Chen *et al. Respiratory Research* (2024) 25:32
https://doi.org/10.1186/s12931-024-02668-7

Respiratory Research

RESEARCH

Open Access

Advancing accuracy in breath testing for lung cancer: strategies for improving diagnostic precision in imbalanced data

Ke-Cheng Chen[1,2], Shuenn-Wen Kuo[1,2], Ruei-Hao Shie[3] and Hsiao-Yu Yang[4,5,6,7*]

Abstract

Background Breath testing using an electronic nose has been recognized as a promising new technique for the early detection of lung cancer. Imbalanced data are commonly observed in electronic nose studies, but methods to address them are rarely reported.

Objective The objectives of this study were to assess the accuracy of electronic nose screening for lung cancer with imbalanced learning and to select the best mechanical learning algorithm.

Methods We conducted a case–control study that included patients with lung cancer and healthy controls and analyzed metabolites in exhaled breath using a carbon nanotube sensor array. The study used five machine learning algorithms to build predictive models and a synthetic minority oversampling technique to address imbalanced data. The diagnostic accuracy of lung cancer was assessed using pathology reports as the gold standard.

Results We enrolled 190 subjects between 2020 and 2023. A total of 155 subjects were used in the final analysis, which included 111 lung cancer patients and 44 healthy controls. We randomly divided samples into one training set, one internal validation set, and one external validation set. In the external validation set, the summary sensitivity was 0.88 (95% CI 0.84–0.91), the summary specificity was 1.00 (95% CI 0.85–1.00), the AUC was 0.96 (95% CI 0.94–0.98), the pAUC was 0.92 (95% CI 0.89–0.96), and the DOR was 207.62 (95% CI 24.62–924.64).

Conclusion Electronic nose screening for lung cancer is highly accurate. The support vector machine algorithm is more suitable for analyzing chemical sensor data from electronic noses.

Keywords Volatile metabolite, Breathomics, Imbalanced learning, Electronic nose

*Correspondence:
Hsiao-Yu Yang
hyang@ntu.edu.tw
[1] Division of Thoracic Surgery, Department of Surgery, National Taiwan University Hospital, Taipei, Taiwan
[2] National Taiwan University College of Medicine, Taipei, Taiwan
[3] Green Energy and Environmental Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan
[4] Institute of Environmental and Occupational Health Sciences, National Taiwan University College of Public Health, No. 17 Xuzhou Road, Taipei 10055, Taiwan
[5] Department of Public Health, National Taiwan University College of Public Health, Taipei, Taiwan
[6] Population Health Research Center, National Taiwan University, Taipei, Taiwan
[7] Department of Environmental and Occupational Medicine, National Taiwan University Hospital, Taipei, Taiwan

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Lung cancer is the leading cause of cancer-related deaths globally, with 2.2 million new cases and 1.8 million deaths from lung cancer reported in 2020 [1]. Large numbers of lung cancer screening tools are being developed, but effective screening methods are still lacking. Chest radiography screening has now been found not to reduce lung cancer mortality [2]. Low-dose CT-based lung cancer screening reduces lung cancer mortality; however, the adoption of lung cancer screening programs has been slow [1]. A need exists to develop new lung cancer screening tools.

Breath tests are an innovative tool for lung cancer screening. Because lung cancer patients can exhale specific volatile organic compounds [3–5], studies have used gas sensing array technology (also known as an electric nose) to analyze the exhaled gases of lung cancer patients [6, 7]. The electronic nose uses an array of sensors to measure volatile metabolites in exhaled breath, offering the advantages of a short analysis time, low cost and ease of operation [8]. However, imbalanced data are commonly observed in electronic nose studies, but methods to address imbalanced data have seldom been reported [35]. Imbalanced data can negatively affect the accuracy of a diagnostic test by leading to biased models that perform poorly on minority classes [9]. When a diagnostic test is trained on an imbalanced dataset, the model may overfit to the majority class and underfit to the minority class, leading to better performance on the former. As a method to address this issue, we can balance the dataset by oversampling the minority population or undersampling the majority population to improve the performance of the model on the minority group [10]. Balancing the dataset can help improve performance; however, the accuracy must be corrected when sensitivity or specificity is exceptionally high to generalize the data in community-based screening [11].

The objectives of this study were to assess the accuracy of electronic nose screening for lung cancer with imbalanced learning and to select the best mechanical learning algorithm.

Methods

This study followed the STARD guidelines for reporting diagnostic accuracy [12].

Participants

We conducted a case–control study between October 2020 and July 2023, recruiting lung cancer patients from the outpatient department of National Taiwan University Hospital. The case group comprised those patients who underwent surgery and were diagnosed with lung cancer based on a pathological report. Lung cancer staging is based on the American Joint Committee on Cancer's TNM classification (7th edition) [13]. We recruited a healthy control group of health care workers who underwent health screening and chest X-rays in the same hospital during the same period and reported no lung cancer.

Exclusion criteria

At the screening stage, we excluded pregnant women and people younger than 20 years of age. At the final data analysis stage, we excluded patients with the following characteristics: 1. pathology reported as nonmalignant (e.g., hyperplasia, thymoma); 2. metastatic cancers; and 3. carcinoma in situ or minimally invasive adenocarcinoma.

Test methods

Collection of breath samples

We collected alveolar air samples using a standardized procedure [14]. As concentrations of volatile metabolites could be influenced by the flow rate, diet, and anatomical dead space [15, 16], we sampled alveolar air using specially designed equipment (Fig. 1). When expiratory carbon dioxide concentrations reached a high level representative of the expiratory alveolar phase, we collected alveolar air to prevent contamination of the respiratory or digestive dead space [17]. All subjects were not allowed to eat or smoke for 12 h prior to sampling. We used a fixed flow rate to obtain stable volatile metabolite concentrations and prevent the influence of the flow rate [16]. We followed a standardized cleaning protocol according to recommendations from the European Respiratory Society [18]. Each bag (SKC, Inc., USA) was flushed with nitrogen five times and then heated to 45 °C for approximately 12 h to prevent the influence of contaminated sampling bags.

Measurement

We collected alveolar air and analyzed samples within two hours. We used an electronic nose chemosensor Cyranose320 (Sensigent, California, USA) consisting of 32 nanocomposite conductive polymer sensors to analyze the breath samples. The breath sample in each bag was analyzed 10 times. The flow rate of the electronic nose was set to 120 ml/min, with a baseline purge for 10 s and a sample purge for 40 s, followed by a 10 s wash-out. Cyranose320 uses a 32 conductive polymer chemo-resistive sensor that is sensitive to temperature and humidity [19]. We operated the electronic nose in the same room, which was maintained at a stable temperature (mean 22.6 °C, standard deviation 2.4 °C) and ambient humidity (55%, standard deviation 7%).

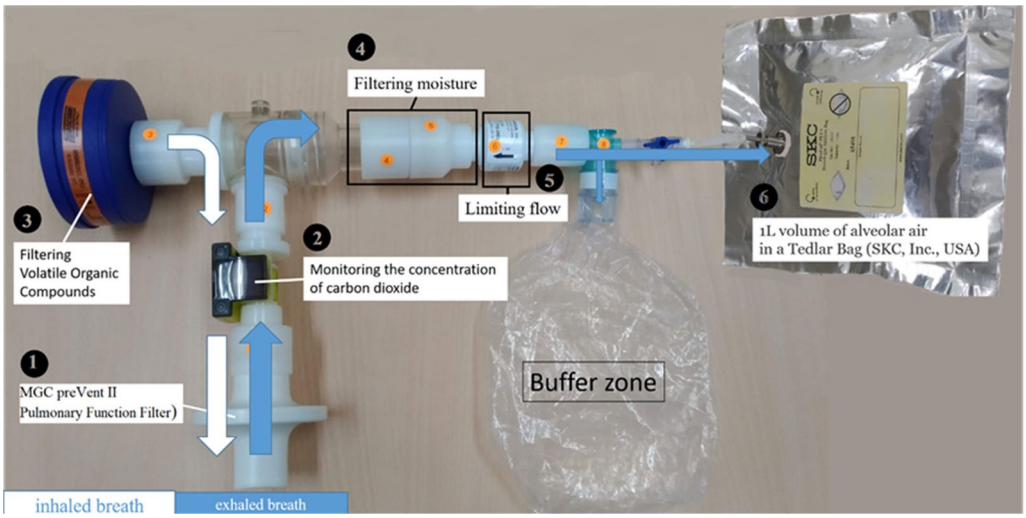


Fig. 1 Schematic of the system framework and sample collection. We sampled alveolar air from the collecting device with a volatile organic compound filter, capnometer, flow meter, nonrebreathing bag, and a three-way control valve. The capnometer (Masimo, CA, USA) was monitored [49]. Breath samples were stored in a 1-L Tedlar bag (SKC Inc., PA, USA)

Preprocessing sensor data

The sensor resistance resulting from pumping indoor air into the electronic nose was used as the baseline response to generate sensor response data. We normalized and autoscaled the raw sensor response data to remove background noise and exclude outliers [20, 21]. Then, the data were autoscaled to the unit variance that refers to mean centering and then divided by the standard deviation. After autoscaling, the value distribution of each sensor across the database was set to have a mean value of zero and one unit standard deviation [20]. We carefully inspected the raw sensor responses and removed sensor data showing S5 and S31 that showed data drift. Then, we obtained a mean value for each sensor response [22].

Statistics

The data were randomly divided into a training set used to derive the model (3/5 of the observations), an internal validation set (1/5 of the observations), and an external validation set (1/5 of the observations). The data used for validation were never used in model training. We prevented the effects of unequal case proportions in each group by using a synthetic minority oversampling technique (SMOTE) to balance the data by replicating minority observations [23]. We used five algorithms to compare which algorithm was suitable for electronic nose data and provided average values to allow readers to judge the potential range of accuracy when applied to screening. The five machine learning algorithms used to build predictive models included k-nearest neighbors, decision trees, neural networks, support vector machines (SVMs) (including linear kernels, polynomial kernels, and radial basis kernels), and random forests. [24]. We used the R packages "class" to model k-nearest neighbors, "C50" to model decision trees, "neuralnet" to model neural networks, "kernlab" to model SVMs, and "randomForest" to model random forest (RF). We used the modelLookup function of the caret package for automatic parameter tuning to improve model performance [25]. Using bootstrapping, we calculated the accuracy over 100 iterations to determine the parameters of the machine learning method with the highest prediction accuracy. The optimized model was then further tested using an internal validation set and an external validation set to assess its accuracy.

Using the pathology report as a reference standard, we determined the validity of the breath test by calculating accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating curve (AUC). AUC values of 0.7–0.8, 0.8–0.9, and 0.9–1.0 represent good, very good, and excellent diagnostic accuracy for lung cancer, respectively. [26]. We used 2000 bootstrap replicates to calculate confidence intervals for the AUC [27]. Based on the case–control study design, the proportion of patients in our study population will be higher than the proportion of lung cancer patients in the community. Partial AUC (pAUC) consists of analyzing only a region of special interest in the ROC curve and allows the selection of models with high specificity or sensitivity, rather than models with better than average performance but a potentially lower clinical value [28]. We calculated the partial AUC for 90–100 specificities and 90–100%

sensitivity to estimate the accuracy of breath test application in community screening [27, 29].

We calculated the diagnostic odds ratio (DOR) to compare the accuracy of different machine learning algorithms. A DOR value ranges from 0 to infinity, with higher values indicating better discriminatory test performance [30]. A test with a DOR of 10 is considered an excellent test [31]. We conducted a meta-analysis to derive summary estimates of DOR, sensitivity, and specificity for different machine learning algorithms. We plotted summary receiver operating characteristic (SROC) curves to generate a point estimate of overall accuracy and compared the overall accuracy between the internal validation and external validation sets [32].

We used the kappa statistic to assess reliability. Kappa indicates the degree to which the observed consistency exceeds the expected level of chance; kappa greater than 0.75 indicates excellent consistency beyond chance, kappa between 0.40 and 0.75 indicates moderate to good consistency, and kappa less than 0.40 indicates poor consistency [33].

Sensitivity analysis

Because smoking may affect exhaled volatile organic compounds [34, 35], we performed a sensitivity analysis that excluded current smokers.

Sample size estimation

We calculated the needed sample size using the following formula [36]:

$$\mathrm{SE} = \sqrt{\frac{C(100 - C)}{n}} \tag{1}$$

where SE is the standard error, C is the percentage of correctly classified patients, and n is the estimated sample size. An SE of 3 was used to limit the standard error to no more than 3%, and the acceptable accuracy (C) was 90 based on our goals. The number of needed samples for the training set is 100 [34, 35].

Results

We recruited 190 subjects between October 2020 and July 2023. We excluded 2 subjects who were unable to perform the breath test, 11 subjects without pathology reports, 5 subjects with pathology reports that were not diagnosed with lung cancer, 6 subjects with metastatic cancers, and 11 subjects with carcinoma in situ. Finally, 155 subjects were included in the final analysis, including 111 lung cancer patients and 44 healthy controls (Fig. 2). Table 1 shows the demographic characteristics of the study subjects, as well as the staging and histologic types of lung cancer in patients. Most of the lung cancers were stage I adenocarcinoma. The summary sensitivities for the internal and external validation sets were 0.86 (95% CI 0.82–0.89) and 0.88 (95% CI 0.84–0.91), respectively. Sensitivity ranged from 0.77 to 0.92, specificity ranged from 0.58 to 1.00, PPV ranged from 0.66 to 1.00, NPV ranged from 0.83 to 0.92, kappa ranged from 0.47 to 0.90, and AUC ranged from 0.91 to 0.94 (Table 2). The summary sensitivities for the internal and external validation sets were 0.86 (95% CI 0.82–0.89) and 0.88 (95% CI 0.84–0.91), respectively. The summary specificities for the internal and external validation sets were 0.99 (95% CI 0.81–1.00) and 1.00 (95% CI 0.85–1.00), respectively. The summary DORs for the internal and external validation sets were 110.97 (95% CI 31.77–387.62) and 206.95 (95% CI 51.84–826.19), respectively.

SROC curves showed higher overall sensitivity for the external validation set but also higher false-positive rates for the external validation set (Fig. 3). The area under the curve (AUC) for both the internal and external validation sets showed an increase in variability when assay specificity exceeded 90% for both the internal (Additional file 1: Figure S1) and external validation sets (Additional file 2: Figure S2). The partial AUCs between 90 and 100% specificity ranged from 0.74 to 0.83 in the internal validation set (Additional file 3: Figure S3) and from 0.83 to 0.86 in the external validation set (Additional file 4: Figure S4).

The summary DOR for the breath tests was 146.65 (95% CI 59.01–364.47) (Fig. 4), and the DORs for the internal validation and external validation sets were 110.97 (95% CI 31.77–387.62) and 206.95 (95% CI 51.84–826.19), respectively. The support vector machine using radial kernel had the highest DOR (Fig. 3) and the highest kappa value.

In the sensitivity analysis, although smoking is likely to influence the volatile organic compounds in breath, this study used a carbon black polymer sensor does not seem to be significantly influenced by smoking (Fig. 3).

Discussion

Principal findings

Although electronic noses have been used in diagnostic lung cancer studies, most of these studies have utilized a case–control study design with a high prevalence of lung cancer in the study population [37, 38], leading to a potential overestimation of diagnostic accuracy. The choice of study design (e.g., case–control studies conducted in hospitals or cross-sectional studies conducted in the community) can affect the assessment of accuracy, rendering the results not directly applicable, but few reports have been published in the field of machine learning. This study describes a method to correct the high accuracy obtained in a case–control study to estimate the real accuracy in a community setting. This study shows how the SMOTE approach can be applied to

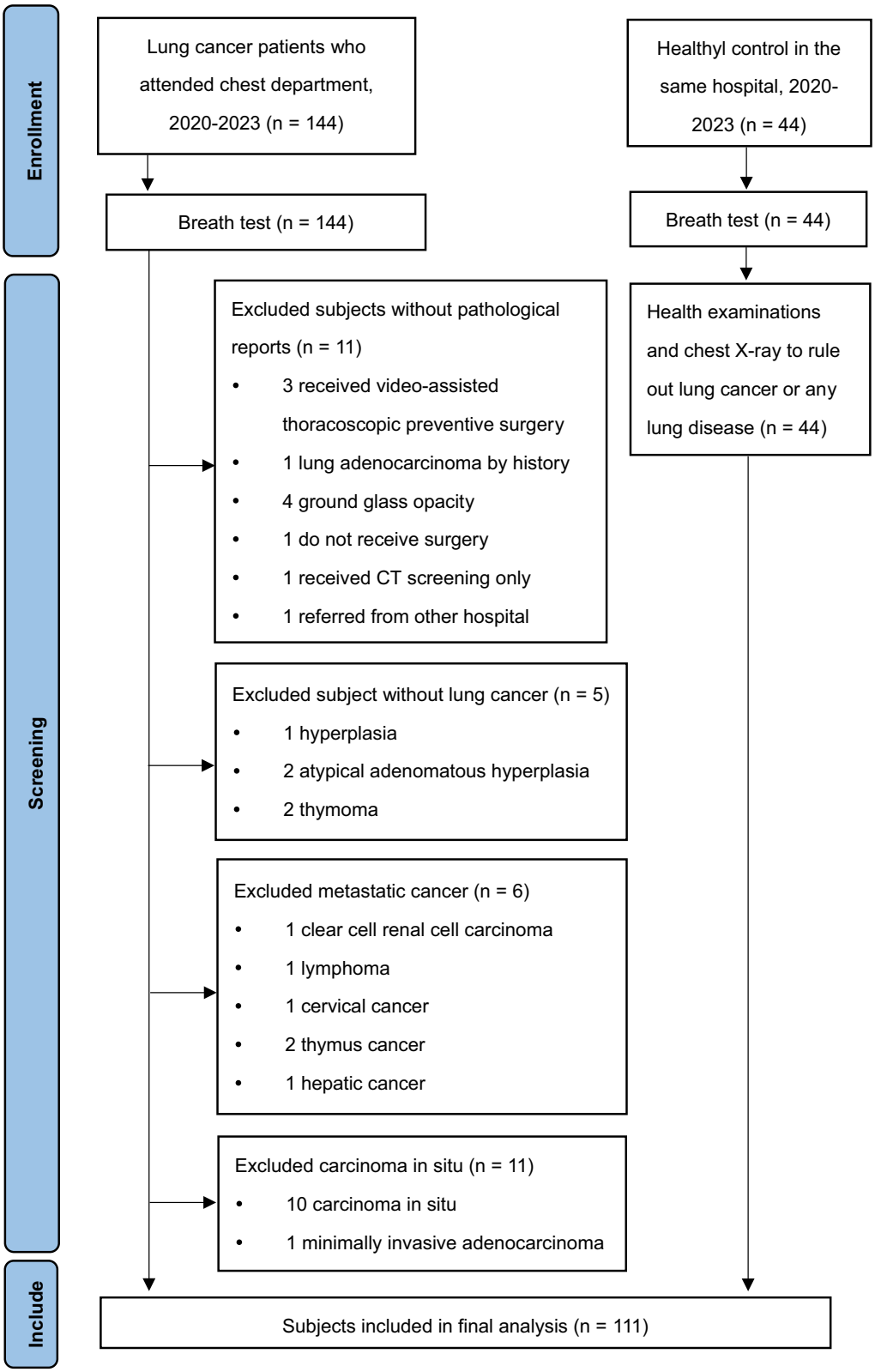


Fig. 2 Flow chart of the screening process

Table 1 Characteristics of the study subjects

Characteristics	Lung cancer Patients (*n* = 111)	Healthy controls (*n* = 44)
Sex (female) (%)	72 (64.86)	20 (45.45)
Age (years), mean (SD)	64.41 (11.85)	33.23 (8.35)
Never smoked (%)	83 (74.77)	42 (95.45)
Habitual cooking (%)	52 (47.27)	6 (13.64)
Burning incense (%)	37 (33.64)	0 (0)
Burning essential oils (%)	6 (5.45)	1 (2.27)
Family history of lung cancer (%)	30 (27.52)	7 (15.91)
Stage		
I (%)	87 (78.38)	–
II (%)	6 (5.41)	–
III (%)	10 (9.01)	–
IV (%)	8 (7.21)	–
Pathology		
Adenocarcinoma (%)	108 (97.30)	N/A
Squamous cell carcinoma (%)	1 (0.90)	N/A
Adenoid cystic carcinoma (%)	1 (0.90)	
Lymphoepithelial-like carcinoma (%)	1 (0.90)	

prevent a high proportion of cases or a scarcity of cases in a study population from generating spuriously high sensitivity and specificity. Considering the partial AUC, the electronic nose has good accuracy and reliability. To the best of our limited knowledge, this study is the first to provide epidemiologists' perspectives and suggest solutions for scientists from different disciplines. We provide evidence that imbalanced learning can enhance the accuracy of lung cancer screening. Compared to our previous study [39], in which air samples were collected from a tracheal tube placed in anesthetized patients, the AUCs were 0.91 (95% CI = 0.79–1.00) using linear discriminant analysis and 0.90 (95% CI = 0.80–0.99) using the supportive vector machine technique. This study further shows that the sensor array technique combined with machine learning can detect lung cancer in an outpatient setting.

The diagnostic accuracy of a test can be affected by the prevalence of the disease in the screened population [40]. Compared to other similar studies, sensitivity and specificity appear to be related to the proportion of lung cancer patients and nonlung cancer controls. When the proportion of controls is high, the specificity is higher than the sensitivity. When a higher proportion of controls is included, the specificity will be higher than the sensitivity. For example, Rocco et al. included 23 lung cancer patients and 77 nonlung cancer controls and reported a sensitivity and specificity of 86% and 95%, respectively [4]. Gasparri's study involved 70 lung cancer patients and 76 controls, with a sensitivity of 81% and specificity of 91% [5]. Shlomi et al. included 16 lung cancer patients and 30 controls, with a sensitivity of 75% and a specificity of 93% [41]. This difference is primarily due to the way sensitivity and specificity are calculated. Sensitivity is the ability of a test to correctly identify individuals with the disease (true positives). In a population with a higher disease prevalence, more true positive cases are present to detect. In a population with a 100% disease prevalence, even a completely false test that always predicts "having the disease" (regardless of the true disease status) would achieve 100% sensitivity. This outcome is because the test would correctly identify all individuals with the disease, but it would also incorrectly identify all individuals

Table 2 Prediction accuracy of the electronic nose data using machine learning algorithms

Model and parameters	Validation	Accuracy (95% CI)	Sensitivity	Specificity	PPV	NPV	Kappa	AUC (95% CI)
k-nearest neighbors (*k* = 5)	*Internal*	0.87 (0.79–0.93)	0.88	0.87	0.86	0.88	0.74	0.94 (0.89–0.99)
	External	0.94 (0.87–0.98)	0.88	1.00	1.00	0.90	0.88	0.95 (0.90–1.00)
Decision tree (trials = 20, winnow = TRU, model = tree)	*Internal*	0.73 (0.63–0.82)	0.90	0.58	0.66	0.86	0.47	0.91 (0.84–0.98)
	External	0.91 (0.84–0.96)	0.92	0.90	0.90	0.92	0.82	0.94 (0.90–1.00)
Neural network (size = 1, decay = 0)	*Internal*	0.89 (0.81–0.94)	0.77	1.00	1.00	0.83	0.78	0.92 (0.86–0.97)
	External	0.94 (0.87–0.98)	0.88	1.00	1.00	0.90	0.88	0.94 (0.89–0.99)
Support vector machines (linear kernel) (C = 1)	*Internal*	0.92 (0.85–0.96)	0.83	1.00	1.00	0.87	0.84	0.94 (0.89–0.99)
	External	0.92 (0.85–0.96)	0.83	1.00	1.00	0.87	0.84	0.94 (0.89–0.99)
Support vector machines (radial kernel) (C = 1)	*Internal*	0.95 (0.89–0.98)	0.90	1.00	1.00	0.91	0.90	0.94 (0.89–0.99)
	External	0.95 (0.89–0.98)	0.90	1.00	1.00	0.91	0.90	0.94 (0.89–0.99)
Support vector machines (polynomial kernel) (degree = 3, scale = 0.1, C = 1)	*Internal*	0.92 (0.85–0.96)	0.83	1.00	1.00	0.87	0.84	0.94 (0.89–0.99)
	External	0.94 (0.87–0.98)	0.88	1.00	1.00	0.90	0.88	0.94 (0.89–0.99)
Random forest (mtry = 2)	*Internal*	0.91 (0.84–0.96)	0.90	0.92	0.91	0.91	0.82	0.92 (0.86–0.99)
	External	0.79 (0.70–0.87)	0.85	0.73	0.75	0.84	0.58	0.94 (0.90–0.99)

PPV positive predictive value, *NPV* negative predictive value, *AUC* area under the receiver operating curve

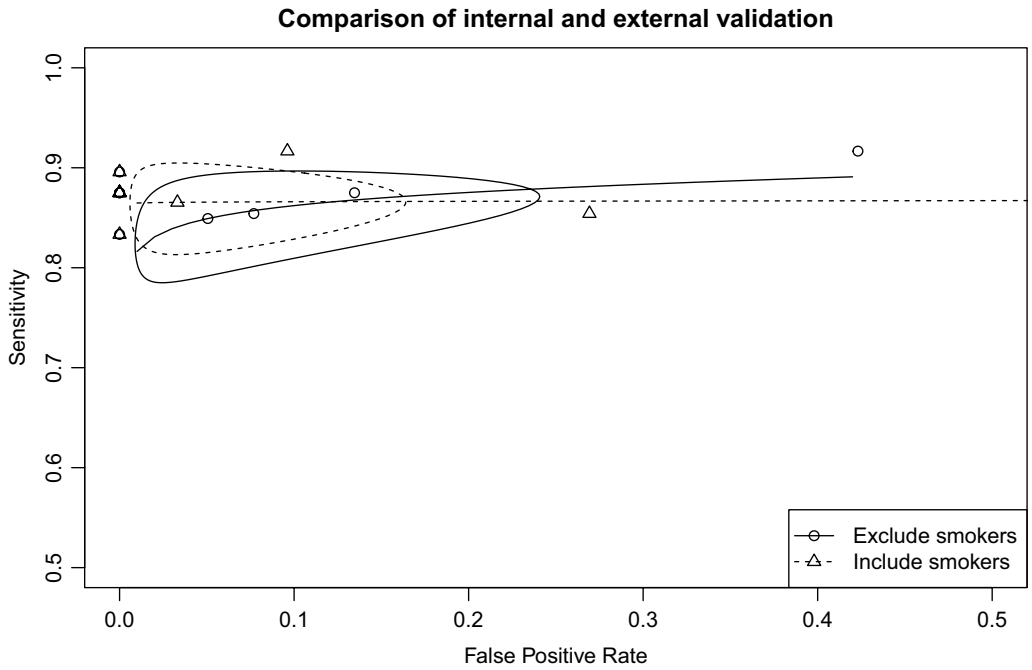
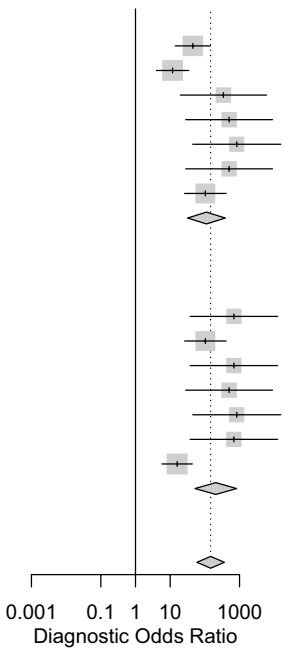


Fig. 3 Summary receiver operating characteristic (SROC) curve showing the diagnostic test accuracy for excluding smokers and including smokers. This figure shows summary receiver operating characteristic (SROC) curves for excluding smokers and including smokers. Scatter points are the accuracy obtained from different machine learning models. The ellipsoid confidence region depicts the 95% confidence region in the SROC space for the summary point estimate of diagnostic performance



Study	Experimental Events	Experimental Total	Control Events	Control Total	OR	95%−CI
validation = internal						
k Nearest Neighbors	42	49	6	51	45.000	[13.985; 144.801]
Decision Tree	43	65	5	35	11.727	[3.994; 34.435]
Neural Network	37	37	11	63	342.391	[19.564; 5992.238]
Support Vector Machines (Linear Kernel)	40	40	8	60	500.294	[28.043; 8925.496]
Support Vector Machines (Radial Kernel)	43	43	5	57	830.455	[44.664; 15441.095]
Support Vector Machines (polynomial Kernel)	40	40	8	60	500.294	[28.043; 8925.496]
Random Forest	43	47	5	53	103.200	[26.023; 409.262]
Random effects model		321		379	110.971	[31.770; 387.618]
Heterogeneity: I^2 = 67%, τ^2 = 1.7209, p < 0.01						
validation = external						
k Nearest Neighbors	42	42	6	58	686.538	[37.597; 12536.435]
Decision Tree	44	49	4	51	103.400	[26.077; 410.007]
Neural Network	42	42	6	58	686.538	[37.597; 12536.435]
Support Vector Machines (Linear Kernel)	40	40	8	60	500.294	[28.043; 8925.496]
Support Vector Machines (Radial Kernel)	43	43	5	57	830.455	[44.664; 15441.095]
Support Vector Machines (polynomial Kernel)	42	42	6	58	686.538	[37.597; 12536.435]
Random Forest	41	55	7	45	15.898	[5.796; 43.603]
Random effects model		313		387	206.947	[51.837; 826.186]
Heterogeneity: I^2 = 70%, τ^2 = 2.0475, p < 0.01						
Random effects model		**634**		**766**	**146.649**	**[59.005; 364.471]**
Heterogeneity: I^2 = 67%, τ^2 = 1.7518, p < 0.01						
Test for subgroup differences: χ^2_1 = 0.43, df = 1 (p = 0.51)						

Fig. 4 Forest plot of diagnostic odds ratios (DORs) for different machine learning algorithms and validation sets. The vertical dotted line represents the odds ratio (OR) and 95% confidence interval (95% CI). The diamond represents the point estimate and confidence intervals when combining and averaging all the machine learning algorithms together

without the disease as having it. As a result, the sensitivity of this test would be 100%, but its specificity, PPV, and NPVV would all be 0%. On the other hand, in a population with a 0% disease prevalence, even a completely false test that always predicts "no disease" (regardless of the true disease status) would achieve 100% specificity. However, its sensitivity, positive PPV, and NPV were all 0%. In other words, while it would be good at correctly identifying individuals without the disease (specificity), it would be entirely useless for correctly identifying individuals with the disease. This result highlights the importance of complete reporting of sensitivity, specificity, PPV, and NPV when evaluating the performance of diagnostic tests.

The choice of the study design can indeed have a significant influence on the detection accuracy of a diagnostic test. A case–control study design conducted in a hospital setting with a high prevalence of the disease in the study population can result in a higher sensitivity for the test. This outcome is because the likelihood of encountering individuals with the disease (cases) is higher in such a setting, which can make it easier for the test to detect the condition, thereby leading to a higher sensitivity. Therefore, while a hospital-based case–control study with a high disease prevalence can increase sensitivity, it can also result in a potential drawback known as "spectrum bias" [42]. Spectrum bias occurs when the study population does not accurately represent the broader population in which the test will eventually be used. A hospital-based case–control study with a high disease prevalence can enhance sensitivity but might not fully represent the test's performance in community-based scenarios with varying disease prevalences. The pAUC can be used as a method to assess the generalizability of the results to a broader population by providing a corrected AUC when sensitivity or specificity is exceptionally high due to imbalanced data, such as in situations where the disease prevalence is extremely high or low in the study population. The pAUC is a valuable tool in cases where traditional AUC measures may not adequately account for the impact of imbalanced data on evaluations of test performance.

SVM has the highest accuracy for analyzing data from electronic nose sensor arrays. Uddin et al. compared different machine learning algorithms used for disease prediction and found that SVM was most frequently used, but the RF algorithm was relatively accurate, followed by SVM. However, Uddin's review included algorithms used to analyze clinical and demographic data and did not specifically consider the electronic nose data, which may have different characteristics and requirements for analysis [43]. In our study, we evaluated the accuracy using DOR. The DOR is commonly used in meta-analyses of diagnostic studies that can combine results from different studies into summary estimates with increased precision [30]. The results show that SVM has the highest accuracy and is suitable for analyzing normalized and centralized electronic nose sensor array data. This finding is consistent with a comparison of machine learning algorithms for the classification of electronic nose data [44–46]. The reasons why SVM has the highest accuracy and is suitable for this type of analysis include (1) Non-Linearity Handling: SVM is capable of handling nonlinear data by mapping it into a higher-dimensional space using a kernel trick. This ability is valuable because electronic nose sensor array data can be complex and nonlinear [47]. (2) Strong Margin Maximization: SVM aims to find a decision boundary that maximizes the margin between different classes [48]. This approach often results in a more robust and accurate classification, especially when the data are well structured, as in normalized and centralized sensor array data. (3) Effective with High-Dimensional Data: SVM is particularly effective when analyzing high-dimensional data, which is often the case in electronic nose sensor array data. It helps find patterns and relationships in these complex datasets. However, importantly, the SVM performance depends on factors such as kernel selection, parameter tuning, and the specific characteristics of the dataset. While SVM can be a powerful tool, other algorithms must be considered and thorough experiments must be conducted to determine the best approach. However, SVM is sensitive to noisy data and outliers that can significantly affect the overall performance of the model. When analyzing electronic nose data, noisy data can come from (1) changes in environmental conditions such as temperature and humidity and (2) sampling variability, such as the way samples are prepared or presented to the electronic nose, which may contribute to noise. Inconsistent sample handling may lead to inconsistent sensor responses. (3) Background Odors: The presence of background odors or contaminants in the measurement environment can interfere with the detection of specific target odors, adding noise to the data. (4) Signal Processing Artifacts: The methods used for data preprocessing can introduce noise if not applied appropriately. Before applying SVM, we recommend that researchers develop standardized procedures for performing the electronic nose analysis in an air-conditioned room with a fixed temperature and humidity. Researchers can use a volatile organic compound filter at the inlet of the sampling device to eliminate background odors (Fig. 1, Part 3). A dehumidification chamber containing silica gel can be added to the inlet of the air samples (Fig. 1, Part 4). When preprocessing sensor data, normalization and autoscalation are

important to remove background noise and exclude outliers. Using standardized procedures, the findings can be generalized to other populations.

Conclusions

In this study, we developed a breath test for lung cancer using chemical sensors and machine learning techniques. Since the study was conducted in a hospital with a case–control study design, we described methods to estimate the accuracy of the breath test in community screening and how to apply the SMOTE approach to imbalanced data. Through a meta-analysis and the DOR, this study shows that the SVM algorithm is more suitable for classifying chemosensor data from electronic noses.

Abbreviations

PPV	Positive predictive value
NPV	Negative predictive value
AUC	Area under the receiver operating curve
pAUC	Partial AUC
DOR	Diagnostic odds ratio
SROC	Summary receiver operating characteristic curve
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
RF	Random forest

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-024-02668-7.

Additional file 1: Figure S1. Area under the curve (AUC) for the internal validation set. (Conti.) (a) k-nearest neighbours, (b) Decision tree, (c) Neural network, (d) Support vector machines, linear kernel, (e) Support vector machines, radial kernel, (f) Support vector machines, polynomial kernel, (g) Random forest.

Additional file 2: Figure S2. Area under the curve (AUC) for the external validation set. (a) k-nearest neighbors, (b) Decision tree, (c) Neural network, (d) Support vector machines, linear kernel, (e) Support vector machines, radial kernel, (f) Support vector machines, polynominal kernel, (g) Random forest.

Additional file 3. Figure S3.

Additional file 4. Figure S4.

Author contributions

Conceptualization, KCC; Methodology, KCC; Software, HYY; Validation, HYY; Formal Analysis, HYY; Investigation, SWK and KCC; Resources, HYY; Data Curation, SWK; Writing–Original Draft Preparation, KCC; Writing–Review & Editing, HYY; Visualization, HYY; Supervision, HYY; Project Administration, HYY.; Funding Acquisition, HYY.

Funding

This study was funded by the National Science and Technology Council of Taiwan (Grant numbers MOST 109-2314-B-002-166-MY3 and NSTC 112-2314-B-002-101) and the Population Health Research Center from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) of Taiwan (grant number NTU-112L9004) provided to the corresponding author. The funding bodies had no role in the design or execution of these studies beyond financial support.

Availability of data and materials

Data are available upon request due to privacy/ethical restrictions.

Declarations

Ethics approval and consent to patients

This study was approved by the Research Ethics Committee (No. 201912253RIND) of the National Taiwan University Hospital, and all subjects provided informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest.

Received: 18 November 2023 Accepted: 2 January 2024
Published online: 16 January 2024

References

1. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol. 2023;20:624–39.
2. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ, et al. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA. 2011;306:1865–73.
3. Hakim M, Broza YY, Barash O, Peled N, Phillips M, Amann A, Haick H. Volatile organic compounds of lung cancer and possible biochemical pathways. Chem Rev. 2012;112:5949–66.
4. Buszewski B, Ligor T, Jezierski T, Wenda-Piesik A, Walczak M, Rudnicka J. Identification of volatile lung cancer markers by gas chromatography–mass spectrometry: comparison with discrimination by canines. Anal Bioanal Chem. 2012;404:141–6.
5. Filipiak W, Filipiak A, Sponring A, Schmid T, Zelger B, Ager C, Klodzinska E, Denz H, Pizzini A, Lucciarini P, et al. Comparative analyses of volatile organic compounds (VOCs) from patients, tumors and transformed cell lines for the validation of lung cancer-derived breath markers. J Breath Res. 2014;8: 027111.
6. Kononov A, Korotetsky B, Jahatspanian I, Gubal A, Vasiliev A, Arsenjev A, Nefedov A, Barchuk A, Gorbunov I, Kozyrev K, et al. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer. J Breath Res. 2020;14:14.
7. Gasparri R, Santonico M, Valentini C, Sedda G, Borri A, Petrella F, Maisonneuve P, Pennazza G, D'Amico A, Di Natale C, et al. Volatile signature for the early diagnosis of lung cancer. J Breath Res. 2016;10: 016007.
8. Wojnowski W, Majchrzak T, Dymerski T, Gebicki J, Namiesnik J. Electronic noses: powerful tools in meat quality assessment. Meat Sci. 2017;131:119–31.
9. Khodabandelu S, Ghaemian N, Khafri S, Ezoji M, Khaleghi S. Development of a machine learning-based screening method for thyroid nodules classification by solving the imbalance challenge in thyroid nodules data. J Res Health Sci. 2022;22: e00555.
10. Prasetiyowati MI, Maulidevi NU, Surendro K. The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy. PeerJ Comput Sci. 2022;8: e1041.
11. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, Manuel DG. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak. 2020;20:4.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Standards for Reporting of Diagnostic A: towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem. 2003;49:1–6.

13. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17:1471–4.
14. Schubert JK, Spittler KH, Braun G, Geiger K, Guttmann J. CO(2)-controlled sampling of alveolar gas in mechanically ventilated patients. J Appl Physiol. 1985;2001(90):486–92.
15. Wang T, Pysanenko A, Dryahina K, Spanel P, Smith D. Analysis of breath, exhaled via the mouth and nose, and the air in the oral cavity. J Breath Res. 2008;2: 037013.
16. Bikov A, Hernadi M, Korosi BZ, Kunos L, Zsamboki G, Sutto Z, Tarnoki AD, Tarnoki DL, Losonczy G, Horvath I. Expiratory flow rate, breath hold and anatomic dead space influence electronic nose ability to detect lung cancer. BMC Pulm Med. 2014;14:202.
17. Lourenco C, Turner C. Breath analysis in disease diagnosis: methodological considerations and applications. Metabolites. 2014;4:465–98.
18. Horvath I, Barnes PJ, Loukides S, Sterk PJ, Hogman M, Olin AC, Amann A, Antus B, Baraldi E, Bikov A, et al. A European Respiratory Society technical standard: exhaled biomarkers in lung disease. Eur Respir J. 2017;49:1600965.
19. Peng G, Trock E, Haick H. Detecting simulated patterns of lung cancer biomarkers by random network of single-walled carbon nanotubes coated with nonpolymeric organic materials. Nano Lett. 2008;8:3631–5.
20. Lu Y, Partridge C, Meyyappan M, Li J. A carbon nanotube sensor array for sensitive gas discrimination using principal component analysis. J Electroanal Chem. 2006;593:105–10.
21. Lewis NS. Comparisons between Mammalian and artificial olfaction based on arrays of carbon black– polymer composite vapor detectors. Acc Chem Res. 2004;37:663–72.
22. Bofan M, Mores N, Baron M, Dabrowska M, Valente S, Schmid M, Trove A, Conforto S, Zini G, Cattani P, et al. Within-day and between-day repeatability of measurements with an electronic nose in patients with COPD. J Breath Res. 2013;7: 017103.
23. Nnamoko N, Korkontzelos I. Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med. 2020;104: 101815.
24. Lantz B. Machine Learning with R. 2nd ed. Birmingham: Packt Publishing Ltd.; 2015.
25. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:1–26.
26. Simundic AM. Measures of diagnostic accuracy: basic definitions. EJIFCC. 2009;19:203–11.
27. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
28. Robin X, Turck N, Hainard A, Lisacek F, Sanchez JC, Muller M. Bioinformatics for protein biomarker panel classification: what is needed to bring biomarker panels into in vitro diagnostics? Expert Rev Proteomics. 2009;6:675–89.
29. Chen CY, Lin WC, Yang HY. Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research. Respir Res. 2020;21:45.
30. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129–35.
31. Blackman NJ. Systematic reviews of evaluations of diagnostic and screening tests. Odds ratio is not independent of prevalence. BMJ. 2001;323:1188.
32. Shim SR, Kim SJ, Lee J. Diagnostic test accuracy: application and practice using R software. Epidemiol Health. 2019;41: e2019007.
33. Tooth LR, Ottenbacher KJ. The kappa statistic in rehabilitation research: an examination. Arch Phys Med Rehabil. 2004;85:1371–6.
34. Papaefstathiou E, Stylianou M, Andreou C, Agapiou A. Breath analysis of smokers, non-smokers, and e-cigarette users. J Chromatogr B Analyt Technol Biomed Life Sci. 2020;1160: 122349.
35. Dragonieri S, Quaranta VN, Buonamico E, Battisti C, Ranieri T, Carratu P, Carpagnano GE. Short-term effect of cigarette smoke on exhaled volatile organic compounds profile analyzed by an electronic nose. Biosensors (Basel). 2022;12:520.
36. Dragonieri S, Quaranta VN, Carratu P, Ranieri T, Resta O. Influence of age and gender on the profile of exhaled volatile organic compounds analyzed by an electronic nose. J Bras Pneumol. 2016;42:143–5.
37. Kort S, Brusse-Keizer M, Schouwink H, Citgez E, de Jongh FH, van Putten JWG, van den Borne B, Kastelijn EA, Stolz D, Schuurbiers M, et al. Diagnosing non-small cell lung cancer by exhaled breath profiling using an electronic nose: a multicenter validation study. Chest. 2023;163:697–706.
38. Tirzite M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. J Breath Res. 2018;13: 016006.
39. Huang CH, Zeng C, Wang YC, Peng HY, Lin CS, Chang CJ, Yang HY. A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer. Sensors (Basel). 2018;18:2845.
40. Willis BH. Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. BMJ Open. 2012;2:e000746.
41. Shlomi D, Abud M, Liran O, Bar J, Gai-Mor N, Ilouze M, Onn A, Ben-Nun A, Haick H, Peled N. Detection of lung cancer and EGFR mutation by electronic nose system. J Thorac Oncol. 2017;12:1544–51.
42. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926–30.
43. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19:281.
44. Gromski PS, Correa E, Vaughan AA, Wedge DC, Turner ML, Goodacre R. A comparison of different chemometrics approaches for the robust classification of electronic nose data. Anal Bioanal Chem. 2014;406:7581–90.
45. Chen K, Liu L, Nie B, Lu B, Fu L, He Z, Li W, Pi X, Liu H. Recognizing lung cancer and stages using a self-developed electronic nose system. Comput Biol Med. 2021;131: 104294.
46. Liu L, Li W, He Z, Chen W, Liu H, Chen K, Pi X. Detection of lung cancer with electronic nose using a novel ensemble learning framework. J Breath Res 2021; 15.
47. Zhang CK, Zhou Y, Guo JW, Wang GQ, Wang X. Research on classification method of high-dimensional class-imbalanced datasets based on SVM. Int J Mach Learn Cybern. 2019;10:1765–78.
48. Wang YY, Chen SC, Xue H. Support Vector Machine incorporated with feature discrimination. Expert Syst Appl. 2011;38:12506–13.
49. Miekisch W, Schubert JK, Noeldge-Schomburg GF. Diagnostic potential of breath analysis–focus on volatile organic compounds. Clin Chim Acta. 2004;347:25–39.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC