Multigroup prediction in lung cancer patients and comparative controls using signature of volatile organic compounds in breath samples

Shesh N. Rai1,2,3,4,5,6*, Samarendra Das1,2,7,8,9*, Jianmin Pan1, Dwijesh C. Mishra9, Xiao-An Fu10

1 Biostatistics and Bioinformatics Facility, Brown Cancer Center, University of Louisville, Louisville, KY, United States of America, 2 School of Interdisciplinary and Graduate Studies, University of Louisville, Louisville, KY, United States of America, 3 Hepatology and Toxicology Center, University of Louisville, Louisville, KY, United States of America, 4 Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, United States of America, 5 Biostatistics and Informatics Facility, Center for Integrative Environmental Research Sciences, University of Louisville, Louisville, KY, United States of America, 6 Christina Lee Brown Envirome Institute, University of Louisville, Louisville, KY, United States of America, 7 ICAR-Directorate of Foot and Mouth Disease, Arugul, Bhubaneswar, Odisha, India, 8 International Centre for Foot and Mouth Disease, Arugul, Bhubaneswar, Odisha, India, 9 ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi, India, 10 Department of Chemical Engineering, University of Louisville, Louisville, KY, United States of America

* These authors contributed equally to this work.

Abstract

Early detection of lung cancer is a crucial factor for increasing its survival rates among the detected patients. The presence of carbonyl volatile organic compounds (VOCs) in exhaled breath can play a vital role in early detection of lung cancer. Identifying these VOC markers in breath samples through innovative statistical and machine learning techniques is an important task in lung cancer research. Therefore, we proposed an experimental approach for generation of VOC molecular concentration data using unique silicon microreactor technology and further identification and characterization of key relevant VOCs important for lung cancer detection through statistical and machine learning algorithms. We reported several informative VOCs and tested their effectiveness in multi-group classification of patients. Our analytical results indicated that seven key VOCs, including C11H22O2, C13H22O, C11H22O, C2H4O2, C7H14O, C6H12O, and C6H8O, are sufficient to detect the lung cancer patients with higher mean classification accuracy (92%) and lower standard error (0.03) compared to other combinations. In other words, the molecular concentrations of these VOCs in exhaled breath samples were able to discriminate the patients with lung cancer (n = 156) from the healthy smoker and nonsmoker controls (n = 193) and patients with benign pulmonary nodules (n = 65). The quantification of carbonyl VOC profiles from breath samples and identification of crucial VOCs through our experimental approach paves the way forward for non-invasive lung cancer detection. Further, our experimental and analytical approach of VOC quantitative analysis in breath samples may be extended to other diseases, including COVID-19 detection.
Introduction

Lung cancer is the most common kind of cancer globally due to poor lifestyle and environmental pollution. In the USA, lung cancer has an incidence of 235,760 new cases with 131,880 deaths in 2020 and is by far the leading cause of cancer deaths among both men and women, making up almost 25% of all cancer fatalities [1]. Every year, a larger number of people die due to lung cancer than colon, breast, and prostate cancers combined [2]. The 5-year survival rate for lung cancer patients is relatively low compared to other diseases. Early detection of lung cancer is crucial to improve the survival of the patients [2]. For this purpose, many screening methods, including chest radiography, sputum cytology, low-dose spiral computer tomography, fluorescence bronchoscopy, and positron emission tomography have been used [3]. However, these procedures are complicated, expensive, and time-consuming, thus making them difficult for the poor or low-income patient groups. Therefore, lung cancer detection through breath analysis using volatile organic compounds (VOCs) can be a vital tool for its early detection in order to increase the survival chance.

VOCs are organic chemicals that exist in air, exhaled breath, etc. and have high vapor pressure at ambient temperature [4]. The analysis of VOCs in exhaled breath plays a vital role in the early detection of lung cancer [4]. Earlier studies showed that there are several thousand(s) of VOCs in human breath, e.g., formaldehyde (CH$_2$O), acetaldehyde (C$_2$H$_4$O), acetone (C$_3$H$_6$O), 2-butanone (C$_4$H$_8$O), to name a few. So far, many VOCs have been identified and reported in breath samples of normal humans and patients with lung cancer [5]. For instance, in 1999, Phillips et al. reported over 3400 different VOCs present in exhaled normal human breath [6].

The combinations of some VOCs reported in the literature can comfortably discriminate the lung cancer patients from the healthy ones [7]. For instance, a recent study showed higher sensitivity of lung cancer detection among patients using the carbonyl VOCs present in the exhaled breath samples [8]. In other words, the molecular concentration data on the VOCs present in exhaled breath samples were collected from the patients to classify them into several groups, i.e., Healthy Control, Cancer, and Benign Nodule. In this direction, Bousamra et al. (2014) developed one quantitative analytical approach to differentiate early lung cancer from benign pulmonary nodules based on the molecular concentration of carbonyl compounds in exhaled breath samples [9]. In the existing analytical approaches, statistical methods including the t-test and Wilcoxon test were used to select the significant VOCs [7, 9]. These methods are univariate in nature and mostly ignore the inter-VOC relationships (i.e., relationship among the VOCs), and there is a chance of spurious association of VOC with the patient classes [16]. Further, Li et al. (2015) developed a technique utilizing quaternary amino-oxy coated silicon micro-reactors [4] for selective capture and quantification of the ketones and aldehydes in the air [10–12] and exhaled breath [7–9]. They used five classification models, including generalized partial least squares, support vector machines, random forests, linear and quadratic discriminant analyses to classify the patients into lung cancer and control groups based on exhaled breath data [4]. Sufficient studies in the literature indicated that the carbonyl compounds in exhaled breath play a significant role in the non-invasive detection of lung cancer [7–12]. Furthermore, the proper identification of key carbonyl compounds through statistical and machine learning techniques requires further advances.

In a typical breath sample study, the molecular concentration data on several hundred(s) of endogenous and exogenous VOCs are usually observed over patients. Sometimes, it may not be experimentally possible to monitor data on all the VOCs and further use them in lung cancer detection. In other words, among these hundred(s) of VOCs, all may not be required for the patient classification or the predictive model building process (i.e., training the machine
learning models and later use them for class label predictions). Therefore, it is pertinent to select/identify a few metabolic VOCs related to lung cancer as key and significant features for cancer detection. The selection of important features (here metabolic VOCs) out of many VOCs is called feature selection in machine learning [13]. Further, it is essential to determine the number of significant VOCs (e.g., feature size or dimension of VOC data), which can be used in the training of the classification model to predict the class type of lung cancer patients. The selection of significant VOCs will save the precious time and cost of data generation for all VOCs present in the breath samples. In other words, the researchers can focus on few VOCs instead of generating data on all the VOCs present in breath samples of the patients.

Therefore, in this study, we endeavor to classify the lung cancer patients based on the VOC molecular concentration data. We present an experimental approach for lung cancer prediction using the carbonyl VOCs present in breath samples. The VOC molecular concentration data are generated from the breath samples of the 414 subjects (156: lung cancer; 65: benign and 193: healthy control) through the unique silicon microreactor technology [8]. The breath samples are collected following the experimental protocol approved by the Institutional Review Board (IRB) at the University of Louisville, USA. We also present an analytical approach involving relevant VOC selection and further used them in lung cancer patient classification model training. This approach of VOC selection is statistically sound, robust, and does not require any probability distributional assumptions about the VOC data (for VOCs testing and selection). Further, we identified several informative VOCs present in exhaled breath samples to detect lung cancer patients. For instance, the developed models provided sufficient classification accuracy for lung cancer detection with a minimum of three VOCs. Also, we studied the effect of the various VOC combinations on the classification of lung cancer patients. Moreover, our developed experimental approach can be applied to detect COVID-19 patients using VOC data from the exhaled breath samples.

The remainder of the paper is organized as follows: (i) the material and methods section deals with detail protocols for the data generation, description of methodology; (ii) the results and discussion section mainly deals with presentation of obtained results along with their discussion; and (iii) the conclusion section summarizes the manuscript along with its future scope.

**Materials and methods**

**Breath sampling and data generation**

This study recruited 156 patients with untreated lung cancer, 65 patients with benign pulmonary nodules, and 193 healthy control subjects to provide exhaled breath samples. The detailed subject demographic characteristics, disease information, and breath analysis data have been published [8]. In brief, there were 103 patients with early stages (Stages 0, I, and II) of lung cancer and 53 patients with late stages of lung cancer. Most lung cancer patients were current or former smokers (149). The healthy controls included 113 current or former smokers and 80 never smokers. The average ages of the lung cancer patients, patients with benign pulmonary nodules, and healthy subjects were 65.1, 54.2, and 49.4 years, respectively. The male percentages of these three subgroups were 51.9%, 49.2%, and 55.7%, respectively.

The detailed research protocol for the collection of exhaled breath samples was approved by the IRB, University of Louisville, USA (IRB #15.0711). The healthy control subjects were recruited from patient family members who were free of lung cancer or other chronic pulmonary disease. All patients with pulmonary nodules were recruited in the James Graham Brown Cancer Center and Jewish Hospital at the University of Louisville, USA. The diagnostic predictions from these breath analyses were confirmed by clinical diagnoses using following-up the
CT scans, positron emission tomography scans or pathology of biopsy or surgically resected specimens.

Exhaled breath samples were collected in one-liter Tedlar bags (Sigma-Aldrich, USA) through normal exhalation allowing for the collection of a mixture of alveolar and tidal breath in one exhaled breath. Ambient clinic exam room air samples (1 L) were also collected to serve as a control of background carbonyl compounds in the collection room. Our previous studies have examined the detailed method of breath sample collection, evacuation of breath samples through the micro-reactors, and analysis of the samples [4, 7–9]. In a brief description, the subjects directly exhaled the breath into Tedlar bags through the Teflon tube from the mouth to provide one liter exhaled breath samples, thus providing a non-invasive collection technique that the patients readily accepted. After the collection of exhaled breath, the Tedlar bags were directly connected to the silicon micro-reactors through silica capillary tubes and septa. A vacuum was applied to draw the collected VOCs from the Tedlar bags through the fabricated microreactor [10–12] at a flow rate of 5 mL/min. The microreactor has thousands of micropillars coated with 2-(aminooxy)-N, N, N-trimethyl-ethan ammonium (ATM) iodide. After complete deflation of the sample bags, ATM and ATM adducts were eluted by flowing methanol (~100 μL) from a pressurized vial through the microreactor and into a collection vial [10]. The eluent methanol solutions were directly analyzed using a hybrid linear ion trap-Fourier transform-ion cyclotron mass spectrometry (FT-ICR-MS) instrument (Finnigan LTQ-FT, Thermo Electron, Bremen, Germany) equipped with a TriVersa NanoMate ion source (Advion Biosciences, Ithaca, NY) and a nano-electrospray chip (inner nozzle diameter 5.5 μm). A 5 μL methanol solution of a known amount of 5 nmol of ATM–acetone-d6 adduct was added to each eluted methanol sample as an internal reference of FT-ICR-MS. The amount of captured carbonyl compounds was then determined by comparing the FT-ICR-MS signal abundance of ATM–acetone-d6 with those of other ATM–carbonyl adducts. The concentration of each compound in exhaled breath detected by FT-ICR-MS was then calculated from the amount of the captured carbonyl compounds with in terms of nanomoles per liter (nmole/L). The micro-reactor’s carbonyl capture efficiencies and validity of the analyses have been characterized in our previous studies [10–12]. A flow chart of the bioassay for generating data is displayed in Fig 1. Further, the workflow of the proposed experimental approach, including data generation, feature selection, and the classification model development is also illustrated in Fig 1.

This study was approved by the IRB at the University of Louisville (IRB #15.0711). An informed written consent form was reviewed by the subjects for participation. A signed

---

**Fig 1. Outline of the experimental approach used in this study.** Various steps undertaken in this study are shown in flow chart form. The various steps include (i) capturing of exhaled breath; (ii) capturing VOCs present in exhaled breath samples; (iii) generation of molecular concentration data (nmole/L) of VOCs through bioassay mass-spectrometry technique; (iv) feature (VOC) selection and classification model training and its application for lung cancer detection; and (v) validation of the selected VOCs through classification accuracy, literature search, and expert opinion.

https://doi.org/10.1371/journal.pone.0277431.g001
consent form was obtained from the subjects before they participated in the study. There was no minor recruited for the study.

**Notation**

Let, $X_{N \times M} = [x_{i, m}]$ be the VOC data matrix, where $x_{i, m}$ represents the observed measurement of the $i^{th}$ ($i = 1, 2, \ldots, N$) VOC for the $m^{th}$ ($m = 1, 2, \ldots, M$) patient; $x_m$ be the $N$-dimensional vector of observed values of $N$ VOCs for $m^{th}$ patient; $y_m$ be the outcome variable for target class label of $m^{th}$ patient (i.e., control and lung cancer) and takes values $\{+1, -1\}$ for lung cancer and control conditions respectively; $M_1$ and $M_2$ be the number of patients in lung cancer and control classes respectively ($M_1 + M_2 = M$).

**Support Vector Machine-Recursive Feature Elimination**

Support Vector Machine-Recursive Feature Elimination (SVM-RFE) can be used for the selection of relevant VOCs (in a two group case) from the lung cancer VOC data [14, 15]. Let, $\{x_m, y_m\} \in \mathbb{R}^N \times \{-1, 1\}$ be the input given to the Support Vector Machine (SVM) model. Here, we wish to find a hyperplane that divides the patients for lung cancer ($y_m = 1$) from that of control class ($y_m = -1$) in such a way that the distance between the hyperplane and the closest point is maximum (Supplementary Document S1 in S1 File). Then the hyperplane can be written as:

$$k \cdot x + b = 0$$

(1)

where, $k_i$ and $b$ are the weight of $i^{th}$ VOC and bias, respectively. Here, we assume that the patients for the two classes are linearly separable. In other words, we can select two parallel support hyperplanes that separate the lung cancer and control classes in such a way that the distance between these two planes is maximized (Supplementary Document S1 in S1 File).

For the lung cancer class ($y_m = 1$), the supporting hyperplane can be written as:

$$k \cdot x + b = 1$$

(2)

Here, the Eq 2 only holds good for the $x$ which are support vectors (i.e., points closest to the separating hyperplane, Eq 1).

For control class ($y_m = -1$), the support hyperplane can be written as:

$$k \cdot x + b = -1$$

(3)

Now, we assume that every point must lie on either side of the respective support hyperplanes (Eq 2 and Eq 3) (their distance to the separating hyperplane is at least same as the distance between the support vectors and the separating hyperplane), which can be expressed as:

$$y_m(k \cdot x_m + b) \geq 1 \quad \forall m = 1, 2, \ldots, M$$

(4)

We wish to maximize the distance between the lung cancer and control support hyperplanes given in Eqs 2 and 3, respectively. Here, special care must be taken to prevent any data points from falling in between these two support hyperplanes.

To maximize the distance between the support hyperplanes (Supplementary Document S1 in S1 File), we need to minimize $\frac{||k||^2}{2}$ under the constraint of Eq 4. Mathematically, the objective function can be written as:

$$L_p = \min_k \left[ \frac{||k||^2}{2} + \sum_{m=1}^{M} \varphi_m \{1 - y_m(k \cdot x_m + b)\} \right] \quad \forall m = 1, 2, \ldots, M$$

(5)

where, $\varphi_m \geq 0$ are Lagrange multipliers. The constraint on $\varphi_m$ follows from the fact that the
constraint is an inequality. Here, \( k_i \)'s are obtained by minimizing the objective function in Eq 5. The objective function (Eq 5) was optimized with respect to \( k, b \) and the following expressions are obtained.

\[
\frac{\partial L_p}{\partial k} = ||k|| - (\sum_{m=1}^{M} \varphi_m y_m x_m) = 0 \quad \text{and} \quad \frac{\partial L_p}{\partial b} = \sum_{m=1}^{M} \varphi_m y_m = 0
\]  

(6)

The value of \( k \) can be obtained through solving the system of linear equations given in Eq 6 and is expressed as:

\[
k = \sum_{m=1}^{M} \varphi_m y_m x_m \quad \text{with} \quad \sum_{m=1}^{M} \varphi_m y_m = 0 \quad \text{and} \quad \varphi_m \geq 0
\]  

(7)

Here, one also need to solve for \( \varphi_m \), but this cannot be done by solving for Eq 6 when its gradient is zero, thus one need to take into account the constraint that \( \varphi_m \geq 0 \). Therefore, SVMs cannot be solved via a linear system of equations. Rather, optimization algorithms (e.g., gradient descent) are commonly used in SVM-based tools. For this purpose, we executed the svm function implemented in e1071 R package.

The \( k_i^2 (\geq 0) \) (i.e., square of the \( i^{th} \) element of \( k \) in Eq 7) is used as a metric for the ranking of the VOCs in the data [15] to evaluate the impact of the \( i^{th} \) VOC on patients’ classification [16]. In the SVM-RFE technique, VOCs are eliminated with the smallest \( k_i^2 \) iteratively in a backward elimination manner and ranked VOC list is prepared at the end [15, 17]. Here, the backward elimination means training the SVM-based machine learning models iteratively after removing the least significant VOC at each step. Moreover, most feature selection methods, such as SVM-RFE, are sensitive to slight permutations of the class labels [18]. The ranking of VOCs may also lead to the selection of spuriously associated VOCs and make the selection process unreliable [13, 19, 20]. Therefore, it is essential to select VOCs based on statistical testing instead of their ranks. For this purpose, we used the Bootstrap SVM-RFE (Boot-SVM-RFE) technique developed by Das et al. (2017) to select the most relevant VOCs for patient classification [19]. The Boot-SVM-RFE method is briefly described in the following section.

**Bootstrap SVM-RFE**

In the usual supervised setting, the \( M \) patients/samples (as columns) in the \( X \) data matrix, either belonging to lung cancer or control, can be considered subjects/units in a population model, as shown in Eq 8.

\[
(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m), \ldots, (x_{M-1}, y_{M-1}), (x_M, y_M)
\]  

(8)

where, \( x_m \) is the \( N \)-dimensional vector of the VOC measurements of the \( m^{th} \) patient/sample and \( y_m \) the class label (e.g., cancer vs. control) \( m^{th} \) patient/sample. Here, it is assumed that the \( M \) measurements of the different patients/samples (i.e., \( x_m, \forall m = 1, 2, \ldots, M \)) are independent and identically distributed (iid), but the VOCs within the same sample may be correlated. In the bootstrap procedure, \( M \) units are randomly drawn from \( M \) population units in Eq 8 with replacement to constitute a bootstrap data matrix, i.e., \( X_{N \times M}^{(b)} \) (\( M \) units serve as \( M \) columns of \( X \)). This process is repeated \( B \) times to get \( B \) bootstrap data matrices, i.e., \( X_{N \times M}^{(1)}, X_{N \times M}^{(2)}, \ldots, X_{N \times M}^{(b)}, \ldots, X_{N \times M}^{(B)} \). Here, \( B \) (i.e., number of bootstrap samples) depends on several factors, such as the number of units in the population model in Eq 8 [15, 19]. So, we set \( B = 500 \) as the literature showed that the number of bootstrap samples required for obtaining the distribution of the test statistic(s) must be sufficiently large [18, 21]. Furthermore, the values of \( M \) for different classification problems including Cancer vs. Control, Cancer vs. Benign,
Benign vs. Control, (Benign + Cancer) vs. Control, and Cancer vs. (Benign + Control) are 349, 221, 258, 414, and 414 respectively.

The $B$ bootstrap data matrices are given as input to the SVM-RFE technique to compute the $k_i^2$ scores (given in Eq 7), and subsequently, VOC ranking was performed on each of the $B$ bootstrap data matrices. Let, $P_{ib}$ be a random variable (rv) that indicates the position of $i^{th}$ VOC (i.e., ranks obtained from the SVM-RFE) in the $b^{th}$ bootstrap data matrix. Then, another rv can be defined based on $P_{ib}$ (without loss of generality), given as:

$$R_{ib} = \frac{N + 1 - P_{ib}}{N}; \quad 0 \leq R_{ib} \leq 1$$  \hspace{1cm} (9)

where, $R_{ib}$ in Eq 9 is the rank score of $i^{th}$ VOC in the $b^{th}$ bootstrap data matrix. Here, it may be noted that the distribution of the rank scores of VOCs, computed from a bootstrap data matrix, is symmetric around the median value, 0.5 (as rank scores are functions of ranks).

To decide whether $i^{th}$ VOC is relevant or not for the patient classification, the following null hypotheses are framed.

$H_0$: $R_i \leq 0.5$ ($i^{th}$ VOC is not so relevant to patient classification)

$H_1$: $R_i > 0.5$ ($i^{th}$ VOC is relevant to patient classification)

where, $R_i$ is the rank score for $i^{th}$ VOC over all possible bootstrap samples.

To obtain the distribution of test statistic under $H_0$, we defined another rv $Z_{ib}$, as:

$$Z_{ib} = \begin{cases} 
1 & \text{if } (R_{ib} - 0.5) > 0 \\
0 & \text{if } (R_{ib} - 0.5) \leq 0 
\end{cases}$$

Let $r_{ib}$ be another rv representing the rank assigned to $(R_{ib} - 0.5)$ (after arranging in ascending order of their magnitudes). In other words, for each row (which denotes the VOC), $(R_{ib} - 0.5)$ was computed and corresponding rank, $r_{ib}$, was assigned to $(R_{ib} - 0.5)$.

To test $H_0$ vs. $H_1$, the test statistic for $i^{th}$ VOC, $W_i$, is developed and is given as:

$$W_i = \sum_{b=1}^{B} U_{ib}$$  \hspace{1cm} (11)

where, $U_{ib} = Z_{ib}r_{ib}$.

In other words, $W_i$ (Eq 11) is the sum of the ranks of positive $(R_{ib} - 0.5)$ for $i^{th}$ VOC over $B$ bootstrap samples. Further, $U_{ib}$ in Eq 11 is a Bernoulli rv, and its probability mass function can be given as:

$$P[U_{ib} = u_{ib}] = \begin{cases} 
\frac{1}{2} & \text{if } u_{ib} = 0 \\
\frac{1}{2} & \text{if } u_{ib} = b 
\end{cases}$$

(12)

Here, the expected value and variance of $W_i$ in Eq 11 under $H_0$ can be obtained as:

$$E(W_i) = \sum_{b=1}^{B} E(U_{ib}) = \frac{1}{2} \sum_{b=1}^{B} b = \frac{B(B + 1)}{4}$$

(13)
The variance of $W_i$ becomes:

$$V(W_i) = \sum_{b=1}^{n} \{E(U_{ab}^2) - E(U_{ab})^2\}$$

$$= \frac{1}{4} \sum_{b=1}^{n} b^2 = \frac{B(B + 1)(2B + 1)}{24} \tag{14}$$

As $B$ is sufficiently large, then under the central limit theorem, the distribution of $W_i$, given in Eq 11, becomes:

$$Z_i = \frac{W_i - E(W_i)}{\sqrt{V(W_i)}} \xrightarrow{d} N(0,1) \tag{15}$$

Through Eq 15, the $p$-value for $i^{th}$ ($i = 1, 2, \ldots, N$) VOC was computed, and similarly, this testing procedure was repeated for the remaining N-1 VOCs. In other words, the above statistical test was repeated for $N$ times to compute the statistical significance values for the VOCs.

Let, $p_1, p_2, \ldots, p_N$ be the corresponding $p$-values for all the VOCs, and $\alpha$ be the desired significance level. Hence, we employed Hochberg’s procedure [1] to correct the multiple testing problem and computed the adjusted (adj.) $p$-values for the VOCs. The algorithm for Hochberg’s procedure [22] is as follows:

First, the $p$-values of the VOCs were sorted in increasing order of their magnitude, shown as: $p_{(1)}, p_{(2)}, \ldots, p_{(N)}$, where, $p_{(i)}$ is the rank ordered $p$-value and $(i)$ stands for the $i^{th}$ ordered value (i.e., $p_{(1)}$: smallest $p$-value, $p_{(2)}$: second smallest $p$-value and so on).

**Step 1.** If $p_{(1)} > \alpha$, then retain corresponding null hypothesis and go to the next step. Else reject it and stop.

**Step $i = 2, 3, \ldots, N-1$.** If $p_{(i-1)} > \alpha/i$, then retain the corresponding null hypothesis and go to the next step. Else reject all remaining hypotheses and stop.

**Step N.** If $p_{(1)} > \alpha/N$, then retain the corresponding null hypothesis. Else reject it.

Now, the adj. $p$-values are given recursively beginning with the largest $p$-value [1]:

$$\hat{p}_{(i)} = \begin{cases} 
   p_{(i)} & \text{if } i = N \\
   \min(\hat{p}_{(i+1)}, (N-i+1)p_{(i+1)}) & \text{if } i = N-1, \ldots, 1 
\end{cases} \tag{16}$$

Based on the computed adj. $p$-values, the relevant VOCs were selected from the data. In other words, a lesser value of adj. $p$-value indicates more relevance of the VOC for the patient classification and vice-versa. Similarly, this procedure was applied to select the significant VOC for the patient classification into classes, such as benign vs. control, (lung cancer + benign) vs. control, (control + benign) vs. lung cancer. The outline and key analytical steps of the VOC selection process are shown in Fig 2. The effects of the significant VOCs on the patient classification were studied through an SVM classifier (with linear basis function). Under this setting, the VOCs (and their molecular concentration data) were given as inputs to the SVM classifier to compute the classification-based performance metrics. Further, impacts of the VOCs on the classification of patients under five different cases, including Case I: Cancer vs. Control; Case II: Cancer vs. Benign; Case III: Benign vs. Control; Case IV: (Benign + Cancer) vs. Control; and Case V: (Control + Benign) vs. Cancer were assessed through performance metrics, such as mean classification accuracy (CA) and standard error (SE) in CA computed through varying sliding window size technique [18, 19]. Here, we used this technique to study the importance of rankings of VOCs (obtained from a feature selection method, e.g., Boot-SVM-RFE) through training a classification model. In other words, the sliding
Fig 2. Outline of the Boot-SVM-RFE technique for significant VOC selection. Data matrix has \(N\) rows as VOCs with concentrations of \(\text{nmole/L}\) and \(M\) columns as patients/samples. \(B\) is the number of bootstrap samples drawn from the data matrix. \(V_b\) and \(R_b\)'s are the VOC ranked list and VOC rank-score for the \(b^{th}\) bootstrap sample. Null hypothesis represents \(i^{th}\) VOC is not important for patients' classification. \(\alpha\) is the desired level of significance.

https://doi.org/10.1371/journal.pone.0277431.g002
windows are VOC intervals that literally "slide" across the whole VOC list, preferably by some constant distance and CA is computed for each sliding window. Sliding windows can overlap or mutually exclusive. A brief description about the varying sliding window size technique is given in Supplementary Document S2 in S1 File. Then, the mean CA and SE in CA were computed over the sliding windows. In other words, the VOC set (of size \( n \leq N \)), which provides maximum discrimination between the subjects/patients of 2 groups through CA, will be considered as the optimal size of the VOC set. The expressions for mean CA and SE in CA computed through varying sliding window size technique are given in Eqs 17 and 18.

\[
\mu_{CA} = \frac{\sum_{k=1}^{K} CA_k}{K}
\]

\[
SE_{CA} = \sqrt{\frac{\sum_{k=1}^{K} (CA_k - \mu_{CA})^2}{K(K-1)}}
\]

The total number of windows denoted as \( K \) in Eqs 17 and 18 can be defined in Eq 19.

\[
K = \frac{(n - S)}{L}
\]

where, \( n \) is the size of the VOC set, \( S \) is the size of the windows (i.e., size refers to the number of ranked VOCs), and \( L \) is the sliding length. The values of \( n, S, L, \) and \( K \) are given in Supplementary Document S2 in S1 File. The optimal size of the VOC set was computed through training the SVM-based classification model (i.e., calculating the indices given in Eqs 17 and 18) using the five-fold cross-validation technique. The outline and workflow of the Boot-SVM-RFE for the key VOC selection are shown in Fig 2.

**Results and discussion**

We observed the carbonyl VOCs having carbon atoms ranging from one (formaldehyde: CH\(_2\)O) to thirteen (tridecanal: C\(_{13}\)H\(_{26}\)O) in exhaled breath samples of healthy controls, benign nodules, and lung cancer patients. Through FT-ICR-MS technology, the molecular concentration data on VOCs with chemical formulas CH\(_2\)O, C\(_2\)H\(_4\)O, C\(_3\)H\(_6\)O, C\(_4\)H\(_8\)O, C\(_5\)H\(_{10}\)O, C\(_6\)H\(_{12}\)O, C\(_7\)H\(_{14}\)O, C\(_8\)H\(_{16}\)O, C\(_9\)H\(_{18}\)O, C\(_{10}\)H\(_{20}\)O, C\(_{11}\)H\(_{22}\)O, C\(_{12}\)H\(_{24}\)O, C\(_{13}\)H\(_{26}\)O, C\(_{14}\)H\(_{28}\)O, C\(_{15}\)H\(_{30}\)O, C\(_{16}\)H\(_{32}\)O, C\(_{17}\)H\(_{36}\)O, C\(_{18}\)H\(_{40}\)O, C\(_{19}\)H\(_{42}\)O, C\(_{20}\)H\(_{44}\)O, C\(_{21}\)H\(_{48}\)O, C\(_{22}\)H\(_{52}\)O, C\(_{23}\)H\(_{56}\)O, C\(_{24}\)H\(_{60}\)O, C\(_{25}\)H\(_{64}\)O, C\(_{26}\)H\(_{68}\)O, C\(_{27}\)H\(_{72}\)O, C\(_{28}\)H\(_{76}\)O, C\(_{29}\)H\(_{80}\)O, C\(_{30}\)H\(_{84}\)O, C\(_{31}\)H\(_{88}\)O, C\(_{32}\)H\(_{92}\)O, C\(_{33}\)H\(_{96}\)O, C\(_{34}\)H\(_{100}\)O, and C\(_{35}\)H\(_{104}\)O, were observed for healthy controls, benign nodules, and lung cancer patients. The common names of these VOCs are provided in Supplementary Document S3 in S1 File. The FT-ICR-MS spectra in Fig 1 showed the capturing of typical relative abundances of these VOCs in breath samples of lung cancer patients, benign, and healthy controls. Chemical structure-wise, isomeric ketones and aldehydes are indistinguishable by direct infusion of one-dimensional FTICR-MS. However, the measured molecular weight at a resolving power of 200,000 provides their accurate chemical formulas. Separation and structure identification of some important isometric ketones and aldehydes was done through FT-ICR-MS/MS and GC–MS technologies. Further, the summary statistic(s) of the 27 considered VOCs for three different patient classes, i.e., control, benign, and lung cancer, are given in S1 Table. The mean values of the VOCs, including C\(_3\)H\(_6\)O, C\(_2\)H\(_4\)O, etc., are higher than others across all the patient classes, which indicated their higher average concentrations in breath samples (S1 Table).

All the 27 VOCs may not be considered as biomarkers for lung cancer detection; therefore, we used the Boot-SVM-RFE technique to identify the significant VOCs for various combinations of patient classes. Here, five different classes were used, i.e., Case I: Cancer (156) vs.
Control (193); Case II: Cancer (156) vs. Benign (65); Case III: Benign (65) vs. Control (193); Case IV: (Benign + Cancer) (221) vs. Control (193); Case V: (Control + Benign) (258) vs. Cancer (156) (Fig 3). Through the Boot-SVM-RFE, the statistical significance values (p-values) for the VOCs were computed for five different classification problems and shown in Table 1 and S2 Table. For instance, the VOCs, including $\text{C}_4\text{H}_8\text{O}$, $\text{C}_6\text{H}_{12}\text{O}$, $\text{C}_7\text{H}_{14}\text{O}$, $\text{C}_8\text{H}_{16}\text{O}$, etc., were found to be statistically significant (at 1% level of significance) for all the cases of classification (Table 1). This finding indicated that these VOCs can be used as biomarkers for the patient classification. Broadly, we found 16 VOCs as statistically significant, at least in one case of patients’ classification (Table 1). The summary statistic(s) for these key VOCs for the three patients’ classes are shown in Table 2. The co-efficient variation of the VOCs, including $\text{C}_3\text{H}_4\text{O}$, $\text{C}_6\text{H}_{10}\text{O}_2$, $\text{C}_9\text{H}_{16}\text{O}_2$, $\text{C}_5\text{H}_8\text{O}$, $\text{C}_7\text{H}_{11}\text{O}$, and $\text{C}_{13}\text{H}_{22}\text{O}$, are quite high as compared to others (Table 2). This observation indicated that the molecular concentrations of these VOCs have greater dispersion levels than others across all the classes (Table 2). Similar interpretations can be made for other VOCs.

The impact of the VOCs on different patients’ classification, such as cancer vs. control, cancer vs. benign, benign vs. control, (benign + cancer) vs. control, and (control + benign) vs. cancer, was studied through the Boot-SVM-RFE technique. The false discovery rates (FDR) for all the VOCs were computed through the Boot-SVM-RFE for different patients’ classifications and shown in Table 1. For cancer vs. control classification, 13 VOCs, including $\text{C}_4\text{H}_8\text{O}$ and $\text{C}_4\text{H}_8\text{O}_2$, are found to be statistically significant when assessed through the FDR values (Fig 3, Table 2). Similarly, for cancer vs. benign, benign vs. control, and (benign + cancer) vs. control classifications, 15 VOCs were found to be statistically significant at the 1% level of significance (Fig 3, Table 1). Further, for (control + benign) vs. cancer classification, 13 VOCs were found to be statistically significant (Table 1). The rankings of the VOCs for five different patient
classification problems are shown in Table 3. For instance, the VOCs such as C₄H₈O and C₄H₈O₂ ranked 1 and 2 respectively for cancer vs. control classification (Table 3). These ranked VOCs can be used as biomarkers for cancer detection with respect to healthy controls. Similar interpretations about the ranking of the VOCs on other classifications can be made, as shown in Table 3. The empirical distributions of the significant VOCs are shown in Supplementary Document S4 in S1 File.

We also studied different combinations of the VOCs on classifying the patients into various classes, such as cancer vs. control, cancer vs. benign, benign vs. control, (benign + cancer) vs. control, and (benign + control) vs. cancer, through the SVM based classification model. Here, the classification accuracy was computed through a five-fold cross-validation technique for each classification problem. The cross-validations of the data were repeated 500 times by taking different combinations of VOCs (based on their ranks) for each classification problem. Then, the mean and standard error of the classification accuracies were computed over the 500 runs through a varying sliding window size technique [13, 16, 19]. The results are shown in

### Table 1. Effects of the VOCs on lung cancer classifications.

|            | Case I Adj. Pval | Case II Adj. Pval | Case III Adj. Pval | Case IV Adj. Pval | Case V Adj. Pval |
|------------|------------------|-------------------|--------------------|-------------------|------------------|
|            | VOCs             | FDR               | FDR                | FDR               | FDR              |
| CH₄O       | NS               | NS                | NS                 | NS                | NS               |
| C₂H₄O      | NS               | NS                | NS                 | NS                | NS               |
| C₃H₆O      | NS               | NS                | NS                 | NS                | NS               |
| C₄H₈O      | 1.58E-82         | 5.6E-83           | 8.20E-31           | 1.2E-31           | 1.06E-06         |
|            | 1.5E-07          | 1.05E-07          | 2.4E-73            | 3.56E-80          | 1.28E-80         |
| C₅H₁₀O     | NS               | NS                | NS                 | NS                | NS               |
| C₆H₁₂O     | 2.43E-49         | 3.8E-50           | 3.54E-45           | 5.9E-46           | 2.23E-15         |
|            | 3.1E-16          | 7.3E-28           | 1.1E-25            | 8.44E-45          | 1.42E-45         |
| C₇H₁₄O     | 2.26E-74         | 4.2E-75           | 4.46E-55           | 9.1E-56           | 5.7E-56          |
|            | 1.1E-56          | 7.23E-70          | 1.5E-70            | 8.24E-50          | 1.51E-50         |
| C₈H₁₆O     | 1.83E-23         | 2.6E-24           | 2.96E-15           | 4.2E-16           | 4.60E-18         |
|            | 6.5E-19          | 9.74E-14          | 1.3E-14            | 7.55E-05          | 1.05E-05         |
| C₉H₁₈O     | NS               | NS                | 0.4940             | 0.13             |
| C₁₀H₂₀O    | NS               | NS                | NS                 | 1.31E-30          | 2.1E-31          |
|            | 4.82E-21         | 6.8E-22           | NS                 | NS               | NS               |
| C₁₁H₂₂O    | 2.41E-80         | 6.8E-81           | 9.94E-15           | 1.4E-15           |
|            | 1.94E-81         | 6.9E-82           | 2.39E-82           | 8.6E-83           |
|            | 2.49E-54         | 5.10E-55          | NS                 | NS               | NS               |
| C₁₂H₂₄O    | NS               | 0.3069            | 1.78E-71           | 6.4E-72           |
|            | 1.44E-59         | 3.4E-60           | 7.01E-29           | 1.1E-29           |
|            | 0.648013         | 0.089725          | NS                 | NS               | NS               |
| C₁₃H₂₆O    | 3.72E-27         | 5.6E-28           | 3.76E-52           | 6.9E-53           |
|            | 9.68E-20         | 1.4E-20           | 4.94E-12           | 6.8E-13           |
|            | 9.41E-28         | 1.41E-28          | NS                 | NS               | NS               |
| C₁₄H₂₈O    | 1.58E-82         | 5.7E-83           | 4.21E-67           | 1.2E-67           |
|            | 1.67E-82         | 8.6E-83           | 1.64E-82           | 8.5E-83           |
|            | 1.84E-82         | 1.21E-82          | NS                 | NS               | NS               |
| C₁₅H₃₀O    | 4.43E-25         | 6.4E-26           | 1.80E-32           | 2.8E-33           |
|            | 5.4E-24          | 1.8E-25           | 1.33E-15           | 1.8E-16           |
|            | 3.2E-08          | 4.51E-09          | NS                 | NS               | NS               |
| C₁₆H₃₂O    | 7.01E-76         | 1.4E-76           | 8.05E-77           | 4.2E-77           |
|            | 0.037348         | 0.00571           | 5.06E-77           | 1.4E-77           |
|            | 2.34E-82         | 1.21E-82          | NS                 | NS               | NS               |
| C₁₇H₃₄O    | 1.25E-54         | 2.1E-55           | 6.15E-12           | 8.5E-13           |
|            | 5.90E-57         | 1.2E-57           | 3.31E-66           | 6.1E-67           |
|            | 6.07E-21         | 8.76E-22          | NS                 | NS               | NS               |
| C₁₈H₃₆O    | NS               | NS                | NS                 | NS               |
| C₁₉H₃₈O    | NS               | NS                | NS                 | NS               |
| C₂₀H₄₀O    | NS               | NS                | NS                 | NS               |
| C₂₁H₄₂O    | 1.75E-14         | 2.4E-15           | 5.06E-82           | 5.1E-82           |
|            | 2.12E-53         | 3.6E-54           | 7.32E-29           | 1.1E-29           |
|            | 3.83E-33         | 6.04E-34          | NS                 | NS               | NS               |
| C₂₂H₴₄O    | NS               | NS                | NS                 | NS               |
| C₂₃H₴₆O    | 0.887738         | 0.12229           | 3.57E-09           | 4.9E-10           |
|            | NS               | NS                | NS                 | NS               |
|            | 0.207624         | 0.028601          | NS                 | NS               | NS               |
| C₂₄H₴₈O    | 1.58E-82         | 5.7E-83           | 3.50E-64           | 8.21E-65          |
|            | 1.67E-82         | 8.7E-83           | 1.64E-82           | 8.54E-83          |
|            | 2.24E-76         | 6.31E-77          | NS                 | NS               | NS               |
| C₂₅H₵₀O    | NS               | NS                | NS                 | NS               |
| C₂₆H₵₂O    | NS               | NS                | NS                 | NS               |
| C₂₇H₵₄O    | NS               | NS                | NS                 | NS               |
| C₂₈H₵₆O    | 1.75E-14         | 2.4E-15           | 5.06E-82           | 5.1E-82           |
|            | 2.12E-53         | 3.6E-54           | 7.32E-29           | 1.1E-29           |
|            | 3.83E-33         | 6.04E-34          | NS                 | NS               | NS               |
| C₂₉H₶₀O    | NS               | NS                | NS                 | NS               |
| C₃₀H₶₂O    | NS               | NS                | NS                 | NS               |

Case I: Cancer (156) vs. Control (193); Case II: Cancer (156) vs. Benign (65); Case III: Benign (65) vs. Control (193); Case IV: (Benign + Cancer) (221) vs. Control (193); Case V: (Control + Benign) vs. Cancer. Adj. P-val.: Adjusted p-values for multiple hypothesis correction; FDR: False Discovery Rate; VOC: Volatile Organic Compounds. NS: highly non-significant.

https://doi.org/10.1371/journal.pone.0277431.t001
### Table 2. Summary statistic(s) of the important VOCs for different patient populations.

| VOC       | Control (n = 193) | Lung Cancer (n = 156) |
|-----------|-------------------|-----------------------|
| C2H6O     | 1.400             | 3.273                 |
| C2H4O     | 0.058             | 1.031                 |
| C2H4O     | 0.014             | 0.111                 |
| C3H6O     | 0.052             | 0.076                 |
| C3H8O     | 0.671             | 1.033                 |
| C3H10O    | 0.112             | 0.111                 |
| C3H12O    | 0.112             | 0.112                 |
| C3H14O    | 0.046             | 0.046                 |
| C3H16O    | 0.087             | 0.087                 |
| C4H10O    | 0.186             | 0.186                 |
| C4H12O    | 0.034             | 0.034                 |
| C4H14O    | 0.003             | 0.003                 |
| C4H16O    | 0.014             | 0.014                 |
| C5H8O     | 0.052             | 0.052                 |
| C5H10O    | 0.119             | 0.119                 |
| C5H12O    | 0.013             | 0.013                 |

#### Benign Pulmonary Nodule (n = 65)

| VOC       | Mean    | SD      | Max    | CV      |
|-----------|---------|---------|--------|---------|
| C2H6O     | 1.941   | 1.454   | 8.210  | 0.44    |
| C2H4O     | 0.060   | 0.014   | 0.33   | 1.07    |
| C2H4O     | 0.037   | 0.008   | 0.028  | 0.019   |
| C3H6O     | 0.014   | 0.014   | 0.014  | 0.11    |
| C3H8O     | 0.062   | 0.069   | 0.069  | 0.027   |
| C3H10O    | 0.973   | 1.036   | 1.036  | 0.11    |
| C3H12O    | 0.955   | 1.007   | 1.007  | 0.11    |
| C3H14O    | 0.114   | 0.046   | 0.046  | 0.11    |
| C3H16O    | 0.222   | 0.025   | 0.025  | 0.09    |
| C4H10O    | 0.227   | 0.047   | 0.047  | 0.32    |
| C4H12O    | 0.059   | 0.058   | 0.058  | 0.11    |
| C4H14O    | 0.122   | 0.086   | 0.086  | 0.11    |
| C4H16O    | 0.227   | 0.141   | 0.141  | 0.32    |
| C5H8O     | 0.014   | 0.032   | 0.032  | 0.11    |
| C5H10O    | 0.059   | 0.058   | 0.058  | 0.11    |
| C5H12O    | 0.005   | 0.008   | 0.008  | 0.03    |
| 95% CI    | (1.314, 1.586) | (0.048, 0.168) | (0.004, 0.006) | (0.032, 0.074) |

---

SD: Standard deviation; Min: Minimum statistic; Max: Maximum statistic; CV: Co-efficient of variation; SE: Standard error; CI: Confidence Interval

https://doi.org/10.1371/journal.pone.0277431.t002
Table 3. Ranking of significant VOCs on different patients’ classification.

| Ranks | Case I | Ranks | Case II | Ranks | Case III | Ranks | Case IV | Ranks | Case V |
|-------|--------|-------|---------|-------|----------|-------|---------|-------|-------|
| 1     | C₆H₆O  | 1     | C₆H₆O  | 1     | C₆H₆O   | 1     | C₆H₆O  | 1     | C₆H₆O |
| 2     | C₆H₄O₂ | 2     | C₆H₄O₂ | 2     | C₆H₄O₂ | 2     | C₆H₄O₂ | 2     | C₆H₄O₂|
| 3     | C₆H₄O₂ | 3     | C₆H₄O₂ | 3     | C₆H₄O₂ | 3     | C₆H₄O₂ | 3     | C₆H₄O₂|
| 4     | C₆H₂O  | 4     | C₆H₂O  | 4     | C₆H₂O  | 4     | C₆H₂O  | 4     | C₆H₂O |
| 5     | C₆H₂O  | 5     | C₆H₂O  | 5     | C₆H₂O  | 5     | C₆H₂O  | 5     | C₆H₂O |
| 6     | C₆H₂O  | 6     | C₆H₂O  | 6     | C₆H₂O  | 6     | C₆H₂O  | 6     | C₆H₂O |
| 7     | C₆H₁₅O | 7     | C₆H₁₅O | 7     | C₆H₁₅O | 7     | C₆H₁₅O | 7     | C₆H₁₅O|
| 8     | C₆H₁₅O | 8     | C₆H₁₅O | 8     | C₆H₁₅O | 8     | C₆H₁₅O | 8     | C₆H₁₅O|
| 9     | C₆H₁₅O | 9     | C₆H₁₅O | 9     | C₆H₁₅O | 9     | C₆H₁₅O | 9     | C₆H₁₅O|
| 10    | C₆H₁₅O | 10    | C₆H₁₅O | 10    | C₆H₁₅O | 10    | C₆H₁₅O | 10    | C₆H₁₅O|
| 11    | C₆H₁₅O | 11    | C₆H₁₅O | 11    | C₆H₁₅O | 11    | C₆H₁₅O | 11    | C₆H₁₅O|
| 12    | C₆H₁₅O | 12    | C₆H₁₅O | 12    | C₆H₁₅O | 12    | C₆H₁₅O | 12    | C₆H₁₅O|
| 13    | C₆H₁₅O | 13    | C₆H₁₅O | 13    | C₆H₁₅O | 13    | C₆H₁₅O | 13    | C₆H₁₅O|
| 14    | C₆H₁₅O | 14    | C₆H₁₅O | 14    | C₆H₁₅O | 14    | C₆H₁₅O | 14    | C₆H₁₅O|
| 15    | C₆H₁₅O | 15    | C₆H₁₅O | 15    | C₆H₁₅O | 15    | C₆H₁₅O | 15    | C₆H₁₅O|

Case I: Cancer (156) vs. Control (193); Case II: Cancer (156) vs. Benign (65); Case III: Benign (65) vs. Control (193); Case IV: (Benign + Cancer) (221) vs. Control (193); Case V: (Control + Benign) vs. Cancer. Ranks of the VOCs based on the computed FDR values.

https://doi.org/10.1371/journal.pone.0277431.t003

Table 4. Classification metrics for different combinations of VOCs.

| #VOCs | Case I | Case II | Case III | Case IV | Case V |
|-------|--------|---------|----------|---------|-------|
|       | Mean.CA | SE.CA | Mean.CA | SE.CA | Mean.CA | SE.CA | Mean.CA | SE.CA | Mean.CA | SE.CA |
| 3     | 92.00   | 0.03   | 76.98   | 0.06   | 86.61   | 0.04   | 87.14   | 0.02   | 91.16   | 0.03   |
| 5     | 92.32   | 0.04   | 76.07   | 0.06   | 86.58   | 0.04   | 88.14   | 0.04   | 92.47   | 0.04   |
| 7     | 92.57   | 0.04   | 76.25   | 0.06   | 86.94   | 0.05   | 88.05   | 0.04   | 92.57   | 0.04   |
| 9     | 92.63   | 0.04   | 78.48   | 0.08   | 88.56   | 0.05   | 89.95   | 0.05   | 91.79   | 0.03   |
| 11    | 92.34   | 0.03   | 76.72   | 0.07   | 87.63   | 0.06   | 87.91   | 0.05   | 92.39   | 0.04   |
| 13    | 92.14   | 0.04   | 77.65   | 0.08   | 87.35   | 0.06   | 87.58   | 0.05   | 92.12   | 0.04   |
| 15    | 91.98   | 0.04   | 77.92   | 0.08   | 89.28   | 0.05   | 89.85   | 0.04   | 91.90   | 0.05   |

Case I: Cancer (156) vs. Control (193); Case II: Cancer (156) vs. Benign (65); Case III: Benign (65) vs. Control (193); Case IV: (Benign + Cancer) (221) vs. Control (193); Case V: (Control + Benign) vs. Cancer. Mean.CA: Mean Classification Accuracy; SE.CA: Standard Error in Classification Accuracy

https://doi.org/10.1371/journal.pone.0277431.t004
C₄H₈O and C₁₀H₂₀O, are clustered separately (Fig 4A). These two VOCs have less similarity with others across all the patients (Fig 4A). The remaining VOCs are clustered together, which indicates their similarity over the samples irrespective of the cancer classes (Fig 4A). The correlation among the VOCs over all the patients is shown in Fig 4B. The correlation analysis indicated that the VOCs, i.e., C₅H₈O, C₁₃H₂₆O, and C₇H₁₁O, are negatively correlated with other VOCs across all the samples (Fig 4B). Further, the remaining VOCs are somewhat positively correlated with others. This analysis indicated a similarity in the molecular concentrations of the VOCs across the samples/patients observed through FT-ICR-MS technology.

This study demonstrates that few detected markers in exhaled breath samples can be used as marker signatures for distinguishing patients with lung cancer from patients with benign pulmonary nodules and healthy subjects (Supplementary Document S5 in S1 File). Here, the VOC markers for cancer detection are identified through the statistically sound Boot-SVM-RFE technique. Through this technique, a statistically meaningful measure, i.e., adjusted p-value or FDR, was computed after correcting the multiple hypothesis testing problems. Then the adjusted p-value or FDR was assigned to each VOC, and the significant VOCs were selected based on these computed values. This measure is easily interpretable by experimental biologists and lab users, as the values are well defined in [0, 1]. In other words, a lower p-value indicates a more informative VOC and vice-versa. The random resampling procedure, i.e., bootstrap method, used in the Boot-SVM-RFE can eliminate the spurious and arbitrary association among the VOCs while detecting marker signatures [13, 16, 19]. Further, the Boot-SVM-RFE is more robust and does not require any distributional assumptions of the VOC data to obtain the distribution of test statistic(s). After selecting a few VOC markers, we trained the machine learning (i.e., SVM) based classification models to establish their relevance in lung cancer patients’ classification. Our study found that instead of using all the VOCs, one can focus on few (e.g., 3, 5, or 7) marker VOCs to detect lung cancer patients more robustly and accurately. This approach will be less time-consuming and require lesser resources to detect lung cancer based on exhaled breath samples using FT-ICR-MS technology. Here, we have narrowed down our search to a few VOCs instead of focusing on all the VOCs present in breath samples, which in turn will save the time and cost of the experiments.
The major advantages of our experimental approach include: first, the microreactors are designed with thousands of micropillars to provide higher capture rates of carbonyl compounds (e.g., VOCs) in breath samples [10–12]. Second, chemo-selective capture of carbonyl compounds through amino-oxy reactions simplifies the spectrum of compounds to be quantitated. Both the steps are well established in the literature and can be further used to detect other diseases, including COVID19 using exhaled breath samples. Third, a statistically efficient technique, Boot-SVM-RFE [19], was used to detect the markers using the VOC molecular concentration data. Fourth, in-silico validation of the VOC signatures through training machine learning-based classification models. In other words, our experimental approach includes VOC molecular data generation through microreactors based FT-ICR-MS technology and statistical analysis of the data using efficient statistical and machine learning techniques.

Conclusion

Early diagnosis of lung cancer is a key factor for increasing its survival rates among the patients. The analysis of carbonyl compounds present in exhaled breath of the patients is a promising non-invasive tool for the diagnosis of lung cancer at the early stage. In other words, the presence of metabolic carbonyl organic compounds in exhaled breath can play a vital role in the early detection of lung cancer patients, which will surely enhance the survival of the patients. Hence, the identification and characterization of the key metabolic VOCs using proper analytical approach and further using them in developing the classification models will play an important role in the quick and non-invasive detection of lung cancer. Therefore, in this study, we proposed an experimental approach to identify the key VOCs through the Boot-SVM-RFE technique and used these key VOCs to distinguish the patients with lung cancer from the benign pulmonary disease and healthy control classes.

Our analytical findings indicated that fewer VOCs can be used for lung cancer detection with sufficient classification accuracy. For instance, seven common key VOCs, including C₄H₈O₂, C₁₃H₂₂O, C₁₁H₂₂O, C₂H₄O₂, C₇H₁₄O, C₆H₁₂O, and C₅H₈O, can be successfully used for classification purposes under the five different settings. In this study, we used linear basis function in Boot-SVM-RFE technique, it will be interesting to study other non-linear basis functions or tree based models (e.g., random forests) to capture non-linear association among the VOCs. Further, our experimental and analytical approach of VOC quantitative analysis in breath samples may be extended to other diseases, including COVID19 detection. Besides, the analytical method used in this study can be applied to high-throughput gene expression studies, including RNA-sequencing and single-cell RNA-sequencing to select gene/bio-markers for the identification of cancer patients or cell types. Also, the reported experimental approach can be applied to other urine, saliva, and blood bio-assays based genetic studies to predict the phenotypes by identifying the organic compound based bio-signatures.

Supporting information

S1 File. Supplementary documents S1-S5. This file contains supporting documents from S1 to S5. Supplementary Document S1: SVM training for two class classification; Supplementary Document S2: Sliding Windows Size technique; Supplementary Document S3: List of VOC and their common names; Supplementary Document S4: Distribution of Key VOCs; Supplementary Document S5: Principal Component plots for visualizing patient classes.

(SDOCX)

S1 Table. Summary statistics of all the VOCs considered in this study.

(XLSX)
S2 Table. Statistical significance values computed through Boot-SVM-RFE for all the VOCs used in the study for all patient class combinations.

(XLSX)

Acknowledgments

In collaboration with Dr. M Bousamra, Dr. X Fu’s lab performed the bioassays. The support to Dr. Samarendra Das from ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, is duly acknowledged.

Author Contributions

Conceptualization: Shesh N. Rai, Samarendra Das.
Data curation: Samarendra Das, Jianmin Pan, Xiao-An Fu.
Formal analysis: Samarendra Das.
Funding acquisition: Shesh N. Rai, Xiao-An Fu.
Investigation: Samarendra Das, Xiao-An Fu.
Methodology: Samarendra Das.
Project administration: Shesh N. Rai, Xiao-An Fu.
Software: Samarendra Das.
Supervision: Shesh N. Rai, Jianmin Pan, Dwijesh C. Mishra, Xiao-An Fu.
Validation: Samarendra Das, Xiao-An Fu.
Visualization: Samarendra Das.
Writing – original draft: Samarendra Das.
Writing – review & editing: Shesh N. Rai, Samarendra Das, Jianmin Pan, Dwijesh C. Mishra, Xiao-An Fu.

References

1. American Cancer Society. Facts & Figures 2021. Atlanta, Ga.; 2021.
2. Chang JE: Lee DS, Ban SW, Oh J, Jung MY, Kim SH, et al. Analysis of volatile organic compounds in exhaled breath for lung cancer diagnosis using a sensor system. Sensors Actuators B Chem. 2018; 255: 800–807. https://doi.org/10.1016/J.SNB.2017.08.057
3. Patz EF: Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tamme à ¤gi MC, et al. Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. JAMA Intern Med. 2014; 174: 269–274. https://doi.org/10.1001/jamainternmed.2013.12736 PMID: 24322569
4. Li M: Yang D, Brock G, Knipp RJ, Bousamra M, Nantz MH, et al. Breath carbonyl compounds as biomarkers of lung cancer. Lung Cancer. 2015; 90: 92–97. https://doi.org/10.1016/j.lungcan.2015.07.005 PMID: 26233967
5. Pauling L: Robinson AB, Teranishi R, Cary P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proc Natl Acad Sci U S A. 1971; 68: 2374–2376. https://doi.org/10.1073/pnas.68.10.2374 PMID: 5289873
6. Phillips M: Herrera J, Krishnan S, Zain M, Greenberg J, Cataneo RN. Variation in volatile organic compounds in the breath of normal humans. J Chromatogr B Biomed Sci Appl. 1999; 729: 75–88. https://doi.org/10.1016/S0378-4347(99)00127-9 PMID: 10410929
7. Fu X-A: Li M, Knipp RJ, Nantz MH, Bousamra M. Noninvasive detection of lung cancer using exhaled breath. Cancer Med. 2014; 3: 174–181. https://doi.org/10.1002/cam4.162 PMID: 24402867
8. Schumer EM: Trivedi JR, van Berkel V, Black MC, Li M, Fu X-A, et al. High sensitivity for lung cancer detection using analysis of exhaled carbonyl compounds. J Thorac Cardiovasc Surg. 2015; 150: 1517–1524. https://doi.org/10.1016/j.jtcvs.2015.08.092 PMID: 26412316

9. Bousamra M: Schumer E, Li M, Knipp RJ, Nantz MH, van Berkel V, et al. Quantitative analysis of exhaled carbonyl compounds distinguishes benign from malignant pulmonary disease. J Thorac Cardiovasc Surg. 2014; 148: 1074–1081. https://doi.org/10.1016/j.jtcvs.2014.06.006 PMID: 25129599

10. Fu X-A: Li M, Biswas S, Nantz MH, Higashi RM. A novel microreactor approach for analysis of ketones and aldehydes in breath. Analyst. 2011; 136: 4662–4666. https://doi.org/10.1039/C1AN15618G PMID: 21897949

11. Li M: Biswas S, Nantz MH, Higashi RM, Fu X-A. Preconcentration and Analysis of Trace Volatile Carbonyl Compounds. Anal Chem. 2012; 84: 1288–1293. https://doi.org/10.1021/ac2021757 PMID: 22145792

12. Li M: Biswas S, Nantz MH, Higashi RM, Fu X-A. A microfabricated preconcentration device for breath analysis. Sensors Actuators B Chem. 2013; 180: 130–136. https://doi.org/10.1016/j.snb.2012.07.034

13. Das S: Rai A, Mishra DC, Rai SN. Statistical approach for selection of biologically informative genes. Gene. 2018; 655. https://doi.org/10.1016/j.gene.2018.02.044 PMID: 29458166

14. Duan KB: Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005. https://doi.org/10.1109/tnb.2005.853657 PMID: 16220686

15. Guyon I. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 1998. https://doi.org/10.10109/5254.708428

16. Das S: Rai SN. Statistical approach for biologically relevant gene selection from high-throughput gene expression data. Entropy. 2020; 22. https://doi.org/10.3390/e22111205 PMID: 33286973

17. Guyon I: Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002. https://doi.org/10.1023/A:1012487302797

18. Wang J: Chen L, Wang Y, Zhang J, Lian Y, Xu D. A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice. Xu Y, editor. PLoS One. 2013; 8: e64929. https://doi.org/10.1371/journal.pone.0064929 PMID: 23762267

19. Das S: Meher PK, Rai A, Bhar LM, Mandal BN. Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine max L.). Tian Z, editor. PLoS One. 2017; 12: e0169605. https://doi.org/10.1371/journal.pone.0169605 PMID: 28056073

20. Das S: McClain CJ, Rai SN. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. Entropy. 2020; 22: 427. https://doi.org/10.3390/e22040427 PMID: 33286201

21. Efron B: Tibshirani RJ. An Introduction to the Bootstrap. Boston, MA: Springer US; 1993. https://doi.org/10.1007/978-1-4899-4541-9

22. Benjamin Y: Hochberg Y. Multiple Hypotheses Testing with Weights. Scand J Stat. 1997; 24: 407–418. https://doi.org/10.1111/1467-9469.00072