 2787–2795.

[54] R. Kalidoss, S. Umapathy, R. Kothalam, U. Sakthivelu, Adsorption kinetics feature extraction from breathprint obtained by graphene based sensors for diabetes diagnosis, J. Breath. Res 15 (1) (2020) 016005.

[55] T. Saidi, O. Zaim, M. Moufid, N. el Bari, R. Ionescu, 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, Sens. Actuators B 257 (2018) 178–188.

[56] D. Guo, D. Zhang, N. Li, L. Zhang, J. Yang, A novel breath analysis system based on electronic olfaction, IEEE Trans. Biomed. Eng. 57 (11) (2010) 2753–2763.

[57] U.N. Thakur, R. Bhardwaj, P.K. Ajmera, A. Hazra, ANN based approach for selective detection of breath acetone by using hybrid GO-FET sensor array, Eng. Res. Express 4 (2) (2022) 025008. Available from: ⟨https://iopscience.iop.org/article/10.1088/2631-8695/ac6487⟩.

[58] X. Li, X. Hu, A. Li, R. Kometani, I. Yamada, K. Sashida, et al., Identification of binary gases' mixtures from time-series resistance fluctuations: a sensitivity-controllable SnO2 gas sensor-based approach using 1D-CNN, Sens. Actuators A Phys. 349 (2023) 114070.

[59] B. Wang, J. Zhang, W. Li, Y. Zhang, T. Wang, Q. Lu, et al., Artificial olfaction based on tafel curve for quantitative detection of acetone ethanol gas mixture, Sens. Actuators B Chem. 377 (2023) 133049.

[60] H. Zhu, C. Liu, Y. Zheng, J. Zhao, L. Li, A Hybrid machine learning algorithm for detection of simulated expiratory markers of diabetic patients based on gas sensor array, IEEE Sens. J. (2022).

[61] BS EN ISO 15197:2015 - TC | 30 Jun 2015 | BSI Knowledge [Internet]. [cited 2023 Nov 4]. Available from: ⟨https://knowledge.bsigroup.com/products/in-vitro-diagnostic-test-systems-requirements-for-blood-glucose-monitoring-systems-for-self-testing-in-managing-diabetes-mellitus?version=tracked⟩.

[62] Mixing Alcohol with Your Diabetes.

[63] E. Mansour, R. Vishinkin, S. Rihet, W. Saliba, F. Fish, P. Sarfati, et al., Measurement of temperature and relative humidity in exhaled breath, Sens. Actuators B 304 (2020) 127371.

[64] A. Paleczek, A. Rydosz, Replication data for: the effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: laboratory, Stud. AGH Univ. Krakow (2024), https://doi.org/10.58032/AGH/S0CH7M.

[65] Jones A.W. The Relationship between Blood Alcohol Concentration (BAC) and Breath Alcohol Concentration (BrAC): A Review of the Evidence Forensic Blood Alcohol Calculations View project Theory and practice of forensic breath alcohol analysis View project. [cited 2023 Feb 3]; Available from: ⟨www.dft.gov.uk/pgr/roadsafety/research/rsrr⟩.

[66] A.P. Drummond-Lage, R.G. de Freitas, G. Cruz, L. Perillo, M.A. Paiva, A.J. A. Wainstein, Correlation between blood alcohol concentration (BAC), breath alcohol concentration (BrAC) and psychomotor evaluation in a clinical monitored study of alcohol intake in Brazil, Alcohol 66 (2018) 15–20. ⟨https://pubmed.ncbi.nlm.nih.gov/29277283/⟩ (Available from:).

[67] A. Kaisdotter Andersson, J. Kron, M. Castren, A. Muntlin Athlin, B. Hok, L. Wiklund, Assessment of the breath alcohol concentration in emergency care patients with different level of consciousness, Scand. J. Trauma Resusc. Emerg. Med. 23 (1) (2015) 1–9. Available from: ⟨https://sjtrem.biomedcentral.com/articles/10.1186/s13049-014-0082-y⟩.

[68] Concentration unit conversion | GASTEC CORPORATION [Internet]. [cited 2023 Feb 3]. Available from: ⟨https://www.gastec.co.jp/en/technology/knowledge/concentration/⟩.

[69] Buitinck L., Louppe G., Blondel M., Pedregosa F., Müller A.C., Grisel O., et al. API design for machine learning software: experiences from the scikit-learn project. 2013 Sep 1 [cited 2023 Mar 14]; Available from: ⟨https://arxiv.org/abs/1309.0238v1⟩.

[70] F. Pedregosa, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830. ⟨http://jmlr.org/papers/v12/pedregosa11a.html⟩ [cited 2023 Mar 14].

[71] Chen T., Guestrin C.. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. [cited 2023 Mar 14]; Available from: https://doi.org/10.1145/2939672.2939785.

[72] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., LightGBM: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process Syst. (2017) 30. ⟨https://github.com/Microsoft/LightGBM⟩ (Available from).

[73] Welcome to LightGBM's documentation! — LightGBM 3.3.5.99 documentation [Internet]. [cited 2023 Mar 14]. Available from: ⟨https://lightgbm.readthedocs.io/en/latest/index.html⟩.

[74] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process Syst. (2018) 31. ⟨https://github.com/catboost/catboost⟩ (Available from).

[75] CatBoost [Internet]. [cited 2023 Mar 14]. Available from: ⟨https://catboost.ai/en/docs/⟩.

[76] TGS1820 - Featured Products - FIGARO Engineering inc [Internet]. [cited 2023 Feb 5]. Available from: ⟨https://www.figaro.co.jp/en/product/feature/tgs1820.html⟩.

A. Paleczek is a PhD Student in field of Automatic, Electronics and Electrical Engineering. She graduated in Biomedical Engineering (major Computing and Electronics in Medicine) in 2021. Her master thesis was concerned on Development of algorithms for the detection biomarkers of diabetes in exhaled air. In 2020, she defended her engineering thesis entitled Recognition of sign language images using a neural network, in which she prepared a database of photos of 22 Polish Sign Language characters, designed a GUI and a convolutional neural network that automatically classified images obtained in real-time from a web camera. In 2021, she defended with distinction her master thesis "Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection".

A. Rydosz received his MSc and PhD degrees in electronics engineering from the AGH University of Science and Technology, Krakow, Poland in 2009 and 2014, respectively. In 2019, he received the DSc (habilitation) in automatic control and robotics, electrical engineering. His current research interests include gas sensors and micropreconcentrators, LTCC, MEMS technology, and gas sensors system applications. He is also interested in the PVD method of the fabrication of various sensing materials with special emphasis on the volatile organic compounds detection in exhaled human breath, for example, as a potential tool for noninvasive measurements of several diseases, such as diabetes. Since 2019 he works as a professor in the Department of Electronics AGH and serves as the Vice-chair of the Joint Chapter AP03/AES10/MTT17, IEEE Poland Section. He was awarded several awards for young researchers, including START 2015, START 2016 from the Foundation for Polish Science, Ministry of Science and Higher Education, and recently from the POLITYKA journal. He is a member of several societies, such as the Polish Vacuum Society, the Polish Sensor Society, the IMAPS, the 500 Innovators Society, Section of Microwaves and Radiolocation, Polish Academy of Science.

3.3. Prediction of diabetes state using artificial breath and e-nose system supported by machine learning

Prediction of diabetes state using artificial breath and e-nose system supported by machine learning

Anna Paleczek[1], Dominik Grochala[2] and Artur Rydosz[3], *Senior Member, IEEE*

[1,2,3]Institute of Electronics, Faculty of Computer Science, Electronics and Telecommunications,
AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Krakow, Poland

[1]paleczek@agh.edu.pl, [2]grochala@agh.edu.pl, [3]rydosz@agh.edu.pl

***Abstract*—Human breath contains volatile organic compounds that can be a source of information about the state of human health and can be used to detect diseases such as cancer and metabolic disorders, for example, diabetes. A well-known biomarker of diabetes is acetone. Due to the growing number of people with diabetes, with special emphasis on undiagnosed people, we have developed a system consisting of an array of sensors and machine learning algorithms developed to detect diabetes in human exhaled air. System tests were carried out on gas mixtures that mimic human breath. For multiclass classification tasks (healthy, prediabetes, and diabetes) CatBoost Classifier algorithm outperforms other tree-based machine learning algorithms. The result showed 86% accuracy in multiclass diabetes state prediction.**

Keywords—breath acetone; CatBoost; classification; diabetes; e-nose; machine learning; sensors

I. Introduction

Human breath consists largely of nitrogen (78%), oxygen (15%) and carbon dioxide (4%). A small part of the exhaled air is the dead space (end-tidal part) [1], which contains volatile organic compounds (VOCs). VOCs are present in the breath at very low concentrations of parts per million (ppm) and even parts per trillion (ppt), making them difficult to detect using e-noses with gas sensor arrays [2], [3]. Exogenous VOCs result from external factors, e.g. drug use, smoking or inhaled air pollution, while endogenous VOCs are produced as a result of metabolic processes taking place in the human body [4], [5], diseases such as diabetes [6]–[11], cancer [12], [13], asthma [14] etc. For example, acetone is known from literature studies as a diabetes biomarker [6]–[10]. The World Health Organization (WHO) [15] reports a continuously growing number of people with diagnosed and undiagnosed diabetes. In the case of diabetes, it is important to detect the disease quickly, as well as constant care and blood glucose measurements, which are currently performed mainly using invasive methods [16]–[19]. The development of a noninvasive method for determining blood glucose level (BGL) and detecting diabetes, e.g. based on the analysis of exhaled air, may increase the number of people diagnosed at an early stage of the disease and reduce related complications.

According to the literature review, there is a proven correlation between BGL and acetone concentration in exhaled breath. The review of the literature shows a concentration of acetone in the breath acetone concentration 0.3-0.9 ppm in healthy people 0.3-0.9 ppm, 0.9-1.8 ppm in the breath of prediabetes and more than 1.8 ppm in the breath of diabetes [20], [21]. Mansour *et al.* report relative humidity (RH) of human exhaled air in the range 41.9-91% [22].

Sarno *et al.* used five gas sensors supported by Deep Neural Networks to predict multilevel diabetes from patients based on exhaled breath. The authors achieved 96% accuracy, but used the same data set for testing and training [23].

Guo *et al.* proposed an e-nose system consisting of twelve gas sensors to predict the four-level diabetes stage (BLG levels). The experiments were carried out using 90 breath samples collected from people with diabetes. The authors performed prediction with linear polynomial, quadratic polynomial, and exponential function fitting. The best accuracy for each of the four levels, 75%, 65.31%, 65.31%, and 55%, was obtained with a linear model. In addition, they successfully classified diabetes and healthy samples using Support Vector Machines with accuracy greater than 92% on the test set [24].

In our previous work, we proposed the XGBoost algorithm to classify the state of diabetes based on gas mixtures and 1.5 ppm acetone as a threshold level between healthy and diabetes state. The algorithm achieved 99% binary classification accuracy, but in a narrower acetone range in gas mixtures [24], so a system with a wider acetone range was needed to best mimic human exhaled air in the case of diabetes. Here, we have improved accuracy by using more categories and different acetone ranges for predictions of each diabetes stage.

In this paper, we proposed an e-nose system with metal oxide semiconductor and electrochemical sensors combined with machine learning algorithms to classify the concentration of acetone in simulated human breath to predict the state of diabetes. This could be a preliminary step in developing a noninvasive device for diabetes detection, including prediabetic state.

The work was partially supported by the IDUB AGH 4122 grant and the statutory activity at the Institute of Electronics AGH

II. Methods

A. Measurement system

To simulate the breath of patients with diabetes and prediabetes, as well as healthy people, gas mixtures consisting of ethylbenzene, propane, CO_2 and acetone were prepared. The mixtures were dosed using six GF50 mass flow controllers (MKS, Andover, Massachusetts, USA). The water bubbler was used to simulate the relative humidity of the gas mixture to mimic human exhaled air. The measurement system is shown in Figure 1. The sensors used and their measurement ranges are listed in Table I. Acetone in the mixtures was in the range of 0-5.12 ppm, CO_2 in the range of 2.8-6.8%, propane in the range of 0-4.8 ppm, and ethylbenzene 0-2.6 ppm. The average RH of the mixtures was 75% to best mimic human breathing. Sensor signal acquisition was performed using DAQ970A (Keysight Technologies, Santa Rosa, CA, USA).
As input data from sensors, raw sensor responses R_G and sensitivity S values given by equation 1 were used.

$$S = R_g - R_0 \quad (1)$$

where:

- S – sensor sensitivity,
- R_g – sensor exposed to the gas mixture exposure,
- R_0 – sensor exposed to the synthetic air and RH exposure.

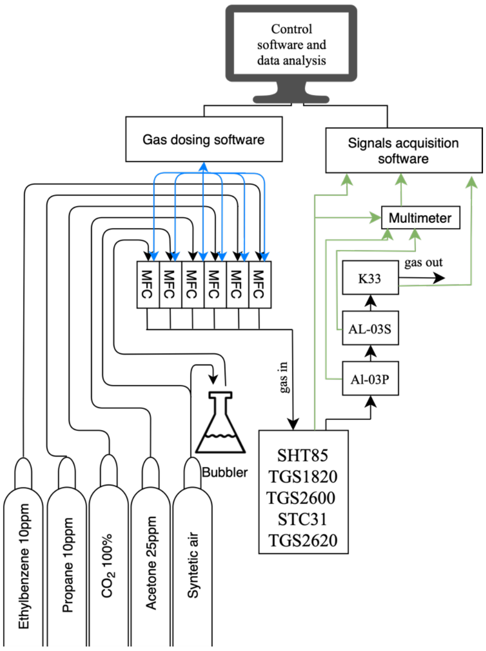


Fig. 1. The sketch of the measurement system.

TABLE I. Sensors used in measurement system

Sensor	*Target gas*	*Typical detection range*
TGS1820 [a.]	$(CH_3)_2CO$	1–20 ppm
TGS2620 [a.]	C_2H_5OH	50–5000 ppm
TGS2600 [a.]	Air Contaminants	1 ~ 30 ppm of H_2
AL-03S [b.]	C_2H_5OH	0 ~ 2.0 mg/L
AL-03P [b.]	C_2H_5OH	0 ~ 2.0 mg/L
K33 [c.]	CO_2	0-10%
STC31[d.]	CO_2	0-100%
SHT85 [d.]	RH	0-100%

[a.] Figaro Engineering Inc, Mino, Osaka, Japan
[b.] MGK SENSOR Co., Ltd., Saitama, Japan
[c.] Senseair, Delsbo, Sweden
[d.] Sensirion, Staefa ZH, Switzerland

B. Algorithm

CatBoost is an open-source supervised machine learning framework based on Gradient Boosted Decision Tree (GBDT) [25] commonly used in medical machine learning tasks [26], [27], especially with heterogeneous data. CatBoost supports categorical features and deals with gradient bias, which leads to reducing overfitting [28].
Duplicates from the data set were removed and then the data set was divided into training and test data in the 80:20 ratio. For algorithms development, scikit-learn Machine Learning in Python [29], [30] and XGBoost [31], LGBM [32], [33], CatBoost [25], [34] software libraries were used.

III. Results

Experiments were performed with several machine learning algorithms such as Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier, LGBM Classifier, and CatBoost Classifier. Algorithms achieved (on the test set) average accuracy: 81%, 83%, 83%, 77%, and 86%, respectively. CatBoost slightly outperforms other algorithms even in the case of homogeneous data. Detailed classification results are shown in Table II.

TABLE II. CatBoost Classifier results

Class	*Precision*	*Recall*	*F1-score*
Healthy	0.95	0.90	0.93
Prediabetes	0.79	0.88	0.83
Diabetes	0.88	0.78	0.82

The confusion matrix is shown in Figure 2. Some cases were incorrectly classified, but only within two adjacent categories, which means that the model did not misclassify diabetes with healthy case. Such a mistake could result in the person with diabetes being omitted and not referred for further examination. If the model got confused between prediabetes and diabetes categories, it would actually have less impact since both patients should be referred for further tests to clarify the disease status and treatment planning.

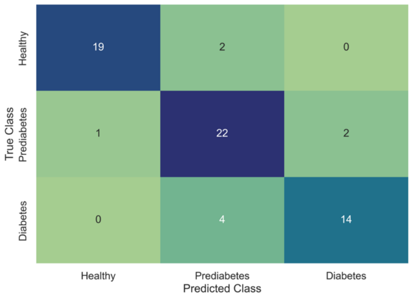


Fig. 2. Confusion matrix.

The feature importance analysis shows that the most important sensor for the multiclass classification of diabetes state was TGS1820. Less important were the data from gas sensors TGS2620, TGS2600, AL03-P, and AL03-S. The results show that the measured relative humidity was less important than the response of TGS1820, but higher than the response values of other sensors.

IV. Conclusions

The early diagnosis of diabetes plays a significant role in its further treatment and reduction of its complications. Therefore, it is important to conduct research and develop a device for non-invasive disease detection. The device based on breath measurements consists of gas sensors, but due to low concentrations of VOCs and poor selectivity of the sensors, it is necessary to use an array of sensors and machine learning algorithms. Most often, binary classifications of healthy/disease are carried out, while in the case of diabetes, it may be insufficient due to small differences in the ranges of acetone concentrations in exhaled air depending on the stage of the disease. The e-nose proposed in the paper enables multiclassification for healthy, prediabetic, and diabetes samples based on gas mixtures with high accuracy. The next step in the research on the development of noninvasive devices should be to test the system using human breaths and to develop algorithms to predict blood glucose levels that can be used both for the detection of diabetes and its subsequent monitoring.

References

[1] J. D. Pleil and A. B. Lindstrom, "Collection of a single alveolar exhaled breath for volatile organic compounds analysis," *Am J Ind Med*, vol. 28, no. 1, pp. 109–121, Jul. 1995, doi: 10.1002/AJIM.4700280110.

[2] H. Kiss *et al.*, "Exhaled Biomarkers for Point-of-Care Diagnosis: Recent Advances and New Challenges in Breathomics," *Micromachines 2023, Vol. 14, Page 391*, vol. 14, no. 2, p. 391, Feb. 2023, doi: 10.3390/MI14020391.

[3] Y. H. Ochoa-Muñoz, R. M. de Gutiérrez, and J. E. Rodríguez-Páez, "Metal Oxide Gas Sensors to Study Acetone Detection Considering Their Potential in the Diagnosis of Diabetes: A Review," *Molecules 2023, Vol. 28, Page 1150*, vol. 28, no. 3, p. 1150, Jan. 2023, doi: 10.3390/MOLECULES28031150.

[4] A. Sharma, R. Kumar, and P. Varadwaj, "Smelling the Disease: Diagnostic Potential of Breath Analysis," *Molecular Diagnosis & Therapy 2023*, pp. 1–27, Feb. 2023, doi: 10.1007/S40291-023-00640-7.

[5] B. Buszewski, M. Kesy, T. Ligor, and A. Amann, "Human exhaled air analytics: Biomarkers of diseases," *Biomedical Chromatography*, vol. 21, no. 6, pp. 553–566, Jun. 2007, doi: 10.1002/BMC.835.

[6] Z. Wang and C. Wang, "Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements," *J. Breath Res.*, vol. 7, no. 3, p. 037109, Sep. 2013, doi: 10.1088/1752-7155/7/3/037109.

[7] T. D. C. Minh, D. R. Blake, and P. R. Galassetti, "The clinical potential of exhaled breath analysis for diabetes mellitus," *Diabetes Res. Clin. Pract.*, vol. 97, no. 2, pp. 195–205, Aug. 2012, doi: 10.1016/j.diabres.2012.02.006.

[8] K. Yan and D. Zhang, "A novel breath analysis system for diabetes diagnosis," *ICCH 2012 Proceedings - International Conference on Computerized Healthcare*. IEEE Computer Society, pp. 166–70, 2012. doi: 10.1109/icch.2012.6724490.

[9] A. Rydosz, "Diabetes Without Needles: Non-invasive Diagnostics and Health Management," *Diabetes Without Needles: Non-invasive Diagnostics and Health Management*, pp. 1–302, Jan. 2022, doi: 10.1016/B978-0-323-99887-1.01001-3.

[10] G. Rooth and S. Ostenson, "Acetone in alveolar air, and the control of diabetes," *Lancet*, vol. 2, no. 7473, pp. 1102–5, 1966, doi: 10.1016/s0140-6736(66)92194-5.

[11] S. Neupane *et al.*, "Exhaled breath isoprene rises during hypoglycemia in type 1 diabetes," *Diabetes Care*, vol. 39, no. 7, pp. e97–e98, Jul. 2016, doi: 10.2337/dc16-0461.

[12] M. H. M. C. Scheepers, Z. Al-Difaie, L. Brandts, A. Peeters, B. van Grinsven, and N. D. Bouvy, "Diagnostic performance of electronic noses in cancer diagnoses using exhaled breath: a systematic review and meta-analysis," *JAMA Netw Open*, vol. 5, no. 6, p. E2219372, Jun. 2022, doi: 10.1001/jamanetworkopen.2022.19372.

[13] H. Yang *et al.*, "The investigation of volatile organic compounds in diagnosing (early) esophageal squamous cell carcinoma and gastric adenocarcinoma," *J Cancer Res Clin Oncol*, pp. 1–13, Mar. 2023, doi: 10.1007/S00432-023-04595-4/FIGURES/3.

[14] P. Xepapadaki *et al.*, "Utility of biomarkers in the diagnosis and monitoring of asthmatic children," *World Allergy Organization Journal*, vol. 16, no. 1, p. 100727, Jan. 2023, doi: 10.1016/J.WAOJOU.2022.100727.

[15] "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed Mar. 05, 2023).

[16] F. J. Carrasco-Sánchez, J. M. Fernández-Rodríguez, J. Ena, R. Gómez-Huelgas, and J. Carretero-Gómez, "Medical treatment of type 2 diabetes mellitus: recommendations of the diabetes, obesity and nutrition group of the spanish society of internal medicine," *Revista Clínica Española (English Edition)*, vol. 221, no. 2, pp. 101–8, Feb. 2021, doi: 10.1016/j.rceng.2020.06.009.

[17] A. D. Association, "2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2021," *Diabetes Care*, vol. 44, pp. S15–S33, Jan. 2021, doi: 10.2337/dc21-s002.

[18] S. H. Ley, O. Hamdy, V. Mohan, and F. B. Hu, "Prevention and management of type 2 diabetes: dietary components and nutritional strategies," *Lancet*, vol. 383, no. 9933, pp. 1999–2007, 2014, doi: 10.1016/s0140-6736(14)60613-9.

[19] A. D. Association, "Screening for diabetes," *Diabetes Care*, vol. 25, no. SUPPL. 1, pp. s21–s24, 2002, doi: 10.2337/diacare.25.2007.s21.

[20] C. Deng, J. Zhang, X. Yu, W. Zhang, and X. Zhang, "Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization," *J Chromatogr B Analyt Technol Biomed Life Sci*, vol. 810, no. 2, pp. 269–275, Oct. 2004, doi: 10.1016/J.JCHROMB.2004.08.013.

[21] D. Wang, "Investigation of Different Materials as Acetone Sensors for Application in Type-1 Diabetes Diagnosis," *Biomed J Sci Tech Res*, vol. 14, no. 5, Feb. 2019, doi: 10.26717/BJSTR.2019.14.002619.

[22] E. Mansour *et al.*, "Measurement of temperature and relative humidity in exhaled breath," *Sens. Actuators B*, vol. 304, p. 127371, Feb. 2020, doi: 10.1016/j.snb.2019.127371.

[23] R. Sarno, S. I. Sabilla, and D. R. Wijaya, "Electronic Nose for Detecting Multilevel Diabetes using Optimized Deep Neural Network," *Engineering Letters*, vol. 28, no. 1, 2020.

[24] D. Guo, D. Zhang, N. Li, L. Zhang, and J. Yang, "Diabetes identification and classification by means of a breath analysis system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6165 LNCS, pp. 52–63, 2010, doi: 10.1007/978-3-642-13923-9_6/COVER.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv Neural Inf Process Syst*, vol. 31, 2018, Accessed: Mar. 14, 2023. [Online]. Available: https://github.com/catboost/catboost

[26] P. B. Dash, J. Nayak, C. R. Kishore, M. Mishra, and B. Naik, "Efficient Ensemble Learning Based CatBoost Approach for Early-Stage Stroke Risk Prediction," *Smart Innovation, Systems and Technologies*, vol. 317, pp. 475–483, 2023, doi: 10.1007/978-981-19-6068-0_46/COVER.

[27] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, pp. 1–45, Dec. 2020, doi: 10.1186/S40537-020-00369-8/FIGURES/9.

[28] W. Chang, X. Wang, J. Yang, and T. Qin, "An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries," *Sensors (Basel)*, vol. 23, no. 4, p. 1811, Feb. 2023, doi: 10.3390/S23041811/S1.

[29] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," Sep. 2013, doi: 10.48550/arxiv.1309.0238.

[30] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Mar. 14, 2023. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[31] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, doi: 10.1145/2939672.

[32] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv Neural Inf Process Syst*, vol. 30, 2017, Accessed: Mar. 14, 2023. [Online]. Available: https://github.com/Microsoft/LightGBM.

[33] "Welcome to LightGBM's documentation! — LightGBM 3.3.5.99 documentation." https://lightgbm.readthedocs.io/en/latest/index.html (accessed Mar. 14, 2023).

[34] "CatBoost." https://catboost.ai/en/docs/ (accessed Mar. 14, 2023).

4. The exhaled breath analysis results obtained during the medical experiment

This chapter presents the results of analysis of breath samples collected during the medical experiment using an electronic nose system designed and optimised based on previous laboratory studies. The study was conducted in collaboration with the Department of Prosthodontics and Orthodontics, Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland, and with the approval of the Jagiellonian University Bioethical Committee (KBET: 1072.6120.40.2023). The study targeted participants aged 45 years or older to identify those at risk of developing metabolic syndrome features. The study participants completed a questionnaire regarding gender, weight, height, age, medications, past and present illnesses, and a questionnaire related to denture wear and dental cavities.

The study was conducted on the breath samples collected from 151 subjects, including 92 women and 59 men. The participants had a mean age of 67 years, with a standard deviation (SD) of 9.3 years, an average body weight of 77 kg (SD = 15.5), an average height of 166 cm (SD = 9.4), and an average body mass index (BMI) of 27.7 kg/m^2 (SD = 4.11), suggesting a tendency towards overweight. Additionally, capillary blood tests were performed, analysing metabolic parameters using test strips and dedicated diagnostic devices. The mean values of the obtained results are as follows:

- Glucose: 110.5 mg/dL (SD = 31.32)
- Uric acid: 5.66 mg/dL (SD = 1.49)
- Total cholesterol: 174.33 mg/dL (SD = 39.02)
- Triglycerides: 124.5 mg/dL (SD = 84.53)

These data indicate that the participants comprised a group with diverse metabolic parameters, with results ranging from normal to potentially elevated values.

These experiments aimed to test the practical application of the developed e-nose system for predicting metabolic and biochemical parameters based on analysis of exhaled air composition. This part of the project was especially important because it demonstrated how the system performs in real clinical settings and tested the ideas developed from the gas mixture experiments. Unlike the earlier stages, which used artificial gas mixtures, this stage utilised only breath samples from real subjects. Therefore, this study enables verification whether previously developed methods for measuring and analysing sensor data, as well as selected configurations of the e-nose system, could predict metabolic parameters based on exhaled breath analysis.

Particular attention was paid to compounds whose concentration in the body has significant diagnostic and preventive significance, such as cholesterol, glucose, and uric acid. These parameters are closely linked to metabolic disorders such as diabetes, hypercholesterolemia, and metabolic syndrome. Their non-invasive assessment can be a valuable tool for diagnosis and patient monitoring in daily clinical practice, given the high prevalence of metabolic disorders. Machine learning algorithms played a key role in analysing data from the medical experiment. Sensor signals measured from the e-nose require advanced processing to extract characteristic features and correlate them with reference results, such as biochemical blood tests. The analysis was performed using regression models, which enabled the prediction of selected health parameters based on breath samples.

The first research paper [AP6] on clinical research analysed the feasibility of predicting total cholesterol levels based on the composition of exhaled air. The study included 151 participants, from whom breath samples were collected and compared with capillary blood test results. Using machine learning, the sensor data were analysed to build a regression model for predicting cholesterol levels. Cholesterol prediction was performed using an e-nose composed of TGS1820, TGS2620, TGS2600, MQ3, Semeatech 7e4 NO2 and 7e4 H2S, SGX_NO2, SGX_H2S, K33, AL-03P, and AL-03S sensors supported by the Light Gradient Boosting Machine (LGBM) Regressor. The model achieved mean absolute percentage error (MAPE) values of 13.7% for the full measurement range and 8% within the norm range (≤200 mg/dL). Feature importance analysis highlighted TGS1820, AL-03P, TGS2620, MQ3 sensors as key contributors. The results demonstrate that gas sensors combined with machine learning enable non-invasive cholesterol estimation from breath. The results showed that the e-nose system can be used to predict total cholesterol levels based on exhaled breath analysis. This work highlights the system’s potential as a non-invasive tool for evaluating the risk of cardiovascular and metabolic disorders.

The second conference research paper [AP7] presented a preliminary evaluation of the system’s potential to predict additional biochemical parameters, such as glucose and uric acid levels, from exhaled breath samples. The results showed a link between the sensor signals and these biochemical markers, suggesting that the method could be developed for non-invasive monitoring of metabolic state. Using data from the three gas sensors included in the e-nose (TGS1820, AL-03P, K33) and LGBMRegressor, the mean absolute error was 19.32 mg/dl for glucose, 31.33 mg/dl for total cholesterol, and 1.43 mg/dl for uric acid. This study also complements earlier work by confirming that the system has applications beyond cholesterol analysis.

Both published papers were significant steps in the development of the e-nose system, starting from laboratory research to clinical validation. Results obtained from breath samples during the medical experiment demonstrated the ability to predict key metabolic parameters, including cholesterol, glucose, and uric acid, using non-invasive methods based on exhaled breath analysis. These studies showed the effectiveness of machine learning algorithms in analysing sensor data. They significantly complement previous laboratory studies and advance the practical application of the system in personalised medicine and the prevention of metabolic diseases. The presented results confirm that it is possible to use the e-nose system supported by machine learning algorithms to estimate health parameters such as blood glucose level, total cholesterol and uric acid level based on the exhaled breath analysis.

The results were presented by the Author at the *47th Annual International Conference of the IEEE Engineering in Medicine and Biology* - 14-17 July 2025, Copenhagen, Denmark.

4.1. Noninvasive Total Cholesterol Level Measurement Using an E-Nose System and Machine Learning on Exhaled Breath Samples

ACS SENSORS

Open Access

pubs.acs.org/acssensors

Article

Noninvasive Total Cholesterol Level Measurement Using an E-Nose System and Machine Learning on Exhaled Breath Samples

Published as part of ACS Sensors special issue "Breath Sensing".

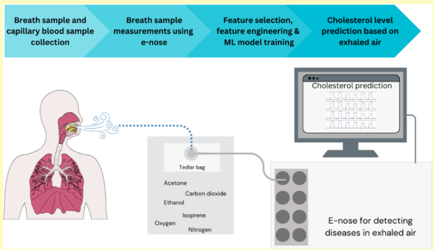
Anna Paleczek,* Justyna Grochala, Dominik Grochala, Jakub Słowik, Małgorzata Pihut, Jolanta E. Loster, and Artur Rydosz

Cite This: *ACS Sens.* 2024, 9, 6630−6637

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: In this paper, the first e-nose system coupled with machine learning algorithm for noninvasive measurement of total cholesterol level based on exhaled air sample was proposed. The study was conducted with the participation of 151 people, from whom a breath sample was collected, and the level of total cholesterol was measured. The breath sample was examined using e-nose and gas sensors, such as TGS1820, TGS2620, TGS2600, MQ3, Semeatech 7e4 NO2 and 7e4 H2S, SGX_NO2, SGX_H2S, K33, AL-03P, and AL-03S. The LGBMRegressor algorithm was used to predict cholesterol level based on the breath sample. Machine learning algorithms were developed for the entire measurement range and for the norm range ≤200 mg/dL achieving MAPE 13.7% and 8%, respectively. The results show that it is possible to develop a noninvasive device to measure total cholesterol level from breath.


Breath sample and capillary blood sample collection
Breath sample measurements using e-nose
Feature selection, feature engineering & ML model training
Cholesterol level prediction based on exhaled air
Cholesterol prediction
Tedlar bag
Acetone
Carbon dioxide
Ethanol
Isoprene
Oxygen
Nitrogen
E-nose for detecting diseases in exhaled air

KEYWORDS: *E-nose system, exhaled breath analysis, gas sensors, LGBMRegressor, machine learning, noninvasive measurement, predictive modeling, total cholesterol level*

■ VOLATILE ORGANIC COMPOUNDS

In recent times, researchers have been working to develop noninvasive methods for measuring and monitoring health parameters, such as blood glucose levels,[1−4] detection of FeNO for asthma and other respiratory diseases,[5,6] SIBO (small intestine bacterial overgrowth), or various types of cancers.[7−9] One possibility is to monitor exhaled breath and the volatile organic compounds (VOCs) contained in it. Human breath (Figure 1) consists mainly of nitrogen (78%−79%), oxygen (13%−16%), and carbon dioxide (4%).[10] The rest of the parts are mostly VOCs. Currently, over 3,000 different VOCs have been identified in breath.[11] Some of them may be of endogenous origin and may be biomarkers of various diseases or conditions of the human body and come from metabolic processes occurring in the body. Exogenous VOCs are the result of external factors, such as smoking, air pollution, or drug metabolism.[12] The relative humidity of human breath is 89%−97%.[13]

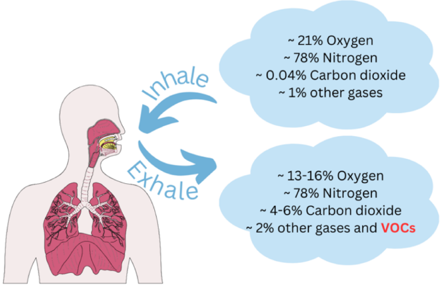
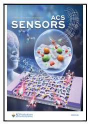


Figure 1. Composition of inhaled and exhaled air.

■ CLINICAL IMPORTANCE OF MONITORING BLOOD CHOLESTEROL LEVELS

Cholesterol performs an essential role in human metabolism and permits homeostatic regulation. It is a crucial component of every cell membrane.[14] As a steroid hormones' precursor,

Received: August 21, 2024
Revised: October 30, 2024
Accepted: November 6, 2024
Published: November 22, 2024



ACS Publications

© 2024 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acssensors.4c02198
ACS Sens. 2024, 9, 6630−6637

cholesterol is responsible for various immune, development and reproductive processes as well as mineral metabolism.[15] Despite this biological significance, hypercholesterolemia contributes to the pathogenesis of cardiovascular diseases (CVDs)—a leading cause of death worldwide. According to the WHO 17.9 million people die of CVDs every year.[16] Cholesterol can accumulate in the walls of arteries and form atheromatous plaques. After long asymptomatic period, plaque can rupture causing intravascular coagulation and ischemia. This phenomenon occurs particularly within the coronary, cerebral, and peripheral circulation, leading respectively to myocardial infarction, stroke, and limb ischemia. It is estimated that up to 90% of CVDs could be avoided by modifying risk factors.[17] Hypercholesterolemia is one of the most important modifiable risk factors for CVDs, so regular assessment of cholesterol levels and early implementation of appropriate treatment are valuable for patients. Although the clinical use of total cholesterol (TC) in relation to the LDL-cholesterol (LDL-C) is very limited, a linear correlation of TC levels with cardiovascular risk has been demonstrated.[18]

■ THE RELATIONSHIP BETWEEN BLOOD CHOLESTEROL AND VOCS

It is presumed that isoprene is formed during cholesterol biosynthesis in nucleated cells by nonenzymatic conversion of DMAPP. Thereafter, it enters the alveoli via the vascular system and is excreted with exhaled air. The metabolic pathway of cholesterol and its relationship to isoprene in breath[19−22] is shown in Figure 2.

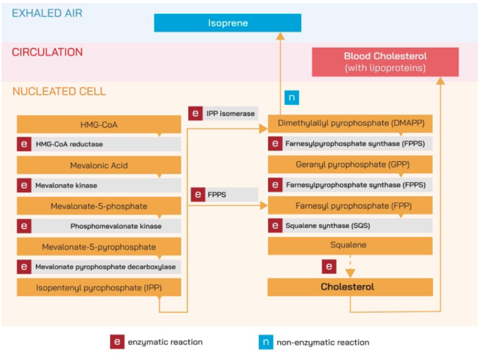

EXHALED AIR
Isoprene
CIRCULATION
Blood Cholesterol
(with lipoproteins)
NUCLEATED CELL
n
e IPP isomerase
HMG-CoA
e HMG-CoA reductase
Mevalonic Acid
e Mevalonate kinase
Mevalonate-5-phosphate
e Phosphomevalonate kinase
Mevalonate-5-pyrophosphate
e Mevalonate pyrophosphate decarboxylase
Isopentenyl pyrophosphate (IPP)
e FPPS
Dimethylallyl pyrophosphate (DMAPP)
e Farnesylpyrophosphate synthase (FPPS)
Geranyl pyrophosphate (GPP)
e Farnesylpyrophosphate synthase (FPPS)
Farnesyl pyrophosphate (FPP)
e Squalene synthase (SQS)
Squalene
e
Cholesterol
e enzymatic reaction
n non-enzymatic reaction

Figure 2. Metabolic pathway of cholesterol and its relationship to isoprene in breath.

■ GAS SENSING METHODS

VOCs in breath occur in concentrations of parts per million (ppm), parts per billion (ppb), or even parts per trillion (ppt); therefore, determining their concentration in exhaled air is difficult using the commercially available gas sensors. There are reference methods, such as gas chromatography coupled with mass spectrometry,[23,24] selected ion flow-tube mass spectrometry,[25] proton-transfer-reaction time-of-flight mass spectrometry,[26] which allow for the separation of gas mixtures into components and their quantitative analysis. The operation of such devices is complicated, they require special storage or long-term start-up procedures, and they are very expensive. Therefore, gas sensors for detecting low concentrations of compounds in gas mixtures have been widely developed. Because sensors can detect multiple substances and exhaled air contains numerous volatile organic compounds, employing a matrix of gas sensors and machine learning algorithms is essential to increase the sensitivity and selectivity of the e-nose systems.

■ BREATH SAMPLING METHODS

Most often, breath is collected in bags specially designed for this purpose, which maintain the initial concentration of compounds contained in the gas mixture for up to several days.[27−29] Such bags include Tedlar Bag, FlexFoil PLUS.[30] It is also possible to supply exhaled air directly to the device.[31,32] Currently, there is no standardized method for storing and collecting breath samples, which leads to problems with reproducing studies and comparing results with those of other researchers.

■ RELATED WORKS

In the related literature, the study of exhaled isoprene and its relationship with cholesterol concentration is often mentioned, but no studies using e-nose to estimate cholesterol from exhaled breath have been presented yet. Gouma et al. proposed a selective nanosensor for exhaled breath analysis, which can be used for noninvasive monitoring of cholesterol levels. They developed sensor arrays for measuring isoprene, carbon dioxide and ammonia gas, however the sensor was tested only on synthetic gases that were composed to mimic human exhaled air.[33] Similar research was conducted by Güntner et al., who developed a Ti-doped ZnO sensor for selective sensing of isoprene for breath diagnosis. This sensor showed a significantly higher response to isoprene than to acetone, ammonia, or ethanol at 90% RH, which is the observed RH of human breath. In this case the authors also tested the sensor only on synthetic gas mixtures.[34]

This paper introduces the first e-nose system combined with a machine learning algorithm for noninvasive measurement of total cholesterol levels using exhaled air samples. The study involved 151 participants from whom a breath sample was collected, and the level of total cholesterol was measured.

■ EXPERIMENTAL SECTION

Information About the Study Involving Human Participants. In collaboration with the Department of Prosthodontics and Orthodontics at the Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland, tests were conducted on breath samples and capillary blood samples collected from 151 individuals (Jagiellonian University bioethical committee approval KBET: 1072.6120.40.2023). The study included patients over the age of 45 to identify those at risk of developing features of metabolic syndrome.

Patients' Information. Each of the 151 participants completed a questionnaire that included questions about gender, weight, height, age, medications taken, past and current illnesses, and well-being related to the use of dentures and dental cavities. 92 women and 59 men participated in the study. Descriptive statistics of the sample population including data on participants' age, height, weight, and BMI are included in Table 1.

Table 1. Descriptive Statistics of the Sample Population

Parameter	Mean	Standard Deviation
Age	67	9.3
Weight	77 [kg]	15.5 [kg]
Height	166 [cm]	9.4 [cm]
BMI	27.7 [kg/m^2]	4.11 [kg/m^2]

Capillary Blood Tests. Participants in the study had their capillary blood samples analyzed by a physician using devices that measure parameters via the strip technique, such as

- Glucose (Accu-Chek Instant, Roche Diabetes Care GmbH, Sandhofer Strasse 116, 68305 Mannheim; www.roche.com.).
- Uric acid (PEMPA 3in1 device, General Life Biotechnology Co., Ltd. 5F., No. 240, Shinshu Rd., Shin Juang Dist., New Taipei City 242, Taiwan; www.BeneCheck.com.tw.).
- Cholesterol (PEMPA 3in1 device, General Life Biotechnology Co., Ltd. 5F., No. 240, Shinshu Rd., Shin Juang Dist., New Taipei City 242, Taiwan; www.BeneCheck.com.tw.).
- Triglycerides (Accutrend Plus, Accutrend Glucose, Roche Diagnostics GmbH, Sandhofer Strasse 116, 68305 Mannheim; www.roche.com).

Descriptive statistics of blood test parameters, including data on measured values of glucose, uric acid, cholesterol, and triglycerides from capillary blood of the participants, are included in Table 2.

Table 2. Descriptive Statistics of Blood Test Parameters

Parameter	Mean	Standard Deviation
Glucose	110.5 [mg/dL]	31.32 [mg/dL]
Uric acid	5.66 [mg/dL]	1.49 [mg/dL]
Cholesterol	174.33 [mg/dL]	39.02 [mg/dL]
Triglycerides	124.5 [mg/dL]	84.53 [mg/dL]

Cholesterol Levels Distribution. In this paper, we focus on predicting cholesterol levels based on exhaled air measurements. The PEMPA 3-in-1 device allows cholesterol to be measured from fresh capillary blood in the range of 100−400 mg/dL (2.59−10.35 mmol/L). With this test, the norm is a result of ≤200 mg/dL (5.17 mmol/L).[35] The distribution of cholesterol values measured in the study participants is presented in Figure 3.

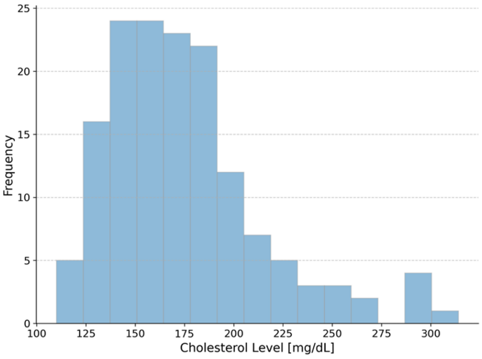


Figure 3. Distribution of cholesterol levels among the study participants.

Breath Tests. Breath samples were collected in Tedlar bags and analyzed using an electronic nose (e-nose) twice. Tedlar bags are specialized bags for collecting and storing breath samples. Their advantage is maintaining high concentrations of the collected substances, which allows the bags to be transported to the external laboratory and to cooperate with remote research centers or hospitals.[28,36−38] However, the e-nose system that we propose is portable and allows for quick testing of the sample in a hospital or medical center (Figure 4).

E-Nose System. The e-nose comprised a system for pumping air from the bags and a set of sensors, including TGS1820, TGS2620, and TGS2600 (Figaro Engineering Inc., Mino, Osaka, Japan), MQ3 (Winsen, ZhengZhou, HeNan, China), 7e4 NO2, 7e4 H2S (SemeaTech, Los Angeles, USA and Shanghai, China), SGX_NO2, SGX_H2S (SGX SENSORTECH, Switzerland), K33 (Senseair, Delsbo, Sweden), and AL-03P, AL-03S (MGK SENSOR Co., Ltd., Saitama, Japan).

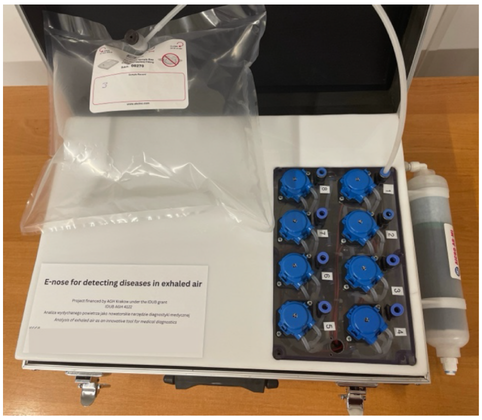


Figure 4. E-nose system used during measurements.

Sensors' Responses. As part of the study, the breath sample collected from each patient in a Tedlar bag was measured twice using the prepared e-nose system. The time of rinsing with ambient air collected through the filter was 10 min between subsequent measurements, and the time of air injection from the bag was 15 min. For each measurement, the R_A (sensor response to purge gas) and R_G (sensor response to breath sample) values were determined (as shown in Figure 5) and the responses of the S and S_1 sensors were calculated (eqs 1 and 2).

$$S = R_G - R_A \quad (1)$$

$$S = \frac{R_G}{R_A} \quad (2)$$

Gas sensor data typically consist of electrical values affected by measurement errors, noise, or drift[39,40] due to changes in sensor layer properties. These quality issues can impact model training and performance, so researchers use signal processing techniques like

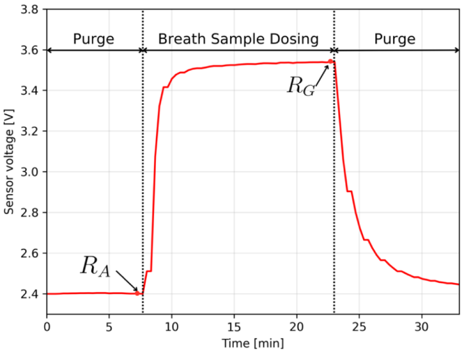


Figure 5. Stages of the breath sample measurement using the developed e-nose system.

filtering[41,42] and baseline normalization,[39,43,44] to prepare the data for next steps in the processing pipeline. In our solution, we used a mean filter[41] to reduce noise, calculating an average of 10 samples, while calculating the sensor response (eqs 1 and 2), which takes into account the baseline value (in the sensor response in the purge stage), allows to minimize the influence of drift. Additionally, before testing using human breaths, the sensors were tested on synthetic mixtures and results were published in our previous papers.[43,45]

Outliers Handling. Based on the sensor responses, four outliers were removed for the K33 (CO_2 sensor) and AL-03P (ethanol sensor) sensors. Measurements were removed where the K33 sensor measured a CO_2 value lower than 2%, which means that the breath sample was incorrectly collected, and measurements where the AL-03P sensor indicated a response indicating the presence of ethanol in exhaled air, which could come from the mouthwash.

Train Test Split. The data set was divided into training and test sets in a ratio of 90:10 so that both measured values of the breath sample of one patient were located in only one of the sets. The training set included breath sample measurements collected from 136 patients, and the test set included 15 patients. This means that when two measurements from each patient were used, the training and test sets included 272 and 30 samples, respectively.

Machine Learning Algorithms. The aim of the study was to develop an algorithm that would allow prediction of cholesterol concentration in blood using e-nose and breath sample. The R_G, S, and S_1 data from sensors available in e-nose and BMI were taken as features. For this purpose, machine learning algorithms were used for the regression problem. The study tested machine learning algorithms: linear regression, lasso regression, ridge regression, random forest, LGBM regressor, XGB regressor, CatBoost regressor, KNN regressor, and neural networks. The results for all algorithms were compared (Table 3) and the best results were obtained using LGBM regressor. For each algorithm, the hyperparameter space for searching was determined. The best hyperparameters were determined using the RandomSearchCV[46,47] method from the scikit-learn library (30 splits, negative mean absolute error optimization)

Table 3. Comparison of Machine Learning Algorithm Performance (Measured as Mean Absolute Error) in Total Cholesterol Level Prediction (Norm Range)

Algorithm	Mean absolute error
Linear Regression	17.02
Lasso Regression	20.82
Ridge Regression	16.43
Random Forest	17.11
LightGBM Regressor	12.94
XGBoost Regressor	19.41
CatBoost Regressor	16.84
KNN Regressor	19.11

LightGBM Regressor Model. LightGBM is a gradient boosting framework that employs tree-based learning algorithms designed for distribution and efficiency. It offers several key benefits including faster training speed, higher efficiency, and lower memory usage. Additionally, it provides better accuracy and supports parallel, distributed, and GPU-based learning, making it capable of handling large-scale data sets effectively.[48] Linear regression models, lasso, and ridge, assume linear relationships between variables, which is a major limitation in the case of sensors' data processing. LGBM, like random forest, CatBoost, and XGBRegressor, is a tree model that can better handle nonlinear relationships in the data.[49] LightGBM handles large numbers of features very well, which can lead to more accurate predictions, even when other models may struggle to maintain performance. LightGBM has parameters that allow for overfitting control (e.g., max_depth, num_leaves, and feature_fraction). This makes it easy to tune to generalize well to the data, which is an advantage oversimpler models, such as linear regression, that have limited overfitting control. Additionally, LGBMRegressor has built-in function for feature importance calculation and analysis.[50]

Metrics. The following metrics were used to evaluate the effectiveness of regression algorithms: mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and R^2 coefficient.

RESULTS AND DISCUSSION

Cholesterol Level Distribution Analysis. The distribution of measured cholesterol levels in the patients is previously shown in Figure 3. The analysis of the histogram and the values of mean (174.33), median (166.0), and calculated skewness index (1.18) shows that the distribution of cholesterol level among the patients participating in the study is right-skewed (the skewness coefficient is greater than 0 and the mean is greater than the median). Twenty-six patients had a score above 200 mg/dL (norm result) and only 7 above 260 mg/dL.

Considering the aforementioned problem, we decided to train two separate algorithms.

Prediction of cholesterol level in the entire range.

Prediction of cholesterol level within the norm (≤200 mg/dL).

Additionally, the predicted value logarithm technique was used to limit the influence of skewness.[51] For prediction over the full range, we obtained better results using only one measurement for each patient. The results for both cases are compared in Table 4.

Table 4. Comparison of Metrics for the Entire Range and Norm Range Prediction using LGBM Regressor

Metric	Entire range	Norm range
MAE	21.2	12.9
RMSE	26.4	15.8
R^2	0.22	0.52
MAPE	13.7%	8%

Prediction of Cholesterol Level in the Entire Range. On average, the predicted cholesterol levels deviate from the actual values by about 21.22 mg/dL. The R-squared value indicates how well the model explains the variance in the target variable. An R^2 of 0.224 means that the model explains about 22.4% of the variance in cholesterol levels, which is relatively low. A MAPE of 13.73% means that, on average, the model's predictions are about 13.73% off from the actual cholesterol levels. A comparison of the values predicted by the machine learning algorithm based on breath sample testing and the values measured using the test strip and capillary blood is shown in Figure 6.

Prediction of Cholesterol Level within the Norm. The performance metrics obtained from the prediction model for total cholesterol levels in the norm range based on exhaled air are indicative of a quite successful model. On average, the predicted cholesterol levels deviate from the actual values by approximately 12.94 mg/dL. This RMSE value indicates that the typical prediction error is around 15.79 mg/dL, providing a more substantial penalty for larger errors. The R-squared (R^2) value of 0.522 signifies that the model explains about 52.2% of the variance in the cholesterol levels, which is moderately good but also highlights that there is room for improvement. Additionally, the model's predictions are, on average, within

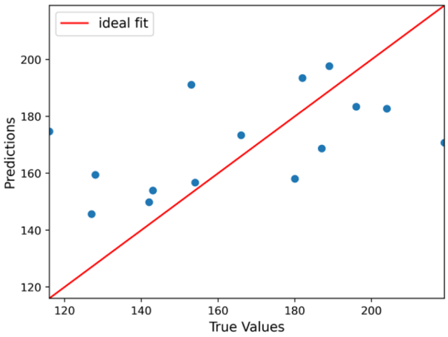


Figure 6. Results of prediction of the total cholesterol level in the entire range.

7.99% of the actual values. These results suggest that while the model has a reasonable predictive capability, further refinement, additional features, and additional data could enhance its accuracy and reliability. A comparison of the values predicted (in norm range) by the LGBMRegressor algorithm based on breath sample testing and the values measured using the test strip and capillary blood is shown in Figure 7.

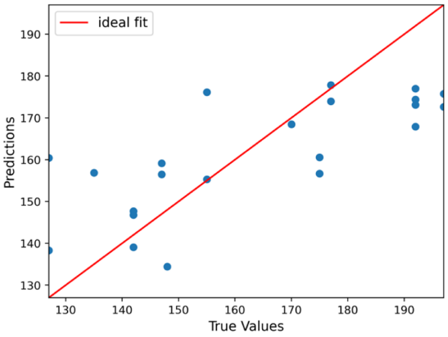


Figure 7. Results of prediction of the total cholesterol level in the norm range.

The Bland— Altman plot analysis provides further insight into the agreement between the predicted and actual total cholesterol levels. The mean difference (or bias) between the predictions and the actual measurements is 2.42 mg/dL. This small mean difference indicates that, on average, the model slightly overestimates the cholesterol levels by 2.42 mg/dL. The limits of agreement (LOA) are defined as the mean difference plus and minus 1.96 times the standard deviation of the differences. The upper LOA is 33.00 mg/dL, and the lower LOA is −28.15 mg/dL. This range suggests that 95% of the differences between the predicted and actual cholesterol levels fall within this interval. The Bland-Altman plot is illustrated in Figure 8.

Features Importance. Analysis of the most important features showed that the most important for prediction were the responses of the TGS1820, AL-03P, TGS2620, and MQ3 sensors. These are mainly sensors for acetone, ethanol, and VOCs. Isoprene, which is most often observed as a cholesterol biomarker in breath, is also a volatile organic compound and can be detected by semiconductor sensors, such as TGS1820 or TGS2620. Gas sensors, especially those based on metal oxides (e.g., SnO_2), operate on the principle of electrical conductivity change in the presence of volatile organic compounds. Isoprene, being a VOC, can cause a change in conductivity similar to that of acetone or ethanol. Due to the cross-selectivity of sensors and the large number of VOCs in exhaled air, it is necessary to use a gas sensor matrix and machine learning algorithms.

CONCLUSIONS

In this paper, we proposed the first e-nose for prediction of total cholesterol concentration in blood based on the exhaled breath analysis. Machine learning algorithms were developed for the entire measurement range and for the norm range ≤200 mg/dL achieving MAPE 13.7% and 8%, respectively. These are the first results allowing further development of the solution and achieving better results. One of the limitations of our study was that only 151 people participated in the study, which is a good introduction to research, while a larger population would improve the results. Total cholesterol level values observed in patients have a right skewed distribution and a small number of people achieved results above the norm, which was difficult for the model to generalize; however, the results in the norm range, where the number of patients was higher, show that such prediction is possible, and it is possible to achieve smaller errors with a larger population. One of the disadvantages of our study is that as a method of determining total cholesterol level in blood, we adopted a portable device for a capillary blood test strip and not measurements from venous blood performed in a professional laboratory with venous blood samples. Measurements with such a device are also burdened with measurement errors. Studies and reports show that the mean absolute relative difference of the five cholesterol self-tests ranged from 6 ± 5% (Accutrend Plus) to 20 ± 12% (Mylan Mytest).[52,53] Our study included people who fasted before the test and those who fasted after a meal. Studies show that there are no clinically significant differences in the level of total cholesterol in the blood after fasting and after a meal.[54] Our method copes with both cases.

Human breath is composed of many compounds that reflect the state of the body but also affect the response of sensors and the prediction of algorithms. Factors that can distort the results include external air pollution,[12,55] smoking, drinking, or eating immediately before the test. In addition, medications taken or other co-occurring diseases also have an impact. Often, when patients have metabolic syndrome[56] (as was the case in our studies), a simultaneous increase in blood parameters such as cholesterol, blood glucose level or triglycerides is observed. Therefore, it is important to collect additional data about patients, as well as to determine the patient's behavior before the test, just as is done with standard blood tests.

The next stages of the study development are the development of a portable device that would allow for broader screening of patients in various medical centers and comparison of results with total cholesterol determined in venous blood. One of the possibilities is also the study of additional parameters such as LDL-C and HDL-C levels and an attempt to predict them based on breathing. In summary, our study and the developed e-nose with machine learning algorithms provide a good basis for further research on a larger

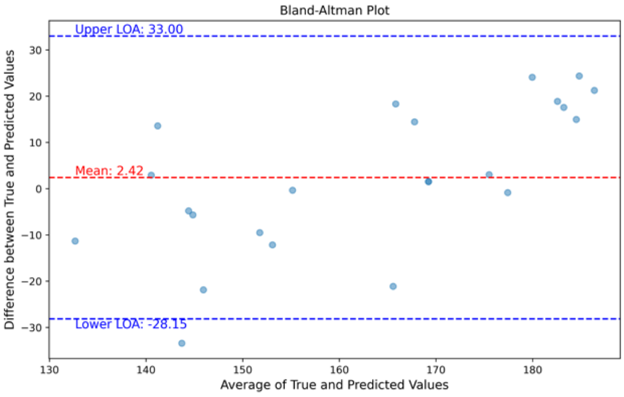


Figure 8. Bland-Altman plot of predictions of total cholesterol level in the norm range.

population and the development of a portable device for noninvasive prediction of total cholesterol, HDL-C and LDL-C levels based on a breath sample.

■ AUTHOR INFORMATION

Corresponding Author

Anna Paleczek − *AGH University of Krakow, Faculty of Computer Science Electronics and Telecommunications, Institute of Electronics, Krakow 30-059, Poland;* orcid.org/0000-0002-1467-3017; Email: paleczek@agh.edu.pl

Authors

Justyna Grochala − *Department of Prosthodontics and Orthodontics, Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Kraków 31-008, Poland*

Dominik Grochala − *AGH University of Krakow, Faculty of Computer Science Electronics and Telecommunications, Institute of Electronics, Krakow 30-059, Poland*

Jakub Słowik − *AGH University of Krakow, Faculty of Computer Science Electronics and Telecommunications, Institute of Electronics, Krakow 30-059, Poland; University Clinical Hospital in Opole, Institute of Medical Sciences, University of Opole, Opole 46-020, Poland*

Małgorzata Pihut − *Department of Prosthodontics and Orthodontics, Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Kraków 31-008, Poland*

Jolanta E. Loster − *Professor Loster's Orthodontics, Private practice, Faculty of Medicine, Jagiellonian University Medical College, Krakow 30-433, Poland*

Artur Rydosz − *AGH University of Krakow, Faculty of Computer Science Electronics and Telecommunications, Institute of Electronics, Krakow 30-059, Poland; The University Hospital in Krakow, Laboratory of Functional and Virtual Medical 3D Imaging [3D-vFMi(maging)/3D-FM], Krakow 30-688, Poland;* orcid.org/0000-0002-9148-1094

Complete contact information is available at: https://pubs.acs.org/10.1021/acssensors.4c02198

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. A.P. worked on conceptualization, methodology, investigation, formal analysis, and data curation, software, visualization, writing of original original draft and review and editing. J.G. conducted conceptualization, methodology, and investigation. D.G. performed conceptualization, methodology, and investigation. J.S. helped in writing of the original draft. M.P. conducted conceptualization, methodology, project administration, and supervision. J.E.L. worked on conceptualization, methodology, project administration, and supervision. A.R. performed project administration, conceptualization, methodology, investigation, writing of the original draft, review, and editing, supervision, funding acquisition

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The local ethical committee approved the study protocol (Jagiellonian University KBET: 1072.6120.40.2023). The experiments were conducted in accordance with the Declaration of Helsinki, principles of Good Clinical Practice, and written informed consent was obtained from all participants before participation. The work was supported by the IDUB AGH 4122 grant and the statutory activity at the Institute of Electronics AGH.

■ ABBREVIATIONS

CVDs, cardiovascular diseases; HDL-C, HDL-cholesterol; LDL-C, LDL-cholesterol; MAE, mean absolute error; MAPE, mean absolute percentage error; ppb, parts per billion; ppm, parts per million; ppt, parts per trillion; RH, relative humidity; RMSE, root mean square error; TC, total cholesterol; VOCs, volatile organic compounds

■ REFERENCES

(1) Wang, Z.; Wang, C. Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements. *J. Breath Res.* **2013**, 7 (3), 037109.

(2) Neupane, S.; Peverall, R.; Richmond, G.; Blaikie, T. P. J.; Taylor, D.; Hancock, G.; et al. Exhaled breath isoprene rises during

hypoglycemia in type 1 diabetes. *Diabetes Care* **2016**, *39* (7), No. e97−8.

(3) Rydosz, A. Diabetes Without Needles: Non-invasive Diagnostics and Health Management [Internet]. *Diabetes Without Needles: Non-invasive Diagnostics and Health Management. Elsevier*; 2022, 1−302. http://www.sciencedirect.com:5070/book/9780323998871/diabetes-without-needles.

(4) Paleczek, A.; Rydosz, A. Review of the algorithms used in exhaled breath analysis for the detection of diabetes. *J. Breath Res.* **2022**, *16* (2), 026003.

(5) Guida, G.; Carriero, V.; Bertolini, F.; Pizzimenti, S.; Heffler, E.; Paoletti, G.; et al. Exhaled nitric oxide in asthma: from diagnosis to management. *Curr. Opin. Allergy Clin. Immunol.* **2023**, *23* (1), 29−35.

(6) Xepapadaki, P.; Adachi, Y.; Pozo Beltrán, C. F.; El-Sayed, Z. A.; Gómez, R. M.; Hossny, E.; et al. Utility of biomarkers in the diagnosis and monitoring of asthmatic children. *World Allergy Organization J.* **2023**, *16* (1), 100727.

(7) Politi, L.; Monasta, L.; Rigressi, M. N.; Princivalle, A.; Gonfiotti, A.; Camiciottoli, G.; Perbellini, L.; et al. Discriminant Profiles of Volatile Compounds in the Alveolar Air of Patients with Squamous Cell Lung Cancer, Lung Adenocarcinoma or Colon Cancer. *Molecules* **2021**, *26* (3), Page 550.

(8) Ratiu, I. A.; Ligor, T.; Bocos-Bintintan, V.; Mayhew, C. A.; Buszewski, B. Volatile organic compounds in exhaled breath as fingerprints of lung cancer, asthma and COPD. *J. Clin. Med.* **2021**, *10* (1), 32.

(9) Chung, J.; Akter, S.; Han, S.; Shin, Y.; Choi, T. G.; Kang, I.; Kim, S.; et al. Diagnosis by Volatile Organic Compounds in Exhaled Breath from Patients with Gastric and Colorectal Cancers. *IJMS* **2023**, *24* (1), 129.

(10) Tortora, G.; Derrickson, B.. *Principles of Anatomy and Physiology*. 15th ed. [Internet] ed., Vol. *53*, John Wiley & Sons, Inc; 2013, pp. 1689−1699.

(11) Smolinska, A.; Klaassen, E. M. M.; Dallinga, J. W.; Van De Kant, K. D. G.; Jobsis, Q.; Moonen, E. J. C.; et al. Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. *PLoS One* **2014**, *9* (4), 2022.

(12) Longo, V.; Forleo, A.; Ferramosca, A.; Notari, T.; Pappalardo, S.; Siciliano, P.; et al. Blood, urine and semen volatile organic compound (VOC) pattern analysis for assessing health environmental impact in highly polluted areas in Italy. *Environ. Pollut.* **2021**, *286*, 117410.

(13) Ferrus, L.; Guenard, H.; Vardon, G.; Varene, P. Respiratory water loss. *Respir Physiol.* **1980**, *39* (3), 367−381.

(14) Paukner, K.; Lesná, I. K.; Poledne, R. Cholesterol in the Cell Membrane—An Emerging Player in Atherogenesis. *Int. J. Mol. Sci.* **2022**, *23* (1), 533.

(15) Schade, D. S.; Shey, L.; Eaton, R. P. Cholesterol Review: A Metabolically Important Molecule. *Endocr Pract.* **2020**, *26* (12), 1514−1523.

(16) Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

(17) McGill, H. C.; McMahan, C. A.; Gidding, S. S. Preventing heart disease in the 21st century: Implications of the pathobiological determinants of atherosclerosis in youth (PDAY) study. *Circulation* **2008**, *117* (9), 1216−1227.

(18) Jung, E.; Kong, S. Y.; Ro, Y. S.; Ryu, H. H.; Shin, S. D. Serum Cholesterol Levels and Risk of Cardiovascular Death: A Systematic Review and a Dose-Response Meta-Analysis of Prospective Cohort Studies. *Int. J. Environ. Res. Public Health* **2022**, *19* (14), 8272.

(19) Deneris, E. S.; Stein, R. A.; Mead, J. F. In vitro biosynthesis of isoprene from mevalonate utilizing a rat liver cytosolic fraction. *Biochem Biophys Res. Commun.* **1984**, *123* (2), 691−696.

(20) Kushch, I.; Arendacká, B.; štolc, S.; Mochalski, P.; Filipiak, W.; Schwarz, K.; et al. Breath isoprene–aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study. *Clin Chem. Lab Med.* **2008**, *46* (7), 1011−1018.

(21) Stone, B. G.; Besse, T. J.; Duane, W. C.; Dean Evans, C.; DeMaster, E. G. Effect of regulating cholesterol biosynthesis on breath isoprene excretion in men. *Lipids* **1993**, *28* (8), 705−708.

(22) Sitaula, S.; Burris, T. P.. Encyclopedia of Cell BiologyBradshaw, R. A.; Stahl, P. D., Eds. Vol. *3*. *Encyclopedia of Cell Biology*; Elsevier Ltd., 2016.

(23) Oliveira, L. F. D.; Mallafré-Muro, C.; Giner, J.; Perea, L.; Sibila, O.; Pardo, A.; Marco, S.; et al. Breath analysis using electronic nose and gas chromatography-mass spectrometry: A pilot study on bronchial infections in bronchiectasis. *Clin. Chim. Acta.* **2022**, *526*, 6−13.

(24) Deng, C.; Zhang, J.; Yu, X.; Zhang, W.; Zhang, X. Determination of acetone in human breath by gas chromatography-mass spectrometry and solid-phase microextraction with on-fiber derivatization. *J. Chromatogr B Analyt Technol. Biomed Life Sci.* **2004**, *810* (2), 269−275.

(25) Markar, S. R.; Chin, S. T.; Romano, A.; Wiggins, T.; Antonowicz, S.; Paraskeva, P.; et al. Breath Volatile Organic Compound Profiling of Colorectal Cancer Using Selected Ion Flow-tube Mass Spectrometry. *Ann. Surg.* **2019**, *269* (5), 903−910.

(26) Jung, Y. J.; Seo, H. S.; Kim, J. H.; Song, K. Y.; Park, C. H.; Lee, H. H. Advanced Diagnostic Technology of Volatile Organic Compounds Real Time analysis Analysis From Exhaled Breath of Gastric Cancer Patients Using Proton-Transfer-Reaction Time-of-Flight Mass Spectrometry. *Front. Oncol.* **2021**, *11*, 1368.

(27) Gilchrist, F. J.; Razavi, C.; Webb, A. K.; Jones, A. M.; Španěl, P.; Smith, D.; et al. An investigation of suitable bag materials for the collection and storage of breath samples containing hydrogen cyanide. *J. Breath Res.* **2012**, *6* (3), 036004.

(28) Steeghs, M. M. L.; Cristescu, S. M.; Harren, F. J. M.; Woollam, M.; Angarita-Rivera, P.; et al. The suitability of Tedlar bags for breath sampling in medical diagnostic research. *Physiol. Meas.* **2006**, *28* (1), 73.

(29) Ghimenti, S.; Lomonaco, T.; Bellagambi, F. G.; Tabucchi, S.; Onor, M.; Trivella, M. G.; et al. Comparison of sampling bags for the analysis of volatile organic compounds in breath. *J. Breath Res.* **2015**, *9* (4), 047110.

(30) Dharmawardana, N.; Goddard, T.; Woods, C.; Watson, D. I.; Ooi, E. H.; Yazbeck, R. Development of a non-invasive exhaled breath test for the diagnosis of head and neck cancer. *Br. J. Cancer* **2020**, *123* (12), 1775−1781.

(31) Vicent-Claramunt, A.; Naujalis, E. Cheap and easy human breath collection system for trace volatile organic compounds screening using thermal desorption − gas chromatography mass spectrometry. *MethodsX* **2021**, *8*, 101386.

(32) Hariyanto, S. R.; Wijaya, D. R. Detection of diabetes from gas analysis of human breath using e-nose. In *Vols. 2018-January, Proceedings of the 11th International Conference on Information and Communication Technology and System, ICTS 2017*; Institute of Electrical and Electronics Engineers Inc., 2017, pp. 241−246.

(33) Gouma, P.; Prasad, A.; Stanacevic, S. A selective nanosensor device for exhaled breath analysis. *J. Breath Res.* **2011**, *5* (3), 037110.

(34) Güntner, A. T.; Pineau, N. J.; Chie, D.; Krumeich, F.; Pratsinis, S. E. Selective sensing of isoprene by Ti-doped ZnO for breath diagnostics. *J. Mater. Chem. B* **2016**, *4* (32), 5358−5366.

(35) 3in1 device − PEMPA. https://pempa.pl/urzadzenie-3w1/.

(36) Mochalski, P.; King, J.; Unterkofler, K.; Amann, A. Stability of selected volatile breath constituents in Tedlar, Kynar and Flexfilm sampling bags. *Analyst* **2013**, *138* (5), 1405−1418.

(37) McGarvey, L. J.; Shorten, C. V. The Effects of Adsorption on the Reusability of Tedlar® Air Sampling Bags. *AIHAJ. - American Industrial Hygiene Association* **2000**, *61* (3), 375−380.

(38) Beauchamp, J.; Herbig, J.; Gutmann, R.; Hansel, A. On the use of Tedlar® bags for breath-gas sampling and analysis. *J. Breath Res.* **2008**, *2* (4), 046001.

(39) Zuppa, M.; Distante, C.; Persaud, K. C.; Siciliano, P. Recovery of drifting sensor responses by means of DWT analysis. *Sens. Actuators, B* **2007**, *120* (2), 411−416.

(40) Dennler, N.; Rastogi, S.; Fonollosa, J.; van Schaik, A.; Schmuker, M. Drift in a popular metal oxide sensor dataset reveals limitations for gas classification benchmarks. *Sens. Actuators, B* **2022**, *361*, 131668.

(41) Liu, L.; Li, W.; He, Z. C.; Chen, W.; Liu, H.; Chen, K.; et al. Detection of lung cancer with electronic nose using a novel ensemble learning framework. *J. Breath Res.* **2021**, *15* (2), 026014.

(42) Polaka, I.; Bhandari, M. P.; Mezmale, L.; Anarkulova, L.; Veliks, V.; Sivins, A. Modular Point-of-Care Breath Analyzer and Shape Taxonomy-Based Machine Learning for Gastric Cancer Detection. *Diagnostics* **2022**, *12* (2), 491.

(43) Paleczek, A.; Grochala, D.; Rydosz, A. Artificial breath classification using XGBoost algorithm for diabetes detection. *Sensors* **2021**, *21* (12), 4187.

(44) Binson, V. A.; Subramoniam, M.; Sunny, Y.; Mathew, L. Prediction of Pulmonary Diseases with Electronic Nose Using SVM and XGBoost. *IEEE Sens. J.* **2021**, *21* (18), 20886−20895.

(45) Paleczek, A.; Rydosz, A. The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies. *Sens. Actuators, B* **2024**, *408*, 135550.

(46) Pedregosa, F.; Michel, V.; Grisel Oliviergrisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; et al. Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12* (85), 2825−2830.

(47) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Müller, A. C.; Grisel, O.et al. *API Design For Machine Learning Software: experiences From The Scikit-Learn Project*. 2013.

(48) Welcome to LightGBM's documentation! — LightGBM 4.0.0 documentation. https://lightgbm.readthedocs.io/en/stable/.

(49) Cherkassky, V.; Ma, Y. Comparison of model selection for regression. *Neural Comput.* **2003**, *15* (7), 1691−1714.

(50) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

(51) Hammouri, H. M.; Sabo, R. T.; Alsaadawi, R.; Kheirallah, K. A. Handling Skewed Data: A Comparison of Two Popular Methods. *Appl. Sci.* **2020**, *10* (18), 6247.

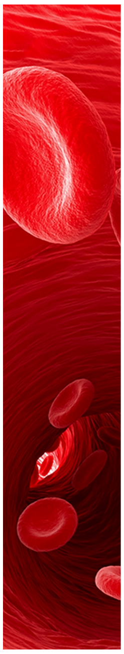
(52) Kurstjens, S.; Gemen, E.; Walk, S.; Njo, T.; Krabbe, J.; Gijzen, K.; et al. Performance of commercially-available cholesterol self-tests. *Ann. Clin Biochem.* **2021**, *58* (4), 289−296.

(53) Suklan, J.; Mutepfa, C.; Dickinson, R.; Hicks, T.; Williams, C.; Nick, H. E.et al. Point of care testing for cholesterol measuring: A rapid review and presentation of the scientific evidence Supporting primary care in the prevention and management of cardiovascular disease.

(54) Craig, S. R.; Amin, R. V.; Russell, D. W.; Paradise, N. F. Blood cholesterol screening: Influence of fasting state on cholesterol results and management decisions. *J. Gen Intern Med.* **2000**, *15* (6), 395−399.

(55) Blanchet, L.; Smolinska, A.; Baranska, A.; Tigchelaar, E.; Swertz, M.; Zhernakova, A.; et al. Factors that influence the volatile organic compound content in human breath. *J. Breath Res.* **2017**, *11* (1), 016013.

(56) Han, T. S.; Lean, M. E. J. Metabolic syndrome. *Medicine* **2015**, *43* (2), 80−87.



4.2. Revolutionizing Health Monitoring: A Three-Gas Sensor System Powered by Machine Learning for Predicting Cholesterol, Glucose, and Uric Acid Levels from Exhaled Breath

Revolutionizing Health Monitoring: A Three-Gas Sensor System Powered by Machine Learning for Predicting Cholesterol, Glucose, and Uric Acid Levels from Exhaled Breath

A. Paleczek, J. Grochala, D. Grochala, J. Słowik, M. Pihut, J. E. Loster and A. Rydosz

Abstract— **Nowadays, more and more people are struggling with elevated levels of total cholesterol, glucose and uric acid. In this study, we developed an electronic nose and examined the patients' breaths with it. Using three gas sensors from our e-nose and machine learning algorithms, we managed to achieve mean absolute error of 19.32 mg/dl, 31.33 mg/dl, 1.43 mg/dl in the prediction of glucose, total cholesterol, uric acid, respectively.**

Clinical Relevance—**This study highlights the potential for using a portable, non-invasive electronic nose to predict glucose, cholesterol, and uric acid levels from breath, enabling early detection of metabolic disorders and improving patient outcomes through timely intervention.**

I. Introduction

Exhaled breath analysis is being studied as a non-invasive method to diagnose and monitor diseases by detecting VOCs like isoprene (linked to heart disease) and acetone (indicating diabetes). While current methods are accurate, they are costly. Researchers aim to improve gas sensors and use machine learning to make detection more affordable and accessible for early disease detection [1]. This study aims to test the feasibility of estimating parameters like glucose, cholesterol, and uric acid from exhaled air using three gas sensors.

II. Methods

The study conducted at the Department of Prosthodontics and Orthodontics at the Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland involved 151 patients who were asked to inflate a Tedlar® bag and had health parameters measured from capillary blood. Based on previous studies [2], 3 sensors were identified as having high significance for cholesterol estimation - TGS1820 (Figaro Engineering Inc, Japan), AL-03P (MGK SENSOR Co., Ltd., Japan), K33 (Senseair, Sweden). Due to outliers in the response of the TGS1820 sensor, the final number of patients included in this study was 123. The participants in the study had an average: glucose level of 110.36 mg/dL, triglycerides 124.63 mg/dL, total cholesterol 170.70 mg/dL, BMI 27.73 kg/m^2, age 67 years. For parameters estimation, the response of three sensors, BMI and machine learning algorithms with optimized hyperparameters were used. The data was divided into training and test sets.

III. Results

For predicting each parameter, the best results were achieved by LGBMRegressor. Results are shown in Table I.

TABLE I. Results of best performing ML algorithms

Target parameter	Algorithm	Results	
		MAE [mg/dL]	*MAPE [%]*
Glucose	LGBMRegressor	19.32	15.58
Total cholesterol	LGBMRegressor	31.33	19.74
Uric acid	LGBMRegressor	1.43	26.56

MAE – mean absolute error, MAPE – mean absolute percentage error.

IV. Discussion & Conclusion

The results from three sensors show promise for developing a portable, non-invasive device for early detection of metabolic disorders, enabling quicker treatment and prevention of complications like cardiovascular diseases or diabetes. Next steps involve testing a larger group, using venous blood, and exploring breath-based detection of other health parameters (e.g. HDL, LDL, Beta-Hydroxybutyrate).

Acknowledgment

The local ethical committee approved the study protocol (Jagiellonian University KBET: 1072.6120.40.2023). The experiments were conducted in accordance with the Declaration of Helsinki, principles of Good Clinical Practice and written informed consent was obtained from all participants before participation.

References

[1] Rydosz, Artur. Diabetes Without Needles: Non-invasive Diagnostics and Health Management. Academic Press, 2022.

[2] Paleczek, Anna, et al. "Noninvasive total cholesterol level measurement using an E-Nose system and machine learning on exhaled breath samples." ACS sensors 9.12 (2024): 6630-6637

* The work was supported by the IDUB AGH 4122 grant and the statutory activity at the Institute of Electronics AGH.

A.Paleczek is with the AGH University of Krakow, al. A. Mickiewicza 30 Krakow, 30-059, Poland, paleczek@agh.edu.pl

J.Grochala is with the Department of Prosthodontics and Orthodontics, Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland, justyna.lemejda@doctoral.uj.edu.pl

D.Grochala is with the AGH University of Krakow, al. A. Mickiewicza 30 Krakow, 30-059, Poland, grochala@agh.edu.pl

J.Słowik is with the AGH University of Krakow, al. A. Mickiewicza 30 Krakow, 30-059, Poland, jakubslowik@student.agh.edu.pl

M.Pihut is with the Department of Prosthodontics and Orthodontics, Dental Institute, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland, malgorzata.pihut@uj.edu.pl

J.E.Loster is with the Professor Loster's Orthodontics, Private practice, Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland, jolanta.loster@uj.edu.pl

A.Rydosz is with the AGH University of Krakow, al. A. Mickiewicza 30 Krakow, 30-059, Poland, rydosz@agh.edu.pl and The University Hospital in Krakow, Laboratory of Functional and Virtual Medical 3D Imaging [3D-vFM], Jakubowskiego 2 Street, 30-688 Krakow, Poland

5. Summary and Conclusions

5.1. Summary of the Dissertation and Research Achievements

Based on the obtained results presented in this dissertation as well as in the reference papers [AP1, AP2, AP3, AP4, AP5, AP6, AP7] it can be stated that the research hypothesis was confirmed. Although, further clinical trials would be required to implement the developed approach in the daily clinical practice.

Chapter 2 presents the current state of knowledge regarding the use of exhaled air analysis in disease diagnosis. The importance of this method as a non-invasive approach is discussed, which can be an alternative or complement to traditional laboratory tests. This chapter presents the diseases currently being investigated for diagnostics using e-nose and machine learning systems, as well as those for which this method has already been approved for clinical use. Particular attention is paid to the role of selected biomarkers, such as acetone, in the diagnosis of diabetes and monitoring metabolic health, which formed the basis for planning further research.

Chapter 3 presents the next stage of the work: laboratory studies. This stage involved research on sensors suitable for detecting acetone in artificial gas mixtures designed to mimic exhaled air in various patient health states, including those containing influencing factors. At this stage of the research, effective classification of gaseous samples containing acetone, compensation for interference resulting from the presence of ethanol, and classification of samples into three classes (healthy, pre-diabetic, and diabetic) were achieved.

Chapter 4 presents the results of the initial clinical validation of the e-nose system supported by machine learning algorithms. Based on the experience and results from laboratory studies, clinical trials were conducted to predict cholesterol, glucose, and uric acid levels. The study was conducted on a group of 151 participants, confirming the method's potential in medical practice.

Summary of the research achievements:

- Designing and testing the e-nose system capable of analysing artificially prepared gas mixtures and clinical samples [AP3, AP4, AP5].
- Developing experimental protocols based on a literature review simulating various biomarker concentrations and the presence of additional influence factors (e.g., ethanol) [AP3, AP4, AP5].
- Selecting and testing multiple machine learning algorithms, both for disease classification and for predicting biomarker concentrations (regression) [AP3, AP4, AP5].
- Effective classification of artificial breath samples into three groups (healthy, pre-diabetic, diabetic) in laboratory conditions [AP5].
- Predicting acetone concentrations by taking into account inference factors present in mixtures, which allowed for the quantitative prediction of biomarkers [AP4].
- Conducting a study on a group of 151 patients, which allowed for the first evaluation of the method's performance in a clinical setting [AP6].
- Using data from the e-nose system and machine learning algorithms to predict cholesterol, glucose, and uric acid concentrations from breath samples [AP7].

5.2. Conclusions

The conducted studies have shown that the e-nose system, supported by machine learning algorithms, is an effective tool for analysing exhaled breath in the context of detecting metabolic diseases.

In both laboratory studies and medical experiments, it was possible to classify and distinguish participants' health states and quantitatively predict selected biomarkers, such as acetone, as well as predict parameters related to the assessment of patients’ metabolic health, such as cholesterol, glucose, and uric acid.

Studies conducted by the Author have shown that the e-nose system, combined with machine learning algorithms, enables noninvasive detection of selected metabolic parameters and health status based on exhaled breath analysis. In studies on diabetes detection, the system successfully classified acetone concentrations in synthetic gas samples, achieving 99% accuracy, 100% sensitivity, and 97.9% specificity for the XGBoost algorithm. XGBoost achieved a 0.245 ppm mean absolute error in the prediction of acetone concentration, while CatBoost achieved an error of 0.568 ppm in mixtures with high-ethanol content. Extended studies presented classification of samples into three health groups (healthy, pre-diabetic, and diabetic individuals), achieving precision of 95%, 79%, and 88%, respectively.

In human studies aimed to predict total cholesterol levels using the e-nose and LGBM Regressor, the MAPE 13.7% was achieved for the full measurement range and 8% within the normal range (≤200 mg/dL). Key sensors (TGS1820, AL-03P, TGS2620, and MQ3) for the prediction were identified by feature importance analysis. Preliminary studies on glucose, cholesterol and uric acid prediction demonstrated the ability to estimate these parameters with mean errors of 19.32 mg/dL, 31.33 mg/dL, and 1.43 mg/dL, respectively, using data from only three gas sensors (TGS1820, AL-03P, and K33).

These results confirm that the e-nose, supported by machine learning algorithms, can effectively predict both the levels of selected metabolites and the patient's health status, opening prospects for noninvasive monitoring of patient metabolic health and the diagnosis of metabolic diseases.

The tested machine learning methods also demonstrated the potential for compensating for the effects of interfering substances, such as ethanol, in artificial breath mixtures, which represents a significant step towards the practical application of this technology. Preliminary results from the initial clinical trials confirm the potential of the e-nose system as a non-invasive diagnostic method, highlighting the need for further development to support metabolic monitoring and early diagnosis.

Despite promising results, the current study has several limitations.

First, the initial clinical trials were conducted at a single medical centre with a small group of 151 participants, which limited the generalizability of the results. Therefore, it is crucial to expand the study to include a larger group of patients from various medical centres and provide detailed information about their health and medications.

Secondly, there was no direct comparison with gas analysis reference methods that would allow for the unambiguous identification of all relevant biomarkers. Further research will also require comparison of results with advanced reference techniques, such as GC/MS, PTR-MS, or SIFT-MS, to identify and quantitatively analyse biomarkers.

Thirdly, the study focused primarily on several metabolic parameters, including total cholesterol, glucose, and uric acid. Furthermore, the reference method for these measurements was capillary blood testing using portable point-of-care strip tests. This measurement is also subject to greater measurement error than venous blood testing in a laboratory.

Furthermore, biological variability between patients and the lack of a standardised breath sampling protocol are significant sources of potential uncertainty in the results and should have been considered. The protocol should include the appropriate preparation of the patient before the examination, such as fasting and refraining from smoking for at least 2 hours prior to the examination. Literature studies report various protocols and methods for breath collection. In some cases, the patient was asked to exhale as much air as possible to capture the end-tidal portion of the breath. At the same time, in other studies, capnometers are used to obtain a sample containing the maximum carbon dioxide concentration and consequently the highest biomarker concentrations. Some studies also include instructions in the protocol for the patient to hold their breath for a specified period before exhaling directly into the device or a special bag, such as a Tedlar® bag. The lack of measurement standardisation makes it difficult to compare study results across different medical centres and research teams.

These limitations indicate the need for further research involving larger and more diverse populations, standardisation of breath sampling and storage procedures, and the use of reference techniques to verify the system's performance.

5.3. Future Work and Perspectives

Analysis of the study's limitations enables the identification of several directions for future research on using an e-nose system supported by machine learning algorithms for disease detection and health monitoring based on exhaled air analysis.

Further research on monitoring metabolic syndrome and diabetes will require expanding diagnostics to include additional biomarkers, such as Low-Density Lipoprotein (LDL), High-Density Lipoprotein (HDL), triglycerides, and inflammatory markers, as well as conducting blood tests using an accurate method, i.e., venous blood sampling. One possible development direction for the conducted research also involves applying the e-nose system and ML to other diseases, such as cancer (lung cancer, colon cancer), neurodegenerative diseases (Alzheimer's, Parkinson's), and infectious diseases (COVID-19, tuberculosis).

Research on both metabolic syndrome and the detection of other diseases requires the development of standardised breath sampling procedures, as well as studies on larger and more diverse patient populations in multicenter, randomised clinical trials. A key aspect will also be comparing e-nose results with reference techniques (GC/MS, PTR-MS, SIFT-MS), which will enable the identification of biomarkers responsible for the studied metabolic states and the identification of interfering factors from other diseases, diet, or individual patient variability.

The study of other sensors and their selection for specific biomarkers or disease entities is also important. The use of Explainable AI algorithms may be beneficial, as they enhance the interpretability of results in clinical practice, thereby increasing patient and physician confidence in the diagnosis.

The design of a detailed patient questionnaire and access to their medical data might also be crucial. Integrating multimodal data, such as e-nose signals, clinical data, and patient demographics, will enable the analysis of the impact of medications, diet, and other diseases on breath test results. This will facilitate the creation of more universal algorithms that achieve high performance regardless of inference factors.

Another aspect that should be considered is the technological development and implementation of the e-nose in clinical practice and everyday use by patients. To this end, it is necessary to conduct research on system miniaturisation, including analysing the impact of sensor responses on model decisions and identifying the most important ones. One aspect of the device's implementation is its potential integration with telemedicine systems, allowing for real-time monitoring (continuous assessment of the patient's health).

Bibliography

[1] B. Buszewski, M. Kesy, T. Ligor, and A. Amann, ‘Human exhaled air analytics: Biomarkers of diseases’, *Biomedical Chromatography*, vol. 21, no. 6, pp. 553–566, Jun. 2007, doi: 10.1002/BMC.835.

[2] N. M. Mule and D. D. Patil, ‘A deep learning approach for chronic obstructive pulmonary disease diagnosis from human exhaled breath gases.’, *Majlesi Journal of Electrical Engineering*, vol. 19, no. 2, pp. 1–12, Jun. 2025, doi: 10.57647/J.MJEE.2025.1902.32.

[3] L. Li *et al.*, ‘Identifying potential breath biomarkers for early diagnosis of papillary thyroid cancer based on solid-phase microextraction gas chromatography-high resolution mass spectrometry with metabolomics’, *Metabolomics*, vol. 20, no. 3, pp. 1–13, Jun. 2024, doi: 10.1007/S11306-024-02119-W/FIGURES/6.

[4] T. C. Setlhare, A. G. Mpolokang, E. Flahaut, and G. Chimowa, ‘Determination of lung cancer exhaled breath biomarkers using machine learning-a new analysis framework’, *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–12, Jul. 2025, doi: 10.1038/s41598-025-11365-4.

[5] K. Schwarz *et al.*, ‘Breath acetone—aspects of normal physiology related to age and gender as determined in a PTR-MS study’, *J. Breath Res.*, vol. 3, no. 2, p. 027003, 2009, doi: 10.1088/1752-7155/3/2/027003.

[6] M. Sun *et al.*, ‘Continuous Monitoring of Breath Acetone, Blood Glucose and Blood Ketone in 20 Type 1 Diabetic Outpatients Over 30 Days’, *J Anal Bioanal Tech*, vol. 8, no. 5, pp. 1–8, Oct. 2017, doi: 10.4172/2155-9872.1000386.

[7] N. Nelson, V. Lagesson, A. R. Nosratabadi, J. Ludvigsson, and C. Tagesson, ‘Exhaled isoprene and acetone in newborn infants and in children with diabetes mellitus’, *Pediatr. Res.*, vol. 44, no. 3, pp. 363–7, 1998, doi: 10.1203/00006450-199809000-00016.

[8] L. R. Narasimhan, W. Goodman, and C. K. N. Patel, ‘Correlation of breath ammonia with blood urea nitrogen and creatinine during hemodialysis’, *Proc Natl Acad Sci U S A*, vol. 98, no. 8, pp. 4617–4621, Apr. 2001, doi: 10.1073/PNAS.071057598.

[9] Q. Jing *et al.*, ‘Ultrasensitive Chemiresistive Gas Sensor Can Diagnose Asthma and Monitor Its Severity by Analyzing Its Biomarker H2S: An Experimental, Clinical, and Theoretical Study’, *ACS Sens*, vol. 7, no. 8, pp. 2243–2252, Aug. 2022, doi: 10.1021/ACSSENSORS.2C00737/ASSET/IMAGES/MEDIUM/SE2C00737_M008.GIF.

[10] F. Monedeiro, M. Monedeiro-Milanowski, I. A. Ratiu, B. Brożek, T. Ligor, and B. Buszewski, ‘Needle Trap Device-GC-MS for Characterization of Lung Diseases Based on Breath VOC Profiles’, *Molecules 2021, Vol. 26, Page 1789*, vol. 26, no. 6, p. 1789, Mar. 2021, doi: 10.3390/MOLECULES26061789.

[11] M. MacIel, S. Sankari, M. Woollam, and M. Agarwal, ‘Optimization of Metal Oxide Nanosensors and Development of a Feature Extraction Algorithm to Analyze VOC Profiles in Exhaled Breath’, *IEEE Sens J*, vol. 23, no. 15, pp. 16571–16578, Aug. 2023, doi: 10.1109/JSEN.2023.3288968.

[12] G. Guida *et al.*, ‘Exhaled nitric oxide in asthma: from diagnosis to management’, *Curr Opin Allergy Clin Immunol*, vol. 23, no. 1, pp. 29–35, Feb. 2023, doi: 10.1097/ACI.0000000000000877.

[13] L. Loewenthal and A. Menzies-Gow, ‘FeNO in Asthma’, *Semin Respir Crit Care Med*, vol. 43, no. 5, pp. 635–645, Oct. 2022, doi: 10.1055/S-0042-1743290/ID/JR220504-25/BIB.

[14] L. Robles and R. Priefer, ‘Lactose Intolerance: What Your Breath Can Tell You’, *Diagnostics 2020, Vol. 10, Page 412*, vol. 10, no. 6, p. 412, Jun. 2020, doi: 10.3390/DIAGNOSTICS10060412.

[15] A. D. Association, ‘2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021’, *Diabetes Care*, vol. 44, no. Supplement 1, pp. S15–S33, Jan. 2021, doi: 10.2337/DC21-S002.

[16] F. J. Carrasco-Sánchez, J. M. Fernández-Rodríguez, J. Ena, R. Gómez-Huelgas, and J. Carretero-Gómez, ‘Medical treatment of type 2 diabetes mellitus: recommendations of the diabetes, obesity and nutrition group of the spanish society of internal medicine’, *Revista Clínica Española (English Edition)*, vol. 221, no. 2, pp. 101–8, Feb. 2021, doi: 10.1016/j.rceng.2020.06.009.

[17] ‘Diabetes’. Accessed: Jan. 22, 2023. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

[18] A. Paleczek *et al.*, ‘Noninvasive Total Cholesterol Level Measurement Using an E-Nose System and Machine Learning on Exhaled Breath Samples’, *ACS Sens*, Nov. 2024, doi: 10.1021/ACSSENSORS.4C02198.

[19] D. Marzorati *et al.*, ‘Metal-oxide gas sensors for exhaled-breath analysis: a review’, *Meas Sci Technol*, vol. 32, no. 10, p. 102004, Jul. 2021, doi: 10.1088/1361-6501/AC03E3.

[20] A. M. Dhanush Gowda, A. D. Dessai, and U. Y. Nayak, ‘Electronic-Nose Technology for Lung Cancer Detection: A Non-Invasive Diagnostic Revolution’, *Lung*, vol. 203, no. 1, pp. 1–19, Dec. 2025, doi: 10.1007/S00408-025-00828-0/TABLES/2.

[21] D. Yang, R. A. Gopal, T. Lkhagvaa, and D. Choi, ‘Metal-oxide gas sensors for exhaled-breath analysis: a review’, *Meas Sci Technol*, vol. 32, no. 10, p. 102004, Jul. 2021, doi: 10.1088/1361-6501/AC03E3.

[22] N. R. Subawickrama Mallika Widanaarachchige, A. Paul, I. K. Banga, A. Bhide, S. Muthukumar, and S. Prasad, ‘Advancements in Breathomics: Special Focus on Electrochemical Sensing and AI for Chronic Disease Diagnosis and Monitoring’, *ACS Omega*, vol. 10, no. 5, pp. 4187–4196, Feb. 2025, doi: 10.1021/ACSOMEGA.4C10008/ASSET/IMAGES/LARGE/AO4C10008_0006.JPEG.

[23] L. L. Liu, S. P. Morgan, R. Correia, and S. Korposh, ‘A single-film fiber optical sensor for simultaneous measurement of carbon dioxide and relative humidity’, *Opt Laser Technol*, vol. 147, p. 107696, Mar. 2022, doi: 10.1016/J.OPTLASTEC.2021.107696.

[24] G. B. Monteiro Fernandes, H. Nascimento, R. M. Santa Cruz, J. L. Brum Marques, and C. da Silva Moreira, ‘Investigation of a Plasmonic Optical Sensor for Acetone Detection in Exhaled Breath and Exhaled Breath Condensate’, *Plasmonics*, vol. 19, no. 5, pp. 2527–2535, Oct. 2024, doi: 10.1007/S11468-023-02190-4/TABLES/2.

[25] R. Zhu *et al.*, ‘Optical chemical gas sensor based on spectral autocorrelation: A method for online detection of nitric oxide and ammonia in exhaled breath’, *Sens Actuators B Chem*, vol. 422, p. 136694, Jan. 2025, doi: 10.1016/J.SNB.2024.136694.

[26] R. A. Sola Martínez *et al.*, ‘Data preprocessing workflow for exhaled breath analysis by GC/MS using open sources’, *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/S41598-020-79014-6.

[27] J. Glöckler *et al.*, ‘Infrared Spectroscopic Electronic Noses: An Innovative Approach for Exhaled Breath Sensing’, *ACS Sens*, vol. 10, no. 1, pp. 427–438, Jan. 2025, doi: 10.1021/ACSSENSORS.4C02725/ASSET/IMAGES/LARGE/SE4C02725_0011.JPEG.

[28] A. A. Bunaciu and H. Y. Aboul-Enein, ‘Breath analysis using FTIR spectroscopy’, *Open Exploration 2019 6:*, vol. 6, pp. 1001308-, Apr. 2025, doi: 10.37349/EMED.2025.1001308.

[29] Z. Jia, W. Q. Ong, F. Zhang, F. Du, V. Thavasi, and V. Thirumalai, ‘A study of 9 common breath VOCs in 504 healthy subjects using PTR-TOF-MS’, *Metabolomics*, vol. 20, no. 4, pp. 1–9, Aug. 2024, doi: 10.1007/S11306-024-02139-6/TABLES/5.

[30] D. Smith, P. Španěl, N. Demarais, V. S. Langford, and M. J. McEwan, ‘Recent developments and applications of selected ion flow tube mass spectrometry (SIFT-MS)’, *Mass Spectrom Rev*, vol. 44, no. 2, pp. 101–134, Mar. 2025, doi: 10.1002/MAS.21835.

[31] F. Röck, N. Barsan, and U. Weimar, ‘Electronic nose: Current status and future trends’, *Chem Rev*, vol. 108, no. 2, pp. 705–725, Feb. 2008, doi: 10.1021/CR068121Q/ASSET/IMAGES/LARGE/CR068121QF00016.JPEG.

[32] J. Sorocki and A. Rydosz, ‘A Prototype of a Portable Gas Analyzer for Exhaled Acetone Detection’, *Applied Sciences 2019, Vol. 9, Page 2605*, vol. 9, no. 13, p. 2605, Jun. 2019, doi: 10.3390/APP9132605.

[33] A. Kononov *et al.*, ‘Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer’, *J Breath Res*, vol. 14, no. 1, p. 016004, Oct. 2019, doi: 10.1088/1752-7163/AB433D.

[34] H. Dong *et al.*, ‘Online Accurate Detection of Breath Acetone Using Metal Oxide Semiconductor Gas Sensor and Diffusive Gas Separation’, *Front Bioeng Biotechnol*, vol. 10, p. 296, Mar. 2022, doi: 10.3389/FBIOE.2022.861950/BIBTEX.

[35] S. Ghimenti *et al.*, ‘Comparison of sampling bags for the analysis of volatile organic compounds in breath’, *J. Breath Res.*, vol. 9, no. 4, p. 047110, Dec. 2015, doi: 10.1088/1752-7155/9/4/047110.

[36] B. Czippelová *et al.*, ‘Impact of breath sample collection method and length of storage of breath samples in Tedlar bags on the level of selected volatiles assessed using gas chromatography-ion mobility spectrometry (GC-IMS)’, *J Breath Res*, vol. 18, no. 3, p. 036004, May 2024, doi: 10.1088/1752-7163/AD4736.

[37] P. Mochalski, J. King, K. Unterkofler, and A. Amann, ‘Stability of selected volatile breath constituents in Tedlar, Kynar and Flexfilm sampling bags’, *Analyst*, vol. 138, no. 5, pp. 1405–18, Mar. 2013, doi: 10.1039/c2an36193k.

[38] S. Ghimenti *et al.*, ‘Comparison of sampling bags for the analysis of volatile organic compounds in breath’, *J. Breath Res.*, vol. 9, no. 4, p. 047110, Dec. 2015, doi: 10.1088/1752-7155/9/4/047110.

[39] J. Beauchamp, J. Herbig, R. Gutmann, and A. Hansel, ‘On the use of Tedlar® bags for breath-gas sampling and analysis’, *J. Breath Res.*, vol. 2, no. 4, p. 046001, 2008, doi: 10.1088/1752-7155/2/4/046001.

[40] 'ReCIVA® Breath Sampler to Collect a Breath Sample'. Accessed: Sep. 20, 2025. [Online]. Available: https://www.owlstonemedical.com/products/reciva/

[41] M. Zuppa, C. Distante, K. C. Persaud, and P. Siciliano, 'Recovery of drifting sensor responses by means of DWT analysis', *Sens Actuators B Chem*, vol. 120, no. 2, pp. 411–416, Jan. 2007, doi: 10.1016/J.SNB.2006.02.049.

[42] S. Lu *et al.*, 'An Improved Algorithm of Drift Compensation for Olfactory Sensors', *Applied Sciences 2022, Vol. 12, Page 9529*, vol. 12, no. 19, p. 9529, Sep. 2022, doi: 10.3390/APP12199529.

[43] H. Cui, X. Dong, and K. Shang, 'An Improved Method for Long-term Drift Compensation of Electronic Nose with Batch Control', *ICEIEC 2022 - Proceedings of 2022 IEEE 12th International Conference on Electronics Information and Emergency Communication*, pp. 145–148, 2022, doi: 10.1109/ICEIEC54567.2022.9835069.

[44] O. A. Ajibola, D. Smith, P. Španěl, and G. A. A. Ferns, 'INNOVATIVE TECHNIQUES Effects of dietary nutrients on volatile breath metabolites', *J Nutr Sci*, vol. 2, pp. 1–15, 2013, doi: 10.1017/jns.2013.26.

[45] A. Krilaviciute *et al.*, 'Associations of diet and lifestyle factors with common volatile organic compounds in exhaled breath of average-risk individuals', *J Breath Res*, vol. 13, no. 2, p. 026006, Mar. 2019, doi: 10.1088/1752-7163/AAF3DC.

[46] V. A. Binson, M. Subramoniam, and L. Mathew, 'Prediction of lung cancer with a sensor array based e-nose system using machine learning methods', *Microsystem Technologies*, vol. 30, no. 11, pp. 1421–1434, Apr. 2024, doi: 10.1007/S00542-024-05656-5/TABLES/4.

[47] 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 Rev Biomed Eng*, vol. 14, pp. 127–138, 2021, doi: 10.1109/RBME.2020.2993591.

[48] R. S. Parte, A. Patil, A. Patil, A. Kad, and S. Kharat, 'Non-Invasive Method for Diabetes Detection using CNN and SVM Classifier', *International Journal of Scientific Research and Engineering Development*, vol. 3, Accessed: Dec. 19, 2021. [Online]. Available: www.ijsred.com

[49] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew, 'Prediction of Pulmonary Diseases with Electronic Nose Using SVM and XGBoost', *IEEE Sens J*, vol. 21, no. 18, pp. 20886–20895, Sep. 2021, doi: 10.1109/JSEN.2021.3100390.

[50] Z. Chen, Y. Zheng, K. Chen, H. Li, and J. Jian, 'Concentration Estimator of Mixed VOC Gases Using Sensor Array with Neural Networks and Decision Tree Learning', *IEEE Sens J*, vol. 17, no. 6, pp. 1884–1892, Mar. 2017, doi: 10.1109/JSEN.2017.2653400.

[51] A. Ogunleye and Q. G. Wang, 'Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease', *IEEE International Conference on Control and Automation, ICCA*, vol. 2018-June, pp. 805–810, Aug. 2018, doi: 10.1109/ICCA.2018.8444167.

[52] L. Li *et al.*, 'Qualitative and Quantitative Transformer-CNN Algorithm Models for the Screening of Exhale Biomarkers of Early Lung Cancer Patients', *Anal Chem*, vol. 97, no. 12, pp. 6651–6660, Apr. 2025, doi: 10.1021/ACS.ANALCHEM.4C06604/ASSET/IMAGES/LARGE/AC4C06604_0008.JPEG.

[53] B. Lee *et al.*, 'Breath analysis system with convolutional neural network (CNN) for early detection of lung cancer', *Sens Actuators B Chem*, vol. 409, p. 135578, Jun. 2024, doi: 10.1016/J.SNB.2024.135578.

[54] W. Li *et al.*, 'A cross-sectional study of breath acetone based on diabetic metabolic disorders', *J Breath Res*, vol. 9, no. 1, p. 016005, Feb. 2015, doi: 10.1088/1752-7155/9/1/016005.

[55] A. T. Güntner, I. C. Weber, S. Schon, S. E. Pratsinis, and P. A. Gerber, 'Monitoring rapid metabolic changes in health and type-1 diabetes with breath acetone sensors', *Sens Actuators B Chem*, vol. 367, p. 132182, Sep. 2022, doi: 10.1016/J.SNB.2022.132182.

[56] L. Politi *et al.*, 'Discriminant Profiles of Volatile Compounds in the Alveolar Air of Patients with Squamous Cell Lung Cancer, Lung Adenocarcinoma or Colon Cancer', *Molecules 2021, Vol. 26, Page 550*, vol. 26, no. 3, p. 550, Jan. 2021, doi: 10.3390/MOLECULES26030550.

[57] I. Oakley-Girvan and S. W. Davis, 'Breath based volatile organic compounds in the detection of breast, lung, and colorectal cancers: A systematic review', *Cancer Biomarkers*, vol. 21, no. 1, pp. 29–39, Jan. 2018, doi: 10.3233/CBM-170177.

[58] J. D. M. Martin, F. Claudia, and A. C. Romain, 'How well does your e-nose detect cancer? Application of artificial breath analysis for performance assessment', *J Breath Res*, vol. 18, no. 2, p. 026002, Jan. 2024, doi: 10.1088/1752-7163/AD1D64.

[59] D. F. Altomare *et al.*, 'Chemical signature of colorectal cancer: case–control study for profiling the breath print', *BJS Open*, vol. 4, no. 6, pp. 1189–1199, Dec. 2020, doi: 10.1002/BJS5.50354.

[60] J. Chung *et al.*, 'Diagnosis by Volatile Organic Compounds in Exhaled Breath from Patients with Gastric and Colorectal Cancers', *International Journal of Molecular Sciences 2023, Vol. 24, Page 129*, vol. 24, no. 1, p. 129, Dec. 2022, doi: 10.3390/IJMS24010129.

[61] R. de Vries *et al.*, 'Prospective Detection of Early Lung Cancer in Patients With COPD in Regular Care by Electronic Nose Analysis of Exhaled Breath', *Chest*, vol. 164, no. 5, pp. 1315–1324, Nov. 2023, doi: 10.1016/J.CHEST.2023.04.050.

[62] N. Fens *et al.*, 'External validation of exhaled breath profiling using an electronic nose in the discrimination of asthma with fixed airways obstruction and chronic obstructive pulmonary disease', *Clinical & Experimental Allergy*, vol. 41, no. 10, pp. 1371–1378, Oct. 2011, doi: 10.1111/J.1365-2222.2011.03800.X.

[63] 'Cancer'. Accessed: Aug. 15, 2025. [Online]. Available: https://www.who.int/health-topics/cancer#tab=tab_1

[64] K. D. McCarthy, 'Detection of Lung, Breast, Colorectal, and Prostate Cancers From Exhaled Breath Using a Single Array of Nanosensors', *Breast Cancer Res Treat*, vol. 29, no. 8, p. 729, 2002.

[65] H. Y. Yang, Y. C. Wang, H. Y. Peng, and C. H. Huang, 'Breath biopsy of breast cancer using sensor array signals and machine learning analysis', *Sci. Rep.*, vol. 11, no. 1, p. 103, Dec. 2021, doi: 10.1038/s41598-020-80570-0.

[66] J. Zhang *et al.*, 'Identification potential biomarkers for diagnosis, and progress of breast cancer by using high-pressure photon ionization time-of-flight mass spectrometry', *Anal Chim Acta*, vol. 1320, Sep. 2024, doi: 10.1016/J.ACA.2024.342883.

[67] V. A. Binson, • M Subramoniam, and • Luke Mathew, 'Prediction of lung cancer with a sensor array based e-nose system using machine learning methods', doi: 10.1007/s00542-024-05656-5.

[68] C. Malagù *et al.*, 'Chemoresistive Gas Sensors for the Detection of Colorectal Cancer Biomarkers', *Sensors 2014, Vol. 14, Pages 18982-18992*, vol. 14, no. 10, pp. 18982–18992, Oct. 2014, doi: 10.3390/S141018982.

[69] G. Zonta *et al.*, 'Use of gas sensors and FOBT for the early detection of colorectal cancer', *Sens Actuators B Chem*, vol. 262, pp. 884–891, Jun. 2018, doi: 10.1016/J.SNB.2018.01.225.

[70] I. Poļaka *et al.*, 'The Detection of Colorectal Cancer through Machine Learning-Based Breath Sensor Analysis', *Diagnostics*, vol. 13, no. 21, p. 3355, Nov. 2023, doi: 10.3390/DIAGNOSTICS13213355/S1.

[71] J.-K. ; Kim *et al.*, 'Machine-Learning-Based Digital Twin System for Predicting the Progression of Prostate Cancer', *Applied Sciences 2022, Vol. 12, Page 8156*, vol. 12, no. 16, p. 8156, Aug. 2022, doi: 10.3390/APP12168156.

[72] D. Acevedo *et al.*, 'Prostate Cancer Detection in Colombian Patients through E-Senses Devices in Exhaled Breath and Urine Samples', *Chemosensors 2024, Vol. 12, Page 11*, vol. 12, no. 1, p. 11, Jan. 2024, doi: 10.3390/CHEMOSENSORS12010011.

[73] C. G. Waltman, T. A. T. Marcelissen, and J. G. H. van Roermund, 'Exhaled-breath Testing for Prostate Cancer Based on Volatile Organic Compound Profiling Using an Electronic Nose Device (Aeonose^{TM}): A Preliminary Report', *Eur Urol Focus*, vol. 6, no. 6, pp. 1220–1225, Nov. 2020, doi: 10.1016/J.EUF.2018.11.006.

[74] M. H. M. C. Scheepers *et al.*, 'Detection of differentiated thyroid carcinoma in exhaled breath with an electronic nose', *J Breath Res*, vol. 16, no. 3, p. 036008, Jun. 2022, doi: 10.1088/1752-7163/AC77A9.

[75] S. R. Markar *et al.*, 'Breath Volatile Organic Compound Profiling of Colorectal Cancer Using Selected Ion Flow-tube Mass Spectrometry', *Ann Surg*, vol. 269, no. 5, pp. 903–910, May 2019, doi: 10.1097/SLA.0000000000002539.

[76] M. Alorda-Clara *et al.*, 'Use of Omics Technologies for the Detection of Colorectal Cancer Biomarkers', *Cancers 2022, Vol. 14, Page 817*, vol. 14, no. 3, p. 817, Feb. 2022, doi: 10.3390/CANCERS14030817.

[77] 'Chronic obstructive pulmonary disease (COPD)'. Accessed: Aug. 15, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[78] 'Asthma'. Accessed: Aug. 15, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma

[79] A. Pardo Martínez, L. F. Romero, F. Sansone, and A. Tonacci, 'Non-Invasive Diagnostic Approaches for Kidney Disease: The Role of Electronic Nose Systems', *Sensors 2024, Vol. 24, Page 6475*, vol. 24, no. 19, p. 6475, Oct. 2024, doi: 10.3390/S24196475.

[80] D. Guo, D. Zhang, N. Li, L. Zhang, and J. Yang, 'A novel breath analysis system based on electronic olfaction', *IEEE Trans Biomed Eng*, vol. 57, no. 11, pp. 2753–2763, 2010, doi: 10.1109/TBME.2010.2055864.

[81] T. Jayasree, M. Bobby, and S. Muttan, ‘Sensor data classification for renal dysfunction patients using support vector machine’, *J Med Biol Eng*, vol. 35, no. 6, pp. 759–764, Dec. 2015, doi: 10.1007/S40846-015-0098-4/TABLES/1.

[82] Z. N. A. Said and A. M. El-Nasser, ‘Evaluation of urea breath test as a diagnostic tool for Helicobacter pylori infection in adult dyspeptic patients’, *World J Gastroenterol*, vol. 30, no. 17, p. 2302, 2024, doi: 10.3748/WJG.V30.I17.2302.

[83] P. Kashyap, P. Moayyedi, E. M. M. Quigley, M. Simren, and S. Vanner, ‘Critical appraisal of the SIBO hypothesis and breath testing: A clinical practice update endorsed by the European society of neurogastroenterology and motility (ESNM) and the American neurogastroenterology and motility society (ANMS)’, *Neurogastroenterology & Motility*, vol. 36, no. 6, p. e14817, Jun. 2024, doi: 10.1111/NMO.14817.

[84] F. Mion, F. Subtil, C. Machon, S. Roman, and A. Mialon, ‘The prevalence of small intestine bacterial overgrowth in irritable bowel syndrome is much higher with lactulose than glucose breath test: Results of a retrospective monocentric study’, *Clin Res Hepatol Gastroenterol*, vol. 48, no. 9, p. 102482, Nov. 2024, doi: 10.1016/J.CLINRE.2024.102482.

[85] R. A. Dweik *et al.*, ‘An Official ATS Clinical Practice Guideline: Interpretation of Exhaled Nitric Oxide Levels (FeNO) for Clinical Applications’, *https://doi.org/10.1164/rccm.9120-11ST*, vol. 184, no. 5, pp. 602–615, Dec. 2012, doi: 10.1164/RCCM.9120-11ST.

[86] N. Murugesan, D. Saxena, A. Dileep, M. Adrish, and N. A. Hanania, ‘Update on the Role of FeNO in Asthma Management’, *Diagnostics 2023, Vol. 13, Page 1428*, vol. 13, no. 8, p. 1428, Apr. 2023, doi: 10.3390/DIAGNOSTICS13081428.

[87] Z. Wang and C. Wang, ‘Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements’, *J. Breath Res.*, vol. 7, no. 3, p. 037109, Sep. 2013, doi: 10.1088/1752-7155/7/3/037109.

[88] N. Teshima, J. Li, K. Toda, and P. K. Dasgupta, ‘Determination of acetone in breath’, *Anal. Chim. Acta*, vol. 535, no. 1–2, pp. 189–99, Apr. 2005, doi: 10.1016/j.aca.2004.12.018.

[89] B. G. Stone, T. J. Besse, W. C. Duane, C. Dean Evans, and E. G. DeMaster, ‘Effect of regulating cholesterol biosynthesis on breath isoprene excretion in men’, *Lipids*, vol. 28, no. 8, pp. 705–708, Aug. 1993, doi: 10.1007/BF02535990/METRICS.

[90] B. A. Marzoog *et al.*, ‘Exhaled Breath Biomarkers Reflect the Inflammasome and Lipidome Changes in Ischemic Heart Disease: A Study Using Machine Learning Models and Network Analysis’, *J Lipid Atheroscler*, vol. 14, p., 2025, [Online]. Available: https://doi.org/10.12997/jla.2025.14.e30

[91] *EXHALED BREATH ANALYSIS : current status, challenges and future perspectives*. ELSEVIER ACADEMIC PRESS, 2025.

Author’s Achievements

Journal papers focused directly on the dissertation‘s subject:

- A. Paleczek and A. Rydosz, ‘Review of the algorithms used in exhaled breath analysis for the detection of diabetes’, *J Breath Res*, vol. 16, no. 2, p. 026003, Jan. 2022, doi: 10.1088/1752-7163/AC4916.
- A. Paleczek, ‘Recent achievements of exhaled breath analysis at the research stage—Artificial intelligence and machine learning algorithms’, *Exhaled Breath Analysis*, pp. 325–355, Jan. 2025, doi: 10.1016/B978-0-443-33796-3.00005-2.
- A. Paleczek, D. Grochala, and A. Rydosz, ‘Artificial breath classification using XGBoost algorithm for diabetes detection’, *Sensors*, vol. 21, no. 12, 2021, doi: 10.3390/s21124187.
- A. Paleczek and A. Rydosz, ‘The effect of high ethanol concentration on E-nose response for diabetes detection in exhaled breath: Laboratory studies’, *Sens Actuators B Chem*, vol. 408, p. 135550, Jun. 2024, doi: 10.1016/J.SNB.2024.135550.
- A. Paleczek *et al.*, ‘Noninvasive Total Cholesterol Level Measurement Using an E-Nose System and Machine Learning on Exhaled Breath Samples’, *ACS Sens*, Nov. 2024, doi: 10.1021/ACSSENSORS.4C02198.

Journal papers not related to the dissertation’s subject:

- A. Paleczek and A. Rydosz, “Medical sensor network and machine learning-enabled digital twins for diagnostic and therapeutic purposes,” in *Sensor networks for smart hospitals*, T. A. Nguyen, Ed., Amsterdam: Elsevier, 2025, pp. 77–94, ISBN: 9780443363702, e-ISBN: 9780443363719.
- D. Grochala, A. Paleczek, K. Staszek, M. Kocoń, K. Segełyn, Ł. Błajszczak, and A. Rydosz, “The impact of the epoxy thin-film layer on microwave-based gas sensor for detection,” *Sens. Actuators A Phys.*, vol. 388, art. no. 116498, pp. 1–8, 2025.
- J. Ramón-Azcón, A. Rydosz, F. De Chiara, J. M. Fernández-Costa, A. Ferret-Miñana, D. Grochala, J. Grochala, G. Lopez-Muñoz, S. Mughal, and A. Paleczek, *Human organs-on-a-chip: novel organ-on-a-chip techniques in medicine*, London: Academic Press, an imprint of Elsevier, 2024.
- A. Paleczek, D. Grochala, K. Staszek, S. Gruszczyński, E. Maciak, Z. Opilski, P. Kałużyński, M. Wójcikowski, T.-V. Cao, and A. Rydosz, “A sensor based on thin films for automotive applications in the microwave frequency range,” *Sens. Actuators B Chem.*, vol. 376 pt. B, art. no. 132964, pp. 1–13, 2023.
- K. Przystalski, A. Paleczek, K. Szustakowski, P. Wawryka, M. Jungiewicz, M. Zalewski, J. Kwiatkowski, A. Gądek, and K. Miśkowiec, “Automated correction angle calculation in high tibial osteotomy planning,” *Sci. Rep.*, vol. 13, no. 1, art. no. 12876, pp. 1–10, 2023.
- Ł. Fuśnik, B. Szafraniak, A. Paleczek, D. Grochala, and A. Rydosz, “A review of gas measurement set-ups,” *Sensors*, vol. 22, no. 7, art. no. 2557, pp. 1–30, 2022.
- D. Grochala, A. Paleczek, J. Bronicki, K. Marszałek, and A. Rydosz, “Wykorzystanie technologii GLAD do zastosowań w przenośnych analizatorach oddechu — The use of GLAD technology for applications in portable respiratory analyzers,” *Przegl. Elektrotech.*, vol. 98, no. 12, pp. 118–120, 2022.

- A. Paleczek, B. Szafraniak, Ł. Fuśnik, A. Brudnik, D. Grochala, S. Kluska, M. Jurzecka-Szymacha, E. Maciak, P. Kałużyński, and A. Rydosz, “The heterostructures of CuO and SnO_x for NO_2 detection,” *Sensors*, vol. 21, no. 13, art. no. 4387, pp. 1–17, 2021.

Conference communicates focused directly on the dissertation’s subject:

- S. Karcz, A. Paleczek, D. Grochala, M. Kocoń, Ł. Błajszczak, K. Staszek, and A. Rydosz, “The progress of the development of the electronic nose based on microwave gas sensors,” in 2025 URSI International Symposium on Electromagnetic Theory (EMTS 2025), Bologna, Italy, 2025, pp. 1–4, doi: 10.46620/URSIEMTS25/LAFF5521.
- D. Grochala, S. Karcz, A. Paleczek, M. Kocoń, M. Dudzik, K. Staszek, and A. Rydosz, “The development of gas-sensing setup for microwave-based e-nose detection system,” in *2024 4th URSI Atlantic Radio Science Meeting (AT-RASC)*, 19–24 May 2024, Meloneras, Spain.
- D. Grochala, A. Paleczek, M. Kocoń, and A. Rydosz, “The portable e-nose system for environmental and medical gas detection—proof of concept,” in *IOS’2024: Integrated Optics - Sensors, Sensing Structures and Methods*, 26 Feb–1 Mar 2024, Szczyrk, Beskidy Mountains, Poland.
- B. Szafraniak, Ł. Fuśnik, D. Grochala, A. Paleczek, J. Grochala, K. Wincza, and A. Rydosz, “SnO_2-based sensor for H_2S detection in exhaled human breath,” in *URSI GASS 2023: XXXVth General Assembly and Scientific Symposium of the International Union of Radio Science*, 19–26 Aug 2023, Sapporo, Japan.
- A. Paleczek, D. Grochala, and A. Rydosz, “Breath acetone classification using XGBoost algorithm for diabetes detection: [poster],” in *Breath Biopsy Conference 2021*, 12–13 Oct 2021, digital

Conference communicates not related to the dissertation’s subject:

- D. Grochala, S. Karcz, A. Paleczek, Ł. Błajszczak, M. Kocoń, K. Marszałek, S. Kern, R. W. Crisp, K. Staszek, and A. Rydosz, “Detection via microwave-based gas sensor with SnO_2 deposited by ALD,” in *GeMiC 2025: 16th German Microwave Conference*, 17–19 Mar 2025, Dresden, Germany.
- S. Karcz, M. Kocoń, A. Paleczek, D. Grochala, K. Staszek, and A. Rydosz, “Multi-sensor system for NO_2 detection in the microwave frequency range—theoretical and experimental results,” in *IMAPS Poland 2024: 46th International Microelectronics and Packaging*, 22–25 Sep 2024, Gdańsk, Poland.
- D. Grochala, A. Paleczek, M. Kocoń, M. Dudzik, Ł. Błajszczak, K. Staszek, M. Wójcikowski, T.-V. Cao, and A. Rydosz, “The impact of the epoxy thin-film layer for microwave-based gas sensors working at high relative humidity levels,” in *2024 4th URSI Atlantic Radio Science Meeting (AT-RASC)*, 19–24 May 2024, Meloneras, Spain.
- D. Grochala, A. Paleczek, J. Lemejda, M. Kajor, and M. Iwaniec, “Evaluation of geometric occlusal conditions based on the image analysis of dental plaster models,” *MATEC Web Conf.*, vol. 357, art. no. 05006, pp. 1–13, 2022.
- D. Grochala, A. Paleczek, K. Staszek, S. Gruszczyński, and A. Rydosz, “Nitrogen dioxide detection by the utilization of MoO_3-based gas sensing layer and eight-port reflectometer in the microwave frequency range,” in *IEEE Sensors 2022 Conference*, Dallas, TX, USA, 30 Oct–2 Nov 2022.

- D. Grochala, A. Paleczek, J. Bronicki, K. Marszałek, and A. Rydosz, "Opracowywanie parametrów technologii GLAD w celu kontrolowania osadzania warstw gazoczułych do zastosowań w przenośnych analizatorach [The development of the GLAD parameters for controlling the deposition process of gas-sensing materials used in the portable analyzers]," in *VIII Congress of the Polish Vacuum Society*, Cracow, Poland, 6–7 Jul 2022.
- A. Paleczek, D. Grochala, K. Staszek, K. Wincza, S. Gruszczyński, and A. Rydosz, "Microwave-based nitrogen dioxide gas sensor for automotive applications," in *ICECCME: Int. Conf. on Electrical, Computer, Communications and Mechatronics Engineering*, 7–8 Oct 2021, Mauritius.

Participation in research projects:

- National Centre for Research and Development (NCBR) - HYDROSTRATEG, "An integrated intelligent monitoring system limiting the migration of compounds of anthropogenic origin in rainwater retention systems", ID: HYDROSTRATEG1/0006/2022.
- National Science Centre (NCN) - SONATA BIS, "Investigation of the possibility of development of metaplasmonic sensors for the detection of diabetes and NAFLD biomarkers in the TDW/GLAD technology", ID: 2022/46/E/ST7/00008.
- National Science Centre (NCN) - OPUS, "Research on microwave gas sensors based on selected metal oxides – theoretical and experimental analysis", ID: 2021/41/B/ST7/0027612.
- National Centre for Research and Development (NCBR) - POL-NOR, "Highly Accurate and Autonomous Programmable Platform for Providing Air Pollution Data Services to Drivers and the Public", ID: NOR/POLNOR/HAPADS/0049/2019.
- National Science Centre (NCN) - SONATA, "Investigation of the influence of GLAD technique for 3S properties (sensitivity, selectivity, stability) of gas sensors with the enhanced response for diabetes biomarkers in exhaled human breath", ID: 2017/26/D/ST7/00355.
- AGH - IDUB Excellence Initiative – Research University ID: IDUB 4122, "Analysis of exhaled air as an innovative tool for medical diagnosis". The research was conducted in cooperation with CMUJ under the medical experiment entitled „Research on selected risk factors of metabolic syndrome and their impact on the condition of the prosthetic base". Bioethics Committee 1072.6120.40.2023 (granted 14/06/2023).

Active participation in scientific conferences:

- *8th International Conference on Bio-Sensing Technology* - 12-15 May 2024, Seville, Spain.
- *47th Annual International Conference of the IEEE Engineering in Medicine and Biology* - 14-17 July 2025, Copenhagen, Denmark.