






Article

Exploring Volatile Organic Compounds in Breath for High-Accuracy Prediction of Lung Cancer

Ping-Hsien Tsou ¹, Zong-Lin Lin ², Yu-Chiang Pan ³, Hui-Chen Yang ¹, Chien-Jen Chang ¹, Sheng-Kai Liang ¹, Yueh-Feng Wen ¹, Chia-Hao Chang ¹, Lih-Yu Chang ¹, Kai-Lun Yu ¹, Chia-Jung Liu ¹, Li-Ta Keng ¹, Meng-Rui Lee ¹, Jen-Chung Ko ^{1,*}, Guan-Hua Huang ^{2,3,*} and Yaw-Kuen Li ^{3,4,*}

- ¹ Department of Internal Medicine, National Taiwan University Hospital, Hsin-Chu Branch, Hsin-Chu 30059, Taiwan; anskyluker@gmail.com (P.-H.T.); isan2383@yahoo.com.tw (H.-C.Y.); cjchangone@yahoo.com.tw (C.-J.C.); paukai1111@yahoo.com.tw (S.-K.L.); freeman0509@gmail.com (Y.-F.W.); ntudcm@gmail.com (C.-H.C.); pagasus2751@yahoo.com.tw (L.-Y.C.); starbox103@gmail.com (K.-L.Y.); m10082100@gmail.com (C.-J.L.); ltkeng@gmail.com (L.-T.K.); sheepman1024@gmail.com (M.-R.L.)
- ² Institute of Statistics, National Yang Ming Chiao Tung University, Hsin-Chu 30010, Taiwan; superstar147896325@yahoo.com.tw
- ³ Center for Emergent Functional Matter Science, National Yang Ming Chiao Tung University, Hsin-Chu 30010, Taiwan; edisonpan0805@gmail.com
- ⁴ Department of Applied Chemistry, National Yang Ming Chiao Tung University, Hsin-Chu 30010, Taiwan
- * Correspondence: e1205470@ms22.hinet.net (J.-C.K.); ghuang@stat.nctu.edu.tw (G.-H.H.); nctuykl@g2.nctu.edu.tw (Y.-K.L.)



Citation: Tsou, P.-H.; Lin, Z.-L.; Pan, Y.-C.; Yang, H.-C.; Chang, C.-J.; Liang, S.-K.; Wen, Y.-F.; Chang, C.-H.; Chang, L.-Y.; Yu, K.-L.; et al. Exploring Volatile Organic Compounds in Breath for High-Accuracy Prediction of Lung Cancer. *Cancers* **2021**, *13*, 1431. <https://doi.org/10.3390/cancers13061431>

Academic Editor: Tsuyoshi Sugiura

Received: 22 February 2021

Accepted: 18 March 2021

Published: 21 March 2021

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



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Simple Summary: Human-exhaled volatile organic compounds (VOCs) can be altered by lung cancer and become identifiable biomarkers. We used selected ion flow tube mass spectrometry (SIFT-MS) to quantitatively analyze 116 kinds of VOCs, which were exhaled by 148 lung cancer patients and 168 healthy individuals and collected from the environment to obtain a group of comprehensive data. A predictive model yielding 0.92 accuracy, 0.96 sensitivity, 0.88 specificity, and 0.98 area under the curve (AUC) was established using an advanced machine learning eXtreme Gradient Boosting (XGBoost) algorithm that considered the influences of exhaled and environmental VOCs.

Abstract: (1) Background: Lung cancer is silent in its early stages and fatal in its advanced stages. The current examinations for lung cancer are usually based on imaging. Conventional chest X-rays lack accuracy, and chest computed tomography (CT) is associated with radiation exposure and cost, limiting screening effectiveness. Breathomics, a noninvasive strategy, has recently been studied extensively. Volatile organic compounds (VOCs) derived from human breath can reflect metabolic changes caused by diseases and possibly serve as biomarkers of lung cancer. (2) Methods: The selected ion flow tube mass spectrometry (SIFT-MS) technique was used to quantitatively analyze 116 VOCs in breath samples from 148 patients with histologically confirmed lung cancers and 168 healthy volunteers. We used eXtreme Gradient Boosting (XGBoost), a machine learning method, to build a model for predicting lung cancer occurrence based on quantitative VOC measurements. (3) Results: The proposed prediction model achieved better performance than other previous approaches, with an accuracy, sensitivity, specificity, and area under the curve (AUC) of 0.89, 0.82, 0.94, and 0.95, respectively. When we further adjusted the confounding effect of environmental VOCs on the relationship between participants' exhaled VOCs and lung cancer occurrence, our model was improved to reach 0.92 accuracy, 0.96 sensitivity, 0.88 specificity, and 0.98 AUC. (4) Conclusion: A quantitative VOCs databank integrated with the application of an XGBoost classifier provides a persuasive platform for lung cancer prediction.

Keywords: volatile organic compounds; SIFT-MS; XGBoost; lung cancer; breath analysis; machine learning

1. Introduction

Scientists have been interested in volatile organic compounds (VOCs) released by human bodies for over five decades. In 1971, Nobel Prize winner Linus Pauling revealed

that breath is a complex mixture comprised of approximately 250 VOCs [1]. In 1999, Phillips et al. [2] detected over 3400 different volatile compounds in exhaled human breath. Metabolic processes of the human body produce these compounds; they enter the lungs via blood and are exhaled. Therefore, variations in exhaled breath compounds' concentrations can be directly linked to a disease, such as cancer [3]. Of all cancer types, there are 1.6 million lung cancer deaths per year, which is more than the sum of the next three most common cancers, i.e., prostate, breast, and colon cancer [4]. Lung cancer is usually quiet in the early stages; patients frequently experience coughing, chest pain, weight loss, etc. These symptoms early tend to be ignored until advanced disease development to be taken seriously. Generally, the 5-year survival related to late diagnosis is around 10–15%. Using conventional diagnostic procedures such as computer tomography (CT), sputum cytology, and biopsy, 85% of lung cancer cases are detected at a phase at which therapy is ineffective for curing the disease [5]. Low-dose thoracic CT can detect tumors at early stages and reduce mortality from lung cancer [6]. With the incidences of lung cancer rising worldwide, early detection techniques are an essential and immediate need. Since the cost and concern regarding radiation exposure mainly impair the applications of currently available examinations, an effective, radiation-free, and less invasive approach for lung cancer screening needs to be established. Some of the current and developing methods used for lung cancer noninvasive detection methods are summarized in Table 1.

Studying VOCs is one of the most interesting strategies and has many advantages (Table 1). Researchers, commonly using gas chromatography-mass spectrometry (GC-MS), have demonstrated the presence of lung-cancer-specific profiles of VOCs [7]. Though GC-MS is an established technique for VOC analysis, compared with selected ion flow tube mass spectrometry (SIFT-MS), GC-MS requires precise calibrations of standard compounds if highly specific and reliable quantification is needed [8]. Though GC-MS analysis integrated with library search is a powerful strategy for compound identification, the quantitative assessment of VOC using GC-MS is far more complicated. On the contrary, SIFT-MS design benefits the quantitative analysis and real-time study of VOCs [9,10]. The research using SIFT-MS has been primarily centered on the esophagus and colorectal cancers [11,12], whereas the breath profile of lung cancer analysis by SIFT-MS has rarely been reported. The features of direct sampling and quantitative VOC estimation of SIFT-MS can provide a large quantity of VOC data useful for additional statistical modeling. For this kind of large data, multivariate and machine learning tools for chemometric applications seem to be the rational way to define, project, model, and interpret the results [7].

The SIFT-MS technique was used in this study for the detection of VOCs in human breath from lung cancer patients and healthy volunteers. A machine learning approach named eXtreme Gradient Boosting (XGBoost) [13] was used to classify participants according to status, cancerous or healthy, based on their VOC analyses. The sampling process and flow chart of SIFT-MS analysis are shown in Figure 1.

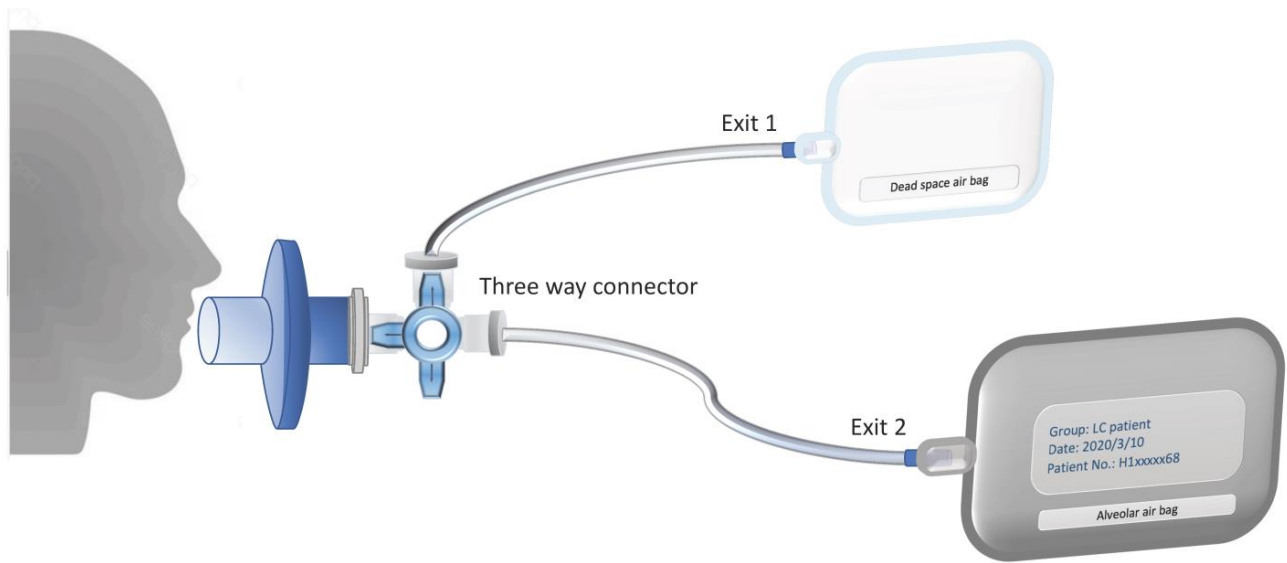
Table 1. Summary of the noninvasive detection methods for lung cancer.

Biomarkers/Specimen	Analytic Platform	Detection Target	Sensitivity (%)	Advantages	Deficiencies	Ref.
CTCs/Blood	IF; FISH	EpCAM, Size-based cells	30.0–69.5	Viable cell, high specificity, high throughput	Limited sensitivity; require enrichment; only detect advanced cancers	[14,15]
Traditional Proteins/Blood	ECLIA	CEA, CYFRA 21-1	22–69	Rapid and common	Limited sensitivity and specificity	[16]
Novel Proteins/EBC, Saliva, Urine, Blood	Microarray; LC-MS/MS	CKAP4, exosomal proteins (NFX1, PKG1, GPC1)	70.0–84.0	Higher sensitivity; high throughput; rapid	Quantity required (MS); validation required	[17–19]

Table 1. Cont.

Biomarkers/Specimen	Analytic Platform	Detection Target	Sensitivity (%)	Advantages	Deficiencies	Ref.
microRNA/Blood	Microarray; RT-PCR; NGS	miRNAs-126, -145, -210 and -205-5p, -17, -190b, -19a, -19b, -26b, -375	80.0–91.5	High throughput, stable	Specialized abilities and facilities are required	[20–24]
Methylated DNA/Blood	NGS; PCR	HOXD10, PAX9, PTPRN2, STAG3, SHOX2	70.0–87.8	High sensitivity and specificity	Require standardization	[25–27]
ctDNA/Blood	NGS; Multiplex-PCR	Genetic mutation, SNVs	48.0–59.0	Target for precision medicine; early detection (~70 days prior to CT image)	Limited sensitivity, require expensive equipment	[28–30]
VOCs/Exhaled Breath	E-Nose sensors; GC-MS; PTR-MS, IMS; LPPI-MS	propanol, isoprene, acetone, pentane, hexanal, toluene, benzene, ethylbenzene, and others	81.0–96.5	Rapid, simple, noninvasive; inexpensive	Require standardization	[7,31,32]

Abbreviations: CTCs (circulating tumor cells); IF (immunofluorescence); FISH (fluorescence in situ hybridization); EpCAM (epithelial cell adhesion molecule); ECLIA (electrochemiluminescence immunoassay); CEA (carcinoembryonic antigen); CYFRA 21-1 (cytokeratin fraction 21-1); EBC (exhaled breath condensate); NGS (next-generation sequencing); CT(computed tomography); RT-PCR (reverse transcription PCR); ctDNA (circulating tumor DNA); SNVs (single-nucleotide variants); GC-MS (gas chromatography mass spectrometry); PTR-MS (proton transfer reaction mass spectrometry); IMS (ion mobility spectrometry); LPPI-MS (low-pressure photoionization mass spectrometry); VOCs (volatile organic compounds).



(a)

Figure 1. Cont.

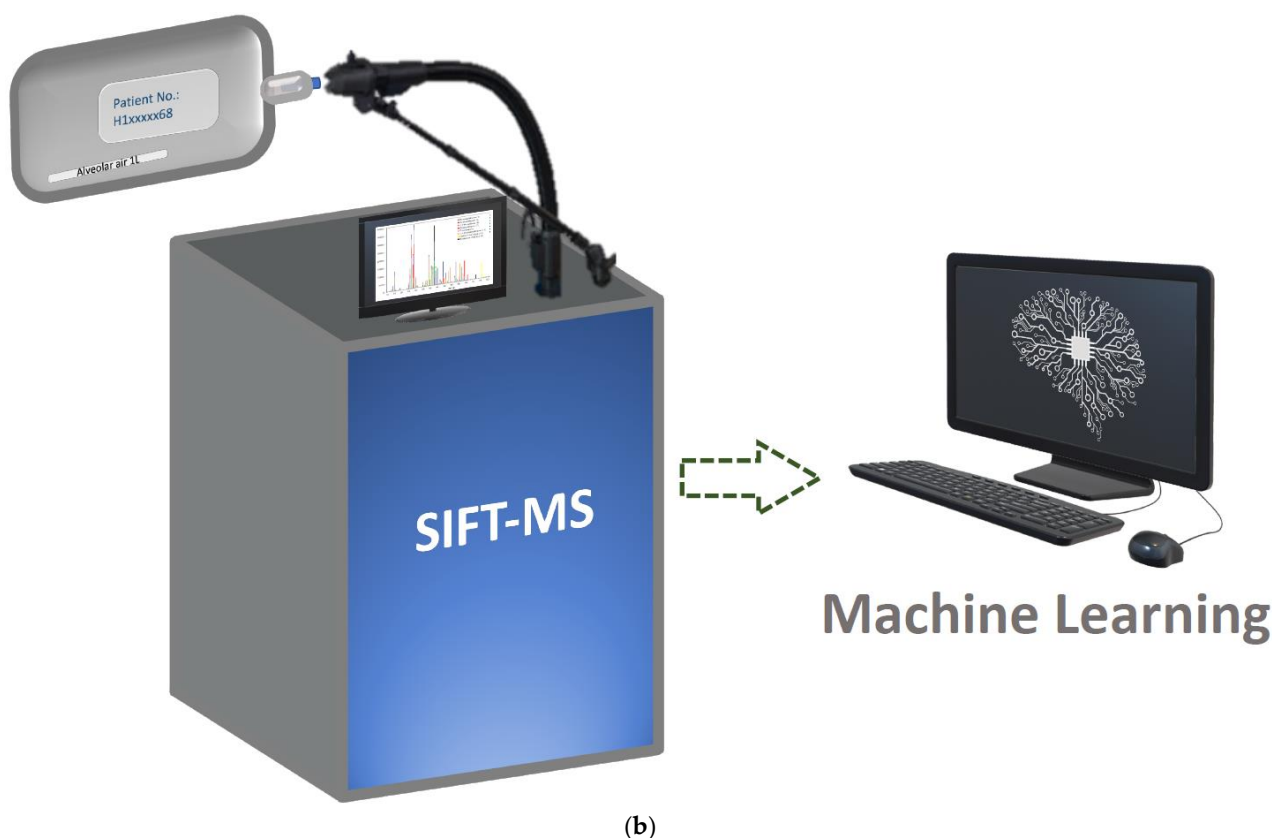


Figure 1. The research flow chart. (a) Collecting the alveolar air. The breath was exhaled through the mouthpiece with a direct-connect three-way valve. At the first stage, the exhaled air flows through Exit 1. When the volume of the front portion of the exhaled air reaches 0.2 L, the three-way valve switches to Exit 2 and starts to collect the rest of the exhaled air, i.e., alveolar air, in a 1.0 L aluminum bag. The process of sample collection may be repeated 2–3 times to collect enough samples for analysis. (b) Delivery and analysis of an exhaled sample. Selected ion flow tube mass spectrometry (SIFT-MS) extracts the exhaled breath from an aluminum bag and analyzes the composition of volatile organic compounds (VOCs). The VOC data are used for model construction, machine learning, and prediction of lung cancer.

2. Results

2.1. Characteristics of Patients with Lung Cancer and Healthy Volunteers

In this study, we enrolled 168 health volunteers (101 women) aged 20 to 74 years (controls) and 148 lung cancer patients (73 men) aged 37 to 90 years (Table 2). Lung cancer patients were older ($p < 0.001$) than the controls. Most histological types of lung cancer were adenocarcinoma (72.9%). The most elevated target driver mutation is exon 19 deletion (22.3%) and exon 21 point mutation (20.3%). Most patients had a nonresectable disease at clinical stage IIIB and C (18.2%), IVA (43.9%), or IVB (27%).

2.2. VOCs for SIFT-MS Analysis

This study investigated 116 specific VOCs previously reported as human breath biomarkers, shown in Table A1. Fifty VOCs showed significant differences between lung cancer patients and healthy controls in all three statistical hypothesis tests adopted (* in Table A1), which can be used as biomarkers for lung cancer. When analyzing the collected background air samples, we identified 57 environmental VOCs whose concentrations were not significantly different between the National Yang Ming Chiao Tung University (NCTU) and the National Taiwan University Hospital Hsin-Chu Branch (NTUH) in all statistical hypothesis tests ([†] in Table A1). The heat map for VOCs of different participants is shown in Figure 2. Fifty-seven percent ($n = 84$) of cancer patients were clustered tightly together (top red in the color bar on the left). Another 24% ($n = 36$) of cancer patients and 38%

($n = 43$) of healthy volunteers of NCTU were grouped closely (middle red and green in the color bar on the left). Healthy volunteers of NTUH were more spread out (blue in the color bar on the left). The hierarchical clustering identified several VOC groups (the dendrogram on the top), where two groups contained dominant features for distinguishing between cancer cases and healthy controls. Ethanol, formic acid, ethanediol, methanol, acetone, butane, and hexane (the far-left brown in the first color bar at the top) had higher values in cancer cases than in healthy controls. Another group of VOCs, including benzoic acid and beta-caryophyllene, (the far-right brown in the first color bar on the top), showed an extremely low concentration for most healthy controls. All these dominant VOCs, except hexane, were significantly different between lung cancer cases and healthy controls. Hexane and beta-caryophyllene were not significantly different between the NCTU and NTUH.

Table 2. Characteristics of patients with lung cancer and healthy volunteers of the study.

Characteristic	Lung Cancer Patients ($n = 148$)	Health Controls ($n = 168$)
Age (years), y^*		
Mean \pm SD	64.5 \pm 11	31.4 \pm 10.4
Range	37–90	20–74
Sex, n (%) [†]		
Female	75 (50.7)	101 (60.1)
Male	73 (49.3)	67 (39.9)
Smoking status, n (%) [*]		
Current smoker	9 (6)	0
Former smoker	47 (31.2)	1
Nonsmoker	92 (62.1)	167 (99)
Lung cancer type, n (%)		
Adenocarcinoma	108 (72.9)	
Squamous cell carcinoma	17 (11.5)	-
Small cell lung cancer	14 (9.5)	
Other lung cancer	8 (5.4)	
Targetable driver mutation, n (%)		
EGFR		
Exon 19 deletion	33 (22.3)	
Exon 21 point mutation	30 (20.3)	
T790M	6 (4.1)	-
ALK	7 (4.7)	
ROS1	3 (2.0)	
Wild type	75 (50.7)	
PD-L1 expression, n (%)		
>50%	18 (12.1)	
1–49%	57 (39.0)	-
<1%	29 (19.6)	
Clinical stage status, n (%)		
IA and B	4 (2.7)	
IIA and B	4 (2.7)	
IIIA	8 (5.4)	
III B and C	27 (18.2)	-
IVA	65 (43.9)	
IVB	40 (27.0)	

* Significantly different between lung cancer patients and healthy controls at p -value < 0.05. [†] Significantly different between lung cancer patients and healthy controls at p -value < 0.1. Abbreviations: EGFR (epidermal growth factor receptor); ALK (anaplastic lymphoma kinase); ROS1 (ROS1 oncogene); NTUH (National Taiwan University Hospital Hsin-Chu Branch); NCTU (National Yang Ming Chiao Tung University).

2.3. XGBoost Prediction Model

For prediction modeling, we first applied XGBoost to all VOC measurements for lung cancer disease state prediction, and its accuracy, sensitivity, specificity, and area under the curve (AUC) were 0.89, 0.82, 0.94, and 0.95, respectively. Using XGBoost with 50 significantly different VOCs between lung cancer cases and healthy controls, we obtained

accuracy, sensitivity, specificity, and AUC of 0.90, 0.84, 0.94, and 0.94, respectively, demonstrating the efficacy of our list of potential lung cancer biomarkers. Notably, our sensitivity and specificity were different, indicating the model's differential prediction ability for lung cancer cases and healthy controls. This might be due to the confounding effect from environmental VOCs, where all cases' breath samples were taken in the hospital, whereas those from controls were taken either in the hospital or in the academic campus.

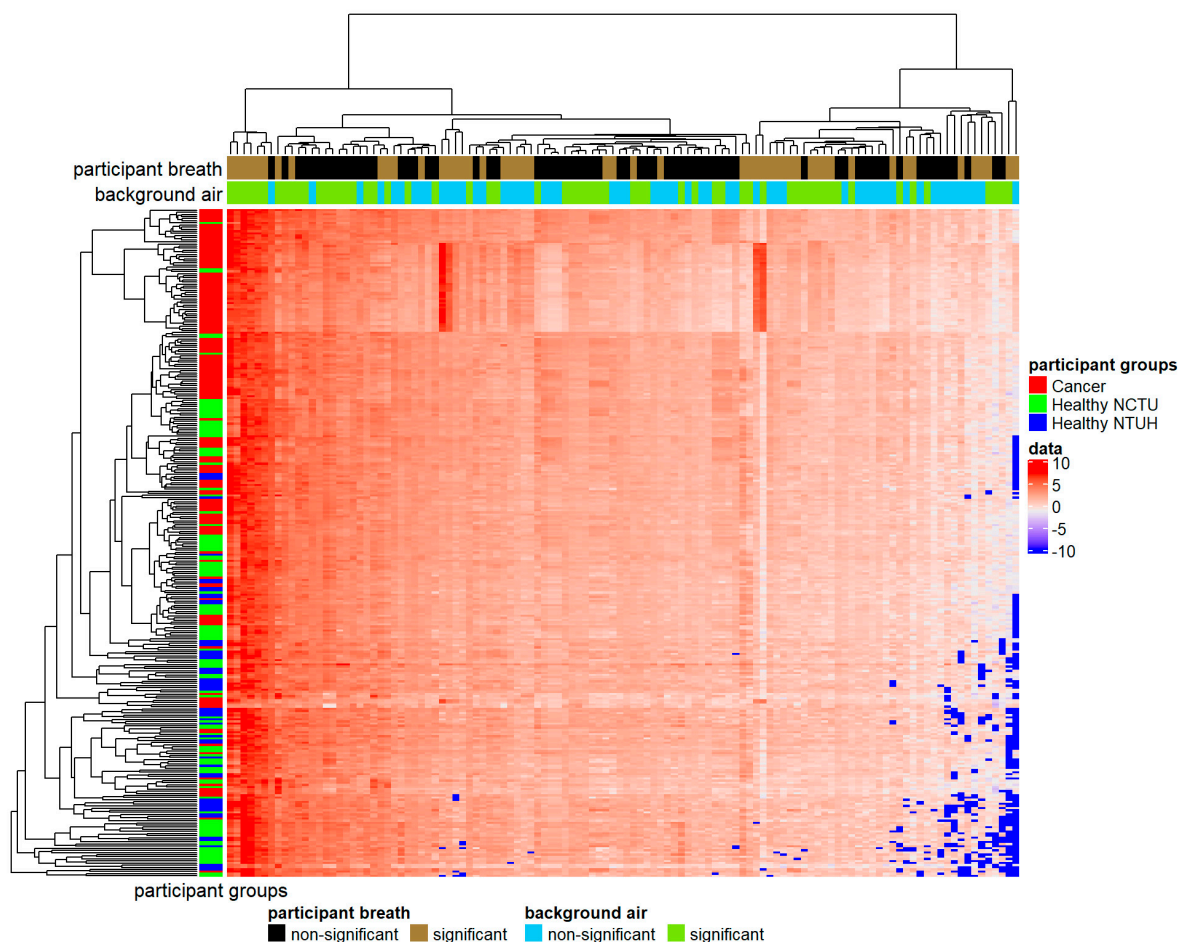


Figure 2. The heat map of 116 volatile organic compounds (VOC) measurements (log concentrations) for 316 participants. In the color matrix at the center, each row represents a participant, and each column represents a single VOC with the diverging color scheme of red (high concentration) and blue (low concentration). VOCs and participants are clustered using the agglomerative hierarchical clustering method. In the color bar on the left, red indicates cancer patients, green indicates healthy volunteers of the National Yang Ming Chiao Tung University (NCTU), and blue indicates healthy volunteers of the National Taiwan University Hospital Hsin-Chu Branch (NTUH). The two-color bars at the top represent the significance of VOCs. The first bar indicates whether VOCs showed significant differences between lung cancer patients and healthy controls (brown for significant and black for nonsignificant). The second bar indicates whether VOCs were significantly different between NCTU and NTUH (green for significantly different and blue for not significantly different).

2.4. Adjust Algorithm for Environmental VOCs

When XGBoost was built on environmentally nondifferential VOCs to eliminate the potential confounding effect from environmental VOCs, the accuracy, sensitivity, specificity, and AUC were 0.88, 0.84, 0.90, and 0.92, respectively, representing a slight improvement in sensitivity but a deterioration in specificity. We further applied SMOTE [33] to the VOC values of the collected background air samples to create synthetic environmental VOCs for each participant. The XGBoost prediction model incorporating participants' exhaled VOCs and these simulated environmental VOCs can account for nonendogenous VOCs present

in the environment. This model achieved better performance with 0.92 accuracy, 0.96 sensitivity, 0.88 specificity, and 0.98 AUC. These results are consistent with our speculation concerning the confounding effect of environmental VOCs. The approaches adopted here which consider environmental VOC effects can also improve prediction accuracy.

3. Discussion

Each whole breath can be divided into three parts according to the pressure of CO₂ in the exhalation. The first and second parts are dead space from the oropharynx and upper respiratory tract. The third part is the air from the alveoli deep inside the lungs that can exchange gases with the blood [34]. Previous breath analyses of lung cancer were performed by collecting the whole breath [35,36]. Some studies collected end-tidal breath by discarding the front of the breath [37,38] or filling the dead space air into other bags [39]. Studies have shown that the concentration of VOC is different in whole breath and end-tidal breath [40]. Our study used three-way connectors to manually fill Tedlar bags with air from dead areas of the mouth and upper respiratory tract and subsequently collect alveolar air from deep in the lungs into aluminum bags. Since VOCs in the alveoli are derived from the blood and gas exchange within the alveoli, this approach can better reflect VOCs' relationship with metabolic state changes caused by disease physiology.

The influence of environmental VOCs at the time of sampling can be considerable in the breathomics of lung cancer. More than 1000 exogenous VOCs are known to be detected in human respiration [41]. The relationship between environmental VOCs and the human body is complex and involves the processes of mixing, diffusion, and distribution in the blood and the metabolism in adipose tissue [42]. The concentration, exposure time, and solubility of the environmental VOCs in the human body and the individual physiology are the important factors that significantly affect the VOC contents of exhaled breath [43]. In past studies, there were no generally applicable rules for considering the influence of environmental VOCs. In addition to using the alveolar gradient concept [44], researchers solved this problem using inspiration filters [45] or having patients spend some time in the ventilation room before collection [38]. Although these methods are effective and widely accepted, we take one step further to eliminate the possible variances caused by environmental factors. Herein, we successfully introduced new algorithms to simulate environmental VOCs at the sampling time and incorporated them into the model to improve prediction accuracy. We showed that our approach could further abrogate the perturbation from environmental VOCs. For further applicability in various environments in the future, we suggest collecting and analyzing VOCs from participants and the environment simultaneously. The XGBoost model can further proceed with the process of learning and tuning, thus minimizing confounding effects. We significantly improved lung cancer prediction accuracy by selecting the phase of breath and calibrating environmental VOCs' impact.

This study describes an innovative machine-learning-based approach that uses SIFT-MS quantitative data to accurately distinguish respiratory samples of lung cancer patients from healthy controls. SIFT-MS quantitative analysis is achieved by applying precisely controlled ultra-soft chemical ionization combined with mass spectrometry detection [8]. The advantage of direct SIFT-MS is that it simplifies the needs for sample preparation, preconcentration, and chromatography. Our study shows that this machine-learning-based breath test is 0.96 sensitive, 0.88 specific, and highly accurate (0.98 area under the curve (AUC)) for identifying lung cancer, whereas the previous studies using multivariate classifiers based on VOCs' chemometrics analyzed by GC-MS demonstrated moderate to high accuracy (AUCs of 0.63–0.9, Table 3). The models reported herein show several advantages. The quantitative characteristics of SIFT-MS are attractive because many quantitative data can enhance model development and fine-tune the XGBoost model to improve prediction accuracy. The adopted XGBoost has been well-recognized and successfully applied in big data analytics. More importantly, our strategy of incorporating environmental VOC factors has unequivocally enhanced the power of XGBoost modeling for prediction.

Limitations of this study include the fact that it was a single-center and case–control study, and most of the patients were elderly and with advanced lung cancer. Our lung cancer patients were significantly older than healthy controls, leading to age mismatches and bias in case–control studies. With aging, a higher degree of oxidative stress occurs, and levels of VOCs in the breath increase, such as isoprene, alkanes, and methylated alkanes [46,47]. Our prediction model built on older and late-stage patients may fail in early detection of the disease. A multicenter study is currently planned to collect young, early, and operable lung cancer patients, aiming to provide an effective approach for early detection of lung cancer and a definitive answer to the test’s accuracy.

The clinical application of breathomics in lung cancer remains challenging up to the present. There are considerable differences in respiration sampling procedures, study designs, and data analysis methods implemented by studies for breathomics of lung cancer, which lead to inconsistent results. The effect of nutritional habits on breath VOCs can be complicated. By modifying metabolism, inflammation, or redox status, or communicating with gut flora, food influences breath VOCs. However, how long it takes for the dietary VOCs to be removed from the breath is not known. The dietary style also has a sustained influence that fasting could not remove [31]. There is no consensus on how to eliminate these dietary effects. We thus did not strictly screen participants’ nutritional status because we wanted to collect data of different dietary habits to establish a big data model that can be universally applied to the general population in the future. Another issue is that there is no validated list of VOC lung cancer biomarkers in the literature [31,48]. The lung cancer biomarkers found in these studies are mostly inconsonant [49]. The mechanism of most VOCs exhaled by the human body remains unclear. The following factors affect the concentration and composition of lung cancer VOCs in the human body: oxidative stress, cytochrome P450, liver enzymes, metabolic carbohydrates (glycolysis/gluconeogenesis pathways), and lipid metabolism [32]. These possible biochemical pathways vary from person to person, resulting in increased or decreased volatile organic compounds concentrations. Under these complex mechanisms, biochemical pathways of VOCs research cannot provide a definite answer. These VOCs that are considered possible and focused can be classified into the following families: hydrocarbons, primary and secondary alcohols, aldehydes and branched aldehydes, ketones, esters, nitriles, and aromatic compounds [32]. Our list of the targeted 116 VOCs was mainly derived from a literature search [32,49]. In this paper, we detail our approaches for VOC breathomics and provide suggested guidelines for further studies. Based on VOC analysis, this is by far the most comprehensive study (Table 3). Our research outcome provides a potentially useful model and platform for nonradiative and noninvasive lung cancer diagnosis.

Table 3. Summary of various algorithms applied to lung cancer diagnosis using multiple VOCs.

Algorithms	Analytical Platform	Patients with Cancer No.	Analyzed VOC No.	Sensitivity %	Specificity %	AUC	Reference/(Year)
Stepwise Discriminant Analysis	GC-MS	67	9	85.1	80.5	NR	[35]/(2003)
Logistic Regression	GC-MS	193	16	84.6	80.0	0.88	[50]/(2007)
Weighted Digital Sum Discriminator	GC-MS	193	30	84.5	81	0.9	[32]/(2008)
Support Vector Machine	GS-MS	107	5	95	89	NR *	[51]/(2016)
Artificial Neural Networks	GC-MS	108	88	86.36	86.36	0.86	[52]/(2019)
K-nearest Neighbor	GC-MS	325	NR	NR	NR	0.63 †	[53]/(2020) This Work
Extreme Gradient Boosting	SIFT-MS	148	116	82	94	0.95	Considering only participants’ VOCs Considering both participants’ VOCs and environmental VOCs
				96	88	0.98	

Abbreviations: AUC, area under the curve; GC-MS, gas chromatography-mass spectrometry; NR, not reported; SIFT-MS, selected ion flow tube mass spectrometry; * Accuracy: 89%, † Classify adenocarcinoma and squamous cell carcinoma patients.

4. Materials and Methods

4.1. Study Participants and Data Collection

Between May 2019 and June 2020, we obtained breath samples from 148 patients with histologically confirmed lung cancers and 168 healthy volunteer staff. All breath samples of lung cancer patients were collected in the National Taiwan University Hospital Hsin-Chu Branch (NTUH), and 112 and 56 healthy volunteers had their breath samples collected in the National Yang Ming Chiao Tung University (NCTU) and NTUH, respectively. Although we collected multiple breath samples from each healthy volunteer of the NCTU, only one randomly selected sample of each volunteer was included in the following analyses. The healthy volunteers had no history of significant pulmonary disease in any other organ and were free of disease. The patients' demographic profiles, history of cigarette smoking, staging, pathological finding, and cancer mutation testing were retrospectively collected from their medical records. The clinical cancer stage was based upon the American Joint Committee on Cancer (AJCC) TNM staging system 8th edition [54]. To evaluate the confounding effect from environmental VOCs, we also collected 18 and 29 background air samples from the NCTU and NTUH, respectively. These environmental VOCs were collected in the same place on the same day as participants' VOCs were obtained. All collected VOCs of the environment and participants were then analyzed. Each patient and volunteer provided written informed consent. This research was approved by the National Taiwan University Hospital Hsin-Chu Branch institutional review board (108-023-E).

4.2. Breath Sampling Methodology

All participants orally rinsed with water before sampling for breathing and stayed in the same place for more than 30 min before collecting the gas. To collect the alveolar breath and remove dead space air, each participant breathed normally through a disposable mouthpiece and into the device. We collected 0.2 L of the front portion of the exhaled breath flow through Exit 1 into a Tedlar bag (SKC Inc., Eighty Four, PA, USA), controlled by a 3-way valve. The remaining part of the exhaled breath, alveolar air, was collected in a 1.0 L aluminum bag, as shown in Figure 1a. A bag of room air was collected concurrently with some breath samples to account for nonendogenous VOCs present in the environment. Before collection, aluminum breath analysis bags were flushed with nitrogen gas at least ten times to remove background VOCs associated with the bags. The sealed samples were kept at room temperature (25 °C) and analyzed within 6 h. (Figure 1b). In some reports [55–57], VOC storage in bags potentially causes VOC content changes in breath samples due to storage and transportation conditions. In our case, we collected the samples from the location within a 15-min driving distance. To confirm our storage condition's feasibility, we performed a time-dependent analysis (twice a day for three days) on ten breath samples. By comparing the quantitative data from each sample, we conclude that most VOCs are considered stable.

4.3. Measurements of VOCs in Exhaled Air

The SIFT-MS theory is based on direct mass spectrometric analysis of VOCs in air or vapor samples by chemical ionization. Selected precursor ions (H_3O^+ , NO^+ , and O_2^+) are injected into the helium carrier gas and ionize the VOCs in the breath samples, generating characteristic productions detected by the downstream quadrupole mass spectrometer. Real-time quantification is achieved by measuring the count rate of both precursor ions and the characteristic product ions in the downstream detection system. The concentration of trace and volatile compounds is achieved at the parts-per-billion or parts-per-million by volume. It enables the simultaneous quantification within a gaseous mixture of several VOCs. A SIFT-MS instrument (VOICE200 ultra, Syft Technologies, Christchurch, New Zealand) was used to analyze the exhaled breath samples applying the selective ion mode (SIM). The 116 compounds shown in Table A1 were separated into seven categories: alkanes, ketones, aldehydes, alcohols, amines, thiols, and others for quantitative analysis. Among these VOCs, some of the product ions from different VOCs, e.g., MH^+ ions at No. 43,

47, 57, 59, 60, 61, 69, 71, 75, 85, 87, 89, 91, 97, 99, 101, and 103, overlap with others. The quantitative estimations of the selected VOCs were performed based on the pre-set protocol of SIFT-MS. This setting combines several product ions derived from three different reagent ions, H_3O^+ , O_2^+ , and NO^+ , with a tolerance feature setting at 20%. The tolerance feature is employed to deal with product ion interferences for every single compound. The final analyte concentration is calculated as an average of the lowest product ion concentration and anything within 20%. Any product ion that falls beyond the 20% tolerance range will not be accounted for in the calculation. Note that part of VOCs with the interference of product ions cannot be resolved by tolerance feature set can only be quantified as a relative scale. Both the accurate measurements and the VOCs' relative scales are collected and employed to construct statistical models.

4.4. Statistical Analysis

All analyses were performed using the R software (version 4.0.2; R Foundation for Statistical Computing). Participants' characteristics were analyzed and compared between lung cancer patients and healthy controls using the t-test or Wilcoxon rank-sum test for continuous variables and the Chi-square test or Fisher's exact test for categorical variables.

The heat map [58] was used to visualize VOCs' changes in different participants, helping establish the initial research hypothesis. We used the Wilcoxon rank-sum test and two-sample t-test with or without equal variance assumption to determine whether VOCs' differences between lung cancer patients and healthy controls were significant. To correct for multiple comparisons, the significance of the difference for each of the 116 VOC measurements was assessed at Bonferroni-corrected p -value = 0.0004 (0.05/116) [59]. We used the XGBoost [13], a machine learning method, to build a prediction model that used VOC measurements to predict lung cancer's disease state. We used 70% of the collected VOC data as the training set to build the prediction model and then used the other 30% to test its performance. The prediction model was established based on either all VOC measurements or those with a significant difference between the two study groups. To eliminate the confounding effect from environmental VOCs, we tried out two approaches: the prediction model built on VOCs whose concentrations were not significantly different among different environments and the model incorporating participants' exhaled VOCs and corresponding environmental VOCs simulated via SMOTE [33] from collected background air samples. The performance of various approaches was evaluated in terms of accuracy (the proportion of correctly classified participants), sensitivity, specificity, and AUC (the area under the ROC curve) on the test data.

5. Conclusions

To summarize, the machine learning model proposed in this study can accurately identify lung cancer using participants' exhaled breath. It is a non-invasive and radiation-free system, which can accelerate the diagnosis of lung cancer. We successfully demonstrated a new approach for disease diagnosis by integrating several techniques, including comprehensive and quantitative VOC analysis and deep learning algorithms for minimizing the interference of environmental factors, which resulted in an accurate prediction model. The development of standardized and automatic breath sampling protocols is an ongoing project that we expect will vastly simplify the process of sample collection and guarantee sample quality. We are confident that these efforts will ultimately unlock the potential and importance of human breathomics.

Author Contributions: Conceptualization, Y.-K.L., J.-C.K. and G.-H.H.; methodology, P.-H.T., Y.-K.L. and G.-H.H.; software, Z.-L.L.; validation, Y.-K.L. and G.-H.H.; formal analysis, G.-H.H.; investigation, P.-H.T.; data curation, P.-H.T., H.-C.Y., L.-T.K., C.-J.C., S.-K.L., Y.-F.W., C.-H.C., L.-Y.C., K.-L.Y., C.-J.L., M.-R.L. and Y.-C.P.; writing—original draft preparation, P.-H.T.; writing—review and editing, P.-H.T. and G.-H.H.; supervision, Y.-K.L., J.-C.K. and G.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of National Taiwan University Hospital Hsin-Chu Branch (108-023-E, approved 31 May 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to technical limitations, the potential issue of intellectual property, and ethics.

Acknowledgments: This work was financially supported by the MOST project 109-2634-F-009-028 and the Center for Emergent Functional Matter Science of National Yang Ming Chiao Tung University from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. The funding source had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

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

Appendix A

Table A1. The 116 VOCs selected for selected ion flow tube mass spectrometry analysis of breath samples in this study.

No.	Compound	No.	Compound	No.	Compound	No.	Compound
1 ^{*,†}	beta-caryophyllene (87-44-5)	30	2-pentanone (107-87-9)	59	diethyl ether (60-29-7)	88 [†]	1,4-diaminobutane (110-60-1)
2	pyrrole (109-97-7)	31 ^{*,†}	(E)-2-heptenal (18829-55-5)	60	isobutyl alcohol (78-83-1)	89	o-xylene (95-47-6)
3 [*]	benzoic acid (65-85-0)	32 [†]	3-buten-2-one (78-94-4)	61 [†]	2-methylpentane (107-83-5)	90 [†]	cyclopentane (287-92-3)
4 ^{*,†}	2,5-dimethylfuran (625-86-5)	33 [†]	butanone (78-93-3)	62	methylcyclopentane (96-37-7)	91	propane (74-98-6)
5 [*]	acetophenone (98-86-2)	34 ^{*,†}	1,5-diaminopentane (462-94-2)	63 [†]	heptanal (111-71-7)	92	heptane (142-82-5)
6	pyridine (110-86-1)	35 ^{*,†}	alpha-terpinene (99-86-5)	64	1-butanol (71-36-3)	93	propanal (123-38-6)
7 [*]	2-methylpyrazine (109-08-0)	36 [*]	1-butyne (107-00-6)	65	3-methyl-2-butenal (107-86-8)	94 [*]	2-propanol (67-63-0)
8 [†]	tridecane (629-50-5)	37	1-methyl-2-pyrrolidinone (872-50-4)	66 [†]	pentanoic acid (109-52-4)	95 ^{*,†}	cyclohexane (110-82-7)
9 [†]	2,5-dimethylpyrazine (123-32-0)	38 [†]	diisopropyl ether (108-20-3)	67 [*]	ethylbenzene (100-41-4)	96	ethane (74-84-0)
10 [†]	1,3-butadiene (106-99-0)	39	2-pentanone new (107-87-9)	68 ^{*,†}	1-heptene (592-76-7)	97	carbon disulfide (75-15-0)
11 ^{*,†}	dodecane (112-40-3)	40 [*]	1,2,4-trimethylbenzene (95-63-6)	69 ^{*,†}	dimethyl sulfide (75-18-3)	98 ^{*,†}	trimethylamine (75-50-3)
12	propyne (74-99-7)	41 [†]	nonane (111-84-2)	70 ^{*,†}	propanoic acid (79-09-4)	99	acetaldehyde (75-07-0)
13 [†]	(E)-2-nonenal (18829-56-6)	42 [*]	propylbenzene (103-65-1)	71	toluene (108-88-3)	100	dimethyl ether (115-10-6)
14 [*]	4-isopropyl toluene (99-87-6)	43	3-butyne-2-ol new (2028-63-9)	72	p-xylene (106-42-3)	101 [†]	acetic acid (64-19-7)
15 [†]	2-hexanone (591-78-6)	44 ^{*,†}	cyclohexanone (108-94-1)	73 [†]	3-methylbutanal (590-86-3)	102	propene (115-07-1)
16	undecane (1120-21-4)	45 ^{*,†}	ethylcyclohexane (1678-91-7)	74	butanal (123-72-8)	103	formaldehyde (50-00-0)
17 [*]	benzaldehyde (100-52-7)	46 [†]	2-methylbutanal (96-17-3)	75	xylenes + ethylbenzene (1330-20-7)	104	furan (110-00-9)
18 [*]	styrene (100-42-5)	47 [*]	nonanal (124-19-6)	76 [*]	isopropylamine (75-31-0)	105 [*]	1-propanol (71-23-8)
19 ^{*,†}	eucalyptol (470-82-6)	48 ^{*,†}	limonene (138-86-3; 7705-14-8)	77 ^{*,†}	methyl acetate (79-20-9)	106 [†]	isobutane (75-28-5)
20 [†]	furfural (98-01-1)	49 [†]	2-pentene (109-68-2)	78 ^{*,†}	1-hexene (592-41-6)	107	isoprene (78-79-5)
21 [*]	1-pentanol (71-41-0)	50	decane (124-18-5)	79 ^{*,†}	1-butene (106-98-9)	108 [*]	formic acid (64-18-6)
22 ^{*,†}	butyl acetate (123-86-4)	51	methyl n-propyl sulfide (3877-15-4)	80 [†]	pentanal (110-62-3)	109	pentane (109-66-0)
23 [*]	octanal (124-13-0)	53 [†]	2-methylpropanal (78-84-2)	81	1-methoxy-2-propanol (107-98-2)	110 [*]	acetonitrile (75-05-8)
24 [*]	3-methyl-1-butanol (123-51-3)	53 ^{*,†}	acetoin (513-86-0)	82	2,3-butanediol (513-85-9; 513-89-3)	111 [*]	ethanol (64-17-5)

Table A1. Cont.

No.	Compound	No.	Compound	No.	Compound	No.	Compound
25 [†]	(E)-2-hexenal (6728-26-3)	54 ^{*,†}	alpha-pinene (80-56-8; 2437-95-8)	83 [†]	hexanal (66-25-1)	112 [†]	hexane (110-54-3)
26 [†]	1,4-butyrolactone (96-48-0)	55 [*]	acrylonitrile (107-13-1)	84 ^{*,†}	acrolein (107-02-8)	113 [*]	methanol (67-56-1)
27 [†]	6-methyl-5-hepten-2-one (110-93-0)	56 ^{*,†}	ethyl acetate (141-78-6)	85 [†]	acetic anhydride (108-24-7)	114 [*]	acetone (67-64-1)
28	benzene (71-43-2)	57 ^{*,†}	2,3-butanedione (431-03-8)	86 [†]	3-methylpentane (96-14-0)	115 [*]	butane (106-97-8)
29 [†]	decanal (112-31-2)	58 ^{*,†}	2-methyl-2-propenal (78-85-3)	87 ^{*,†}	octane (111-65-9)	116 [*]	ethanedial (107-22-2)

* The VOC that showed a significant difference between lung cancer patients and healthy controls in all three statistical hypothesis tests adopted. † VOCs whose concentration was not significantly different between the National Yang Ming Chiao Tung University (NCTU) and the National Taiwan University Hospital Hsin-Chu Branch (NTUH) in all statistical hypothesis tests adopted.

References

- Pauling, L.; Robinson, A.B.; Teranishi, R.; Cary, P. Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 2374–2376. [CrossRef] [PubMed]
- Phillips, M.; Herrera, J.; Krishnan, S.; Zain, M.; Greenberg, J.; Cataneo, R.N. Variation in volatile organic compounds in the breath of normal humans. *J. Chromatogr. B Biomed. Sci. Appl.* **1999**, *729*, 75–88. [CrossRef]
- Hanna, G.B.; Boshier, P.R.; Markar, S.R.; Romano, A. Accuracy and Methodologic Challenges of Volatile Organic Compound-Based Exhaled Breath Tests for Cancer Diagnosis. *JAMA Oncol.* **2019**, *5*, e182815. [CrossRef] [PubMed]
- Weinberg, R.A.; Weinberg, R.A. *The Biology of Cancer*; Garland Science: New York, NY, USA, 2013.
- Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics. *Cancer J. Clin.* **2018**, *68*, 7–30. [CrossRef]
- The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **2011**, *365*, 395–409. [CrossRef]
- Marzorati, D.; Mainardi, L.; Sedda, G.; Gasparri, R.; Spaggiari, L.; Cerveri, P. A review of exhaled breath: A key role in lung cancer diagnosis. *J. Breath Res.* **2019**, *13*, e034001. [CrossRef]
- Syft Technologies Ltd. Technical Comparison SIFT-MS and GCMS full. Available online: <https://www.lqa.com/wp-content/uploads/2017/08/SIFTMS-vs-GCMS.pdf> (accessed on 7 December 2020).
- Španěl, P.; Smith, D. Progress in SIFT-MS: Breath analysis and other applications. *Mass Spectrom. Rev.* **2010**, *30*, 236–267. [CrossRef]
- Španěl, P.; Smith, D. Selected Ion Flow Tube Mass Spectrometry for On-Line Trace Gas Analysis in Biology and Medicine. *Eur. J. Mass Spectrom.* **2007**, *13*, 77–82. [CrossRef]
- Markar, S.R.; Chin, S.-T.; Romano, A.; Wiggins, T.; Antonowicz, S.; Paraskeva, P.; Ziprin, P.; Darzi, A.; Hanna, G.B. Breath Volatile Organic Compound Profiling of Colorectal Cancer Using Selected Ion Flow-tube Mass Spectrometry. *Ann. Surg.* **2019**, *269*, 903–910. [CrossRef]
- Markar, S.R.; Wiggins, T.; Antonowicz, S.; Chin, S.-T.; Romano, A.; Nikolic, K.; Evans, B.; Cunningham, D.; Mughal, M.; Lagergren, J.; et al. Assessment of a Noninvasive Exhaled Breath Test for the Diagnosis of Oesophago-gastric Cancer. *JAMA Oncol.* **2018**, *4*, 970–976. [CrossRef]
- 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. [CrossRef]
- Tanaka, F.; Yoneda, K.; Kondo, N.; Hashimoto, M.; Takuwa, T.; Matsumoto, S.; Okumura, Y.; Rahman, S.; Tsubota, N.; Tsujimura, T.; et al. Circulating Tumor Cell as a Diagnostic Marker in Primary Lung Cancer. *Clin. Cancer Res.* **2009**, *15*, 6980–6986. [CrossRef]
- Sonn, C.-H.; Cho, J.H.; Kim, J.-W.; Kang, M.S.; Lee, J.; Kim, J. Detection of circulating tumor cells in patients with non-small cell lung cancer using a size-based platform. *Oncol. Lett.* **2017**, *13*, 2717–2722. [CrossRef] [PubMed]
- Okamura, K.; Takayama, K.; Izumi, M.; Harada, T.; Furuyama, K.; Nakanishi, Y. Diagnostic value of CEA and CYFRA 21-1 tumor markers in primary lung cancer. *Lung Cancer* **2013**, *80*, 45–49. [CrossRef]
- Nolen, B.M.; Lomakin, A.; Marrangoni, A.; Velikokhatnaya, L.; Prosser, D.; Lokshin, A.E. Urinary Protein Biomarkers in the Early Detection of Lung Cancer. *Cancer Prev. Res.* **2015**, *8*, 111–119. [CrossRef]
- López-Sánchez, L.M.; Jurado-Gámez, B.; Feu-Collado, N.; Valverde, A.; Cañas, A.; Fernández-Rueda, J.L.; Aranda, E.; Rodríguez-Ariza, A. Exhaled breath condensate biomarkers for the early diagnosis of lung cancer using proteomics. *Am. J. Physiol. Cell. Mol. Physiol.* **2017**, *313*, 664–676. [CrossRef] [PubMed]
- Yanagita, K.; Nagashio, R.; Jiang, S.-X.; Kuchitsu, Y.; Hachimura, K.; Ichinoe, M.; Igawa, S.; Fukuda, E.; Goshima, N.; Satoh, Y.; et al. Cytoskeleton-Associated Protein 4 Is a Novel Serodiagnostic Marker for Lung Cancer. *Am. J. Pathol.* **2018**, *188*, 1328–1333. [CrossRef]
- Kim, H.; Yang, J.M.; Jin, Y.; Jheon, S.; Kim, K.; Lee, C.T.; Chung, J.-H.; Paik, J.H. MicroRNA expression profiles and clinico-pathological implications in lung adenocarcinoma according to EGFR, KRAS, and ALK status. *Oncotarget* **2017**, *8*, 8484–8498. [CrossRef] [PubMed]
- Li, N.; Wei, Y.; Wang, N.; Gao, H.; Liu, K. MicroRNA-26b suppresses the metastasis of non-small cell lung cancer by targeting MIEN1 via NF-κB/MMP-9/VEGF pathways. *Biochem. Biophys. Res. Commun.* **2016**, *472*, 465–470. [CrossRef]

22. Dacic, S.; Kelly, L.; Shuai, Y.; Nikiforova, M.N. miRNA expression profiling of lung adenocarcinomas: Correlation with mutational status. *Mod. Pathol.* **2010**, *23*, 1577–1582. [[CrossRef](#)]
23. Lu, S.; Kong, H.; Hou, Y.; Ge, D.; Huang, W.; Ou, J.; Yang, D.; Zhang, L.; Wu, G.; Song, Y.; et al. Two plasma microRNA panels for diagnosis and subtype discrimination of lung cancer. *Lung Cancer* **2018**, *123*, 44–51. [[CrossRef](#)]
24. Leng, Q.; Lin, Y.; Jiang, F.; Lee, C.-J.; Zhan, M.; Fang, H.; Wang, Y.; Jiang, F. A plasma miRNA signature for lung cancer early detection. *Oncotarget* **2017**, *8*, 111902–111911. [[CrossRef](#)] [[PubMed](#)]
25. Wielscher, M.; Vierlinger, K.; Kegler, U.; Ziesche, R.; Gsur, A.; Weinhäusel, A. Diagnostic Performance of Plasma DNA Methylation Profiles in Lung Cancer, Pulmonary Fibrosis and COPD. *EBioMedicine* **2015**, *2*, 929–936. [[CrossRef](#)]
26. Ilse, P.; Biesterfeld, S.; Pomjanski, N.; Wrobel, C.; Schramm, M. Analysis of SHOX2 methylation as an aid to cytology in lung cancer diagnosis. *Cancer Genom. Proteom.* **2014**, *11*, 251–258.
27. Duan, G.-C.; Zhao, Q.-T.; Guo, T.; Wang, H.-E.; Zhang, X.-P.; Zhang, H.; Wang, Z.-K.; Yuan, Z. Diagnostic value of SHOX2 DNA methylation in lung cancer: A meta-analysis. *OncoTargets Ther.* **2015**, *8*, 3433–3439. [[CrossRef](#)]
28. Fiala, C.; Diamandis, E.P. Circulating tumor DNA for personalized lung cancer monitoring. *BMC Med.* **2017**, *15*, 1–3. [[CrossRef](#)]
29. Abbosh, C.; Birkbak, N.J.; Wilson, G.A.; Jamal-Hanjani, M.; Constantin, T.; Salari, R.; Le Quesne, J.; Moore, D.A.; Veeriah, S.; Rosenthal, R.; et al. Erratum: Corrigendum: Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nat. Cell Biol.* **2018**, *554*, 264. [[CrossRef](#)] [[PubMed](#)]
30. Cohen, J.D.; Li, L.; Wang, Y.; Thoburn, C.; Afsari, B.; Danilova, L.; Douville, C.; Javed, A.A.; Wong, F.; Mattox, A.; et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **2018**, *359*, 926–930. [[CrossRef](#)]
31. Jia, Z.; Patra, A.; Kuttu, V.K.; Venkatesan, T. Critical Review of Volatile Organic Compound Analysis in Breath and In Vitro Cell Culture for Detection of Lung Cancer. *Metabolism* **2019**, *9*, 52. [[CrossRef](#)] [[PubMed](#)]
32. Hakim, M.; Broza, Y.Y.; Barash, O.; Peled, N.; Phillips, M.; Amann, A.; Haick, H. Volatile Organic Compounds of Lung Cancer and Possible Biochemical Pathways. *Chem. Rev.* **2012**, *112*, 5949–5966. [[CrossRef](#)] [[PubMed](#)]
33. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
34. D’Mello, J.; Butani, M. Capnography. *Indian J. Anaesth.* **2002**, *46*, 269–278.
35. Phillips, M.; Cataneo, R.N.; Cummin, A.R.; Gagliardi, A.J.; Gleeson, K.; Greenberg, J.; Maxfield, R.A.; Rom, W.N. Detection of Lung Cancer With Volatile Markers in the Breaths. *Chest* **2003**, *123*, 2115–2123. [[CrossRef](#)] [[PubMed](#)]
36. Wehinger, A.; Schmid, A.; Mechtcheriakov, S.; Ledochowski, M.; Grabmer, C.; Gastl, G.A.; Amann, A. Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas. *Int. J. Mass Spectrom.* **2007**, *265*, 49–59. [[CrossRef](#)]
37. Poli, D.; Carbognani, P.; Corradi, M.; Goldoni, M.; Acampa, O.; Balbi, B.; Bianchi, L.; Rusca, M.; Mutti, A. Exhaled volatile organic compounds in patients with non-small cell lung cancer: Cross sectional and nested short-term follow-up study. *Respir. Res.* **2005**, *6*, 71. [[CrossRef](#)] [[PubMed](#)]
38. Wang, Y.; Hu, Y.; Wang, D.; Yu, K.; Wang, L.; Zou, Y.; Zhao, C.; Zhang, X.; Wang, P.; Ying, K. The analysis of volatile organic compounds biomarkers for lung cancer in exhaled breath, tissues and cell lines. *Cancer Biomark.* **2012**, *11*, 129–137. [[CrossRef](#)] [[PubMed](#)]
39. Peng, G.; Hakim, M.; Broza, Y.Y.; Billan, S.; Abdah-Bortnyak, R.; Kuten, A.; Tisch, U.; Haick, H. Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors. *Br. J. Cancer* **2010**, *103*, 542–551. [[CrossRef](#)] [[PubMed](#)]
40. Doran, S.L.F.; Romano, A.; Hanna, G.B. Optimisation of sampling parameters for standardised exhaled breath sampling. *J. Breath Res.* **2017**, *12*, 016007. [[CrossRef](#)] [[PubMed](#)]
41. De Lacy Costello, B.; Amann, A.; Al-Kateb, H.; Flynn, C.; Filipiak, W.; Khalid, T.; Osborne, D.; Ratcliffe, N.M. A review of the volatiles from the healthy human body. *J. Breath Res.* **2014**, *8*, e014001. [[CrossRef](#)] [[PubMed](#)]
42. Phillips, M.; Greenberg, J.; Awad, J. Metabolic and environmental origins of volatile organic compounds in breath. *J. Clin. Pathol.* **1994**, *47*, 1052–1053. [[CrossRef](#)] [[PubMed](#)]
43. Beauchamp, J. Inhaled today, not gone tomorrow: Pharmacokinetics and environmental exposure of volatiles in exhaled breath. *J. Breath Res.* **2011**, *5*, e037103. [[CrossRef](#)]
44. Rudnicka, J.; Kowalkowski, T.; Ligor, T.; Buszewski, B. Determination of volatile organic compounds as biomarkers of lung cancer by SPME–GC–TOF/MS and chemometrics. *J. Chromatogr. B* **2011**, *879*, 3360–3366. [[CrossRef](#)]
45. Machado, R.F.; Laskowski, D.; Deffenderfer, O.; Burch, T.; Zheng, S.; Mazzone, P.J.; Mekhail, T.; Jennings, C.; Stoller, J.K.; Pyle, J.; et al. Detection of Lung Cancer by Sensor Array Analyses of Exhaled Breath. *Am. J. Respir. Crit. Care Med.* **2005**, *171*, 1286–1291. [[CrossRef](#)]
46. Lechner, M.; Moser, B.; Niederseer, D.; Karlseder, A.; Holzknicht, B.; Fuchs, M.; Colvin, S.; Tilg, H.; Rieder, J. Gender and age specific differences in exhaled isoprene levels. *Respir. Physiol. Neurobiol.* **2006**, *154*, 478–483. [[CrossRef](#)] [[PubMed](#)]
47. Phillips, M.; Greenberg, J.; Cataneo, R.N. Effect of age on the profile of alkanes in normal human breath. *Free Radic. Res.* **2000**, *33*, 57–63. [[CrossRef](#)] [[PubMed](#)]
48. Krilaviciute, A.; Heiss, J.A.; Leja, M.; Kupcinskas, J.; Haick, H.; Brenner, H. Detection of cancer through exhaled breath: A systematic review. *Oncotarget* **2015**, *6*, 38643–38657. [[CrossRef](#)] [[PubMed](#)]
49. Saalberg, Y.; Wolff, M. VOC breath biomarkers in lung cancer. *Clin. Chim. Acta* **2016**, *459*, 5–9. [[CrossRef](#)] [[PubMed](#)]
50. Phillips, M.; Altorki, N.; Austin, J.H.; Cameron, R.B.; Cataneo, R.N.; Greenberg, J.; Kloss, R.; Maxfield, R.A.; Munawar, M.I.; Pass, H.I.; et al. Prediction of lung cancer using volatile biomarkers in breath. *Cancer Biomark.* **2007**, *3*, 95–109. [[CrossRef](#)] [[PubMed](#)]

51. 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](#)]
52. Rudnicka, J.; Kowalkowski, T.; Buszewski, B. Searching for selected VOCs in human breath samples as potential markers of lung cancer. *Lung Cancer* **2019**, *135*, 123–129. [[CrossRef](#)]
53. Wang, C.; Long, Y.; Li, W.; Dai, W.; Xie, S.; Liu, Y.; Zhang, Y.; Liu, M.; Tian, Y.; Li, Q.; et al. Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics. *Sci. Rep.* **2020**, *10*, 1–12. [[CrossRef](#)]
54. Amin, M.B.; Greene, F.L.; Edge, S.B.; Compton, C.C.; Gershengwald, J.E.; Brookland, R.K.; Meyer, L.; Gress, D.M.; Byrd, D.R.; Winchester, D.P. *AJCC Cancer Staging Manual*, 8th ed.; Springer: New York, NY, USA, 2017.
55. Mochalski, P.; King, J.; Unterkofler, K.; Amann, A. Stability of selected volatile breath constituents in Tedlar, Kynar and Flexfilm sampling bags. *Analyst* **2013**, *138*, 1405–1418. [[CrossRef](#)] [[PubMed](#)]
56. Van Harreveld, A. (Ton) Odor Concentration Decay and Stability in Gas Sampling Bags. *J. Air Waste Manag. Assoc.* **2003**, *53*, 51–60. [[CrossRef](#)] [[PubMed](#)]
57. Di Gilio, A.; Palmisani, J.; Ventrella, G.; Facchini, L.; Catino, A.; Varesano, N.; Pizzutilo, P.; Galetta, D.; Borelli, M.; Barbieri, P.; et al. Breath Analysis: Comparison among Methodological Approaches for Breath Sampling. *Molecules* **2020**, *25*, 5823. [[CrossRef](#)] [[PubMed](#)]
58. Heat Map—Wikipedia. Available online: https://en.wikipedia.org/wiki/Heat_map (accessed on 22 February 2021).
59. Rosner, B. *Fundamentals of Biostatistics*, 6th ed.; Thomson-Brooks/Cole: Belmont, CA, USA, 2006.