UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

TOWARDS BREATH-BASED DISEASE DIAGNOSIS: A MACHINE LEARNING PIPELINE ON PTR-MS DERIVED VOC SIGNATURES

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

RUCHIKA BHARDWAJ

Norman, Oklahoma

2025

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/341612

TOWARDS BREATH-BASED DISEASE DIAGNOSIS: A MACHINE LEARNING PIPELINE ON PTR-MS DERIVED VOC SIGNATURES

A THESIS APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Thirumalai Venkatesan, Chair

Dr. Naoko Takebe

Dr. Binbin Weng

Dr. Yaser Michael Banad

© Copyright by RUCHIKA BHARDWAJ 2025
All Rights Reserved.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Thirumalai Venkatesan, for his invaluable guidance, encouragement, and support throughout my research journey. His insights, patience, and belief in my work have been instrumental in shaping this thesis.

I would also like to sincerely thank Dr. Velmurugan Thavasi, whose mentorship and collaboration through the Breathomics project provided me with both inspiration and direction. His guidance and thoughtful feedback have been deeply appreciated.

I gratefully acknowledge the Department of Electrical and Computer Engineering at the University of Oklahoma for providing the academic environment and resources necessary for this research.

I would also like to thank the Breathomics data collection team, whose diligent and meticulous efforts in gathering the dataset made this research possible. Your contribution laid the foundation for this work, and I am truly appreciative of your hard work and dedication.

I am forever grateful to my family, especially my parents and my husband, for their unwavering love, encouragement, and sacrifices that made it possible for me to pursue my goals.

Lastly, I would like to thank all the participants and collaborators whose contributions have directly or indirectly helped bring this work to fruition.

Table of Contents

Acknowledgments **iv**

List Of Tables **vii**

List Of Figures **viii**

Abstract **x**

1 Introduction **1**
1.1 Instruments for Spectrometry 2
1.1.1 PT-RMS origin 2
1.1.2 Importance of Breath analysis 3
1.2 Machine learning as a tool for disease detection 4

2 Theory and Background **6**
2.1 Introduction to PT-RMS 6
2.1.1 Working principle of PTR-MS 7
2.1.2 Applications of PT-RMS in Volatomics and Metabolomics 9
2.2 Analyzing Breath Chemistry with PTR-MS 11
2.2.1 Data collection and pre-processing 12
2.2.2 Steps for Pre-processing of data 12
2.3 Statistical methods 13
2.3.1 Log-normal distribution 14
2.3.2 Cohen-D biomarkers 15
2.4 Visualization 16
2.4.1 Principal Component Analysis (PCA) 16
2.4.2 t-distributed Stochastic Neighbor Embedding (t-SNE) 18
2.4.3 Uniform Manifold Approximation and Projection (UMAP) 19
2.5 Machine Learning Algorithms 20
2.5.1 Artificial Neural Network 21
2.5.2 K-Nearest Neighbor (KNN) 26
2.5.3 Random Forest (RF) 26
2.5.4 Support Vector Machine (SVM) 27

3 Results **29**
3.1 Visualization of Data for Classification: PCA, t-SNE, and UMAP . . . 30
3.2 Bio-markers from Statistical Analysis 32
3.2.1 Cohen's d Biomarkers 35
3.3 Classification with select biomarkers 38
3.4 Classification with all features 41

4 Summary and Conclusions **43**
4.1 Model Performance and Classification Results 43
4.2 Impact of Data Volume and Future Scalability 44
4.3 Proposal for On-site Implementation 44

Reference List **46**

List Of Tables

3.1 Top 20 VOCs Differentiating Lung Cancer from Healthy 35
3.2 Top 20 VOCs Differentiating COVID-19 from Healthy (Based on Cohen's d values) 37
3.3 Summary of classification performance for each model. 42

List Of Figures

2.1 Schematic diagram of a PTR-MS system showing the three main operational regions: the reaction region, where VOCs are ionized via proton transfer from H_3O^+ ions in a drift tube; the transfer region, where ionized molecules are guided through a quadrupole ion guide; and the detection region, where the ions are analyzed using a Time-of-Flight Mass Spectrometer (TOF-MS) to determine their mass-to-charge ratios. Taken from Hegen et al. (2023) 8

2.2 Illustration of a quadrupole mass analyzer, where ionized molecules from the ion source are filtered based on their mass-to-charge ratio by an oscillating electric field applied to four parallel rods. Only ions with stable trajectories corresponding to a selected m/z reach the ion detector. 9

2.3 Volatile Organic Compounds (VOCs) reaching the alveoli through bloodstream as a by products of metabolic process in cells. Images taken from Google. 10

2.4 Workflow for PTR-TOF-MS data processing, including calibration, peak identification, and multivariate analysis and other instrument components. 11

2.5 Distribution of m015 VOC among various samples fitted by log-normal distribution (red-line). 15

2.6 Classification of two classes: dog and cat using neural networks. 21

2.7 (a) Schematic representation of ANN with an input layer consisting of N input neurons $\mathbf{v}$, L hidden layers consisting of M_l hidden neurons $\mathbf{h}^{(l)}$ each and an output layer with K output neurons $\mathbf{y}$. Information is processed from left to right, as indicated by the arrows, where each neuron gets as input $a_j^{(l)}$ the value of each neuron $h_i^{(l-1)}$ of the previous layer multiplied by a weight $W_{i,j}^{(l)}$ including a bias $b_j^{(l)}$. (b) Examples of activation functions $\phi(a)$ as a function of neuron input a for (i) sigmoid represented by blue dashed (Eq. 2.13), (ii) rectified linear unit (ReLU) denoted by solid green (Eq. 2.14), and (iii) tanh shown with red dot-dashed (Eq. 2.15). 22

2.8 Schematic diagram of our multi-stage cascading (MSC) model, where on the left, we show the preliminary stages of the model including data acquisition and data preprocessing, while on the right, we show the core of the cascading model where the ML algorithms detect diseases at each stage and follow onto the next stage based on the output at the previous stage. 27

3.1 shows the 2D projections of the dataset using PCA (left), t-SNE (middle), and UMAP (right). Each point represents a subject, and the colors indicate the respective class: green for Healthy, red for Lung Cancer, and purple for COVID-19. 31
3.2 Distribution of VOCs across patients with lung cancer (legend markers). 33
3.3 ROC curve and confusion matrix for COVID-19 classification using both Random Forest (RF) and K-Nearest Neighbors (KNN) on the test set. The left panel shows the ROC curve with an AUC of 1.00, reflecting perfect classification performance. The right panel presents the confusion matrix where both RF and KNN correctly classified all COVID-19 and Healthy samples, achieving 100% sensitivity and 100% specificity. 39
3.4 ROC curves and confusion matrices for Lung Cancer classification using Random Forest (RF) and K-Nearest Neighbors (KNN). Both models produced identical confusion matrices, achieving 90% sensitivity and 100% specificity. However, RF achieved a slightly higher AUC of 0.96 compared to 0.95 for KNN. 40
3.5 Confusion matrix indicating results from the different machine learning algorithms with (a) Artificial Neural network (ANN), (b) K-Nearest neighbor (KNN), (c) Random Forest (RF), and (d) Support vector machine (SVM). 41

Abstract

Diseases such as lung cancer are often difficult to diagnose in their early stages, and by the time detection occurs, the disease may have already progressed to an aggressive phase. Noninvasive methods like exhaled breath analysis offer a promising alternative that can facilitate early detection in a painless and accessible manner. Even established techniques such as Reverse Transcription Polymerase Chain Reaction (RT-PCR), commonly used for COVID-19 detection, fall short in delivering comprehensive diagnostic coverage.

In this study, we leverage Proton Transfer Reaction - Time of Flight - Mass Spectrometry (PTR-TOF-MS) to analyze exhaled breath samples, enabling high-resolution detection of volatile organic compounds (VOCs). By examining the concentration profiles of these VOCs, we aim to identify disease-specific molecular signatures, or "fingerprints," that distinguish between health conditions. Each disease produces a unique VOC signature, making real-time, noninvasive diagnosis possible.

A key contribution of this work is the development of a cascading classification model tailored for real-time disease prediction. Despite working with a relatively limited dataset, the model achieves classification accuracies of up to 96%, highlighting the potential of breath-based diagnostics.

In a separate approach, we employed Cohen's dd effect size method to identify statistically significant VOCs that differentiate disease groups. This statistical technique serves as a powerful tool for biomarker discovery, enabling more interpretable models and guiding future research on disease-specific breath signatures.

These results strongly suggest that expanding the dataset could further enhance model performance and generalizability, paving the way for scalable and rapid clinical screening tools.

Chapter 1

Introduction

The development of non-invasive diagnostic methods has gained significant attention in recent years, with breath analysis emerging as a promising approach for detecting various diseases [Amor et al. (2019)]. We dive into the understanding of the human breath which contains a wide array of volatile organic compounds (VOCs), which forms varying patterns due to diseases, providing meaningful insights into individual's health status. Previous research has highlighted the potential of breath analysis for health monitoring [Ligor et al. (2015)], yet a direct and consistent link between specific VOCs and distinct diseases still remains elusive.

Recent studies have made strides in associating certain VOCs with specific conditions, such as lung cancer, across different age groups and environmental factors [Jia et al. (2024)]. However, these connections are often complex and require advanced techniques to fully decipher. With the advent of machine learning algorithms being used extensively across disciplines, including breath analysis [Cusack et al. (2024)], we aim to further leverage these capabilities of machine learning algorithms to analyze the VOC patterns in human breath and identify bio-markers associated with various diseases simultaneously. By applying these advanced computational tools, we seek to enhance the diagnostic potential of breath analysis.

Our proposed method not only offers a non-invasive and patient-friendly approach but also provides rapid results with an accuracy rate of up to 96%, making it an ideal

tool for early detection and screening. This study demonstrates how combining machine learning with breath analysis can offer a fast, reliable, and scalable method for diagnosing diseases at their initial stages, potentially transforming preventive healthcare.

1.1 Instruments for Spectrometry

Let us begin with how we can analyze the volatile organic compounds (VOCs) present in the exhaled breath which are critical in fields such as medical diagnostics, atmospheric chemistry and food quality assessment. Conventional methods, like Gas Chromatography-Mass Spectrometry (GC-MS), are highly accurate for molecular identification but lack the speed and adaptability required for real-time applications. Proton Transfer Reaction Mass Spectrometry (PTR-MS), introduced in the late 1990s, addresses these gaps with its ability to perform high-sensitivity, real-time measurements of VOCs.

Recent developments and improvements in PTR-TOF-MS (Time-of-Flight) have increased the mass resolution and detection limitations of PTR-MS, hence broadening its use. But these developments also bring with them new difficulties with data processing and instrument complexity. Our study provides a critical overview of these advancements with an emphasis on how they affect industry and research.

1.1.1 PT-RMS origin

With the emergence of Proton Transfer Reaction Mass Spectrometry (PTR-MS) in the late 1990s, a novel technique was designed to overcome the limitations of conventional mass spectrometry for volatile organic compound (VOC) detection, particularly in real-time and trace-level analysis. Developed by Lindinger, Hansel, and Jordan at

the University of Innsbruck [Lindinger et al., 1998]. PTR-MS works on the principle of chemical ionization, which is a soft ionization technique, wherein hydronium ions (H_3O^+) serve as reagent ions to protonate VOC molecules with higher proton affinity than water. As a result the molecular ion is preserved, fragmentation is reduced, and mass spectra can be interpreted more easily thanks to this soft ionization approach.

1.1.2 Importance of Breath analysis

From the pioneering research that led to the identification of over 250 volatile substances in human breath by Linus Pauling and colleagues in 1971 [Pauling et al., 1971], to the more comprehensive study by Phillips et al. in 1999 that revealed over 3,000 VOCs in exhaled breath [Phillips et al., 1999], the importance of breath as a medium for disease detection has grown into a substantial and multidisciplinary field of study. These landmark studies established the presence of a chemically diverse and information-rich medium that reflects metabolic and pathophysiological states, thereby laying the groundwork for breath-based diagnostics.

Breath VOCs are endogenous compounds that originate from normal or disease-related metabolic processes within cells. These compounds enter the bloodstream, diffuse into the alveoli, and are eventually exhaled through the lungs. Their concentrations can vary in response to physiological stress, inflammation, or disease progression. Past studies have demonstrated the potential of breath analysis in detecting diseases such as lung cancer [Phillips et al., 2003], chronic obstructive pulmonary disease (COPD) [Dragonieri et al., 2007], diabetes [Gupta et al., 2010], and more recently, COVID-19 [Steppert et al., 2021]. This non-invasive and painless nature of breath sampling makes it ideal for early screening, continuous monitoring, and longitudinal studies.

Breathomics has advanced significantly with the advent of Proton Transfer Reaction Mass Spectrometry (PTR-MS) and its high-resolution variant, PTR-TOF-MS.

These instruments enable real-time, high-sensitivity detection of trace VOCs without the need for sample pre-concentration or chromatographic separation. However, breathomics still faces notable challenges, including high inter-subject variability, ambient air contamination, and the lack of standardized protocols for sample collection, data processing, and VOC identification. Despite these limitations, breath analysis is steadily evolving into a practical tool for clinical diagnostics, supported by growing evidence and technological refinement.

1.2 Machine learning as a tool for disease detection

The field of breath analysis has witnessed a substantial increase in the volume and dimensionality of data generation with the advancement of analytical technologies such as PTR-MS. There can be hundreds to thousands of volatile organic compounds (VOCs) in each breath sample, making manual interpretation and traditional statistical approaches inadequate for extracting clinically meaningful patterns. This is where machine learning (ML) has emerged as an indispensable tool, capable of navigating the complexity of high-dimensional data to uncovering subtle, non-linear relationships that may be indicative of disease states.

ML techniques have been increasingly adopted in recent years for various biomedical applications, including cancer detection, infectious disease screening, and metabolic disorder classification. By learning from labeled datasets and employing supervised algorithms such as support vector machines (SVM), random forests (RF), k-nearest neighbors (KNN), and deep neural networks, it is possible to build predictive models that distinguish between healthy and diseased individuals based on their breath VOC profile. Unsupervised techniques such as principal component analysis (PCA) and

other clustering algorithms can also play a vital role in visualizing patterns, identifying outliers, and reducing dimensionality before classification.

The integration of PTR-MS with advanced machine learning techniques provides a platform for rapid transformative evolution of non-invasive diagnostics. This combined approach enables the rapid processing and interpretation of complex VOC datasets, facilitating accurate and real-time detection of diseases directly from the exhaled breath. As both analytical instrumentation and computational models continue to improve, this synergy holds immense potential for scalable, point-of-care diagnostics across a wide spectrum of medical conditions. By harnessing the strengths of both domains, we move closer to realizing a future where routine, non-invasive breath tests become a practical tool in early disease detection and public health monitoring.

Chapter 2

Theory and Background

Our study is structured around three core components integral to the process of breath-based disease detection. First, we utilize a Proton Transfer Reaction Mass Spectrometry (PTR-MS) system, which captures exhaled breath and converts its chemical constituents i.e. volatile organic compounds (VOCs) into measurable electrical signals. Second, we analyze the concentration profiles of these VOCs to quantify their presence and variability across samples. Third, we apply machine learning algorithms to these concentration data to classify and identify disease states with predictive accuracy. In the following sections, we provide the theoretical background for each of these components to establish a foundation for interpreting our results, which are presented in the subsequent chapter.

2.1 Introduction to PT-RMS

Proton Transfer Reaction Mass Spectrometry (PTR-MS) is a highly sensitive and real-time analytical technique widely used for detecting and quantifying trace levels of volatile organic compounds (VOCs) in complex gaseous mixtures, including human breath. Introduced in the mid-1990s by [Hansel et al. (1995)], PTR-MS operates on the principle of chemical ionization, where H_3O^+ ions serve as the primary reagent to selectively ionize VOCs with higher proton affinities than water, while leaving major

constituents like N_2, O_2, and CO_2 unreacted. This selective ionization minimizes fragmentation and allows for a more accurate representation of the molecular masses of VOCs present in the sample. Early developments by [Lindinger et al. (1998)] established the method's ability to detect trace levels of VOCs in the pptv range, outperforming traditional techniques like GC-MS in speed and sensitivity. The introduction of PTR-TOF-MS marked a significant leap, achieving mass resolutions up to 6000 m/Δm. This improvement addressed issues of isobaric compound differentiation, a common limitation in earlier PTR-MS systems.

Owing to its high sensitivity (down to pptv levels), fast response time, and minimal sample preparation, PTR-MS has become a powerful tool in environmental monitoring, food science, and medical diagnostics, particularly for non-invasive breath analysis in disease detection [Lindinger et al. (1998); Amann et al. (2014)]. In the context of biomedical applications, PTR-MS enables real-time profiling of VOCs that can serve as potential biomarkers for various diseases, laying the groundwork for machine learning-based classification and diagnosis.

2.1.1 Working principle of PTR-MS

Proton Transfer Reaction Mass Spectrometry (PTR-MS) operates through a multi-stage process that enables the sensitive and selective detection of volatile organic compounds (VOCs) in gaseous samples, such as exhaled human breath. As illustrated in Figure 2.1, the system is divided into three main regions: the reaction region, transfer region, and detection region. In the reaction region, hydronium ions (H_3O^+) are generated from water vapor in the ion source and drift along an electric field through a drift tube. When a sample is introduced through the sample inlet, VOCs with proton affinities higher than that of water undergo soft ionization via proton transfer from H_3O^+, resulting in protonated VOC ions with minimal fragmentation. These ions are

then guided into the transfer region via a quadrupole ion guide, which focuses and transmits them into the detection unit, typically a Time-of-Flight (TOF) mass spectrometer. The TOF analyzer measures the mass-to-charge ratio (m/z) of the ions based on their flight time, enabling rapid and accurate quantification of VOCs [Hansel et al. (1995); Lindinger et al. (1998)].

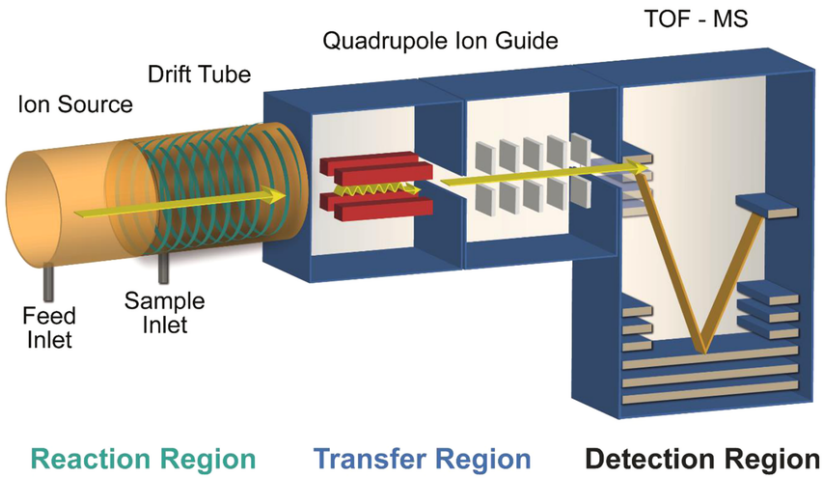


Figure 2.1: Schematic diagram of a PTR-MS system showing the three main operational regions: the reaction region, where VOCs are ionized via proton transfer from H_3O^+ ions in a drift tube; the transfer region, where ionized molecules are guided through a quadrupole ion guide; and the detection region, where the ions are analyzed using a Time-of-Flight Mass Spectrometer (TOF-MS) to determine their mass-to-charge ratios. Taken from Hegen et al. (2023)

In detail, Figure 2.2 provides a schematic of the quadrupole mass filter, which alone can also serve as an alternative to TOF in some PTR-MS systems. After ionization, ions are directed into the quadrupole, which consists of four parallel rods with alternating RF (radio frequency) and DC voltages. This configuration creates a dynamic electric field that selectively stabilizes the trajectory of ions with a specific m/z, allowing only those ions to reach the detector. Ions with unstable trajectories are filtered out. The quadrupole mass analyzer thus acts as a selective gate, scanning through a range of m/z values by varying the applied voltages, and producing a mass spectrum of the sample. While TOF offers higher resolution and faster acquisition, quadrupoles are

more compact and suitable for targeted analyses. Both configurations, however, serve to convert breath-borne molecular information into digital signals that can be analyzed for diagnostic purposes [Amann et al. (2014)].

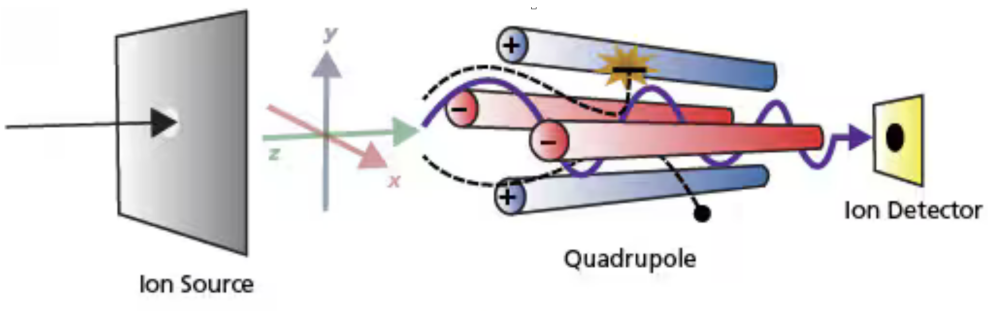


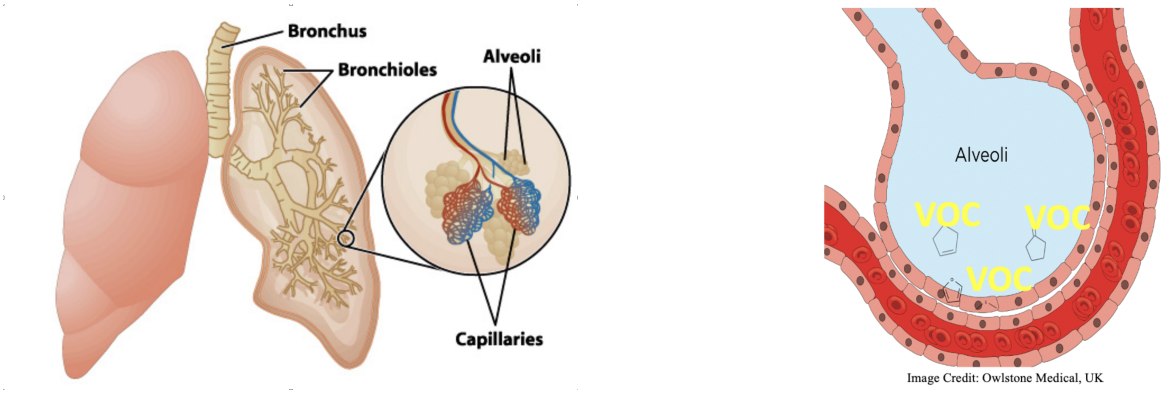
Figure 2.2: Illustration of a quadrupole mass analyzer, where ionized molecules from the ion source are filtered based on their mass-to-charge ratio by an oscillating electric field applied to four parallel rods. Only ions with stable trajectories corresponding to a selected m/z reach the ion detector.

PTR-TOF-MS with its time-of-flight mass analyzer, enhances sensitivity and mass range. Jordan et al. (2009) reported detection limits as low as pptv with significantly improved resolution. These advancements allow for the differentiation of isobaric species [Graus et al. (2010)], a major limitation of earlier quadrupole-based PTR-MS systems. While the increased complexity of PTR-TOF-MS instrumentation may pose challenges for routine use in non-specialized laboratories, its superior analytical performance makes it an invaluable tool for high-resolution VOC analysis in advanced research and clinical applications.

2.1.2 Applications of PT-RMS in Volatomics and Metabolomics

Volatomics and metabolomics are complementary fields aimed at understanding the molecular composition of biological samples. Volatomics specifically focuses on the

comprehensive analysis of volatile organic compounds (VOCs) emitted from biological systems, such as exhaled breath, as shown in Figure 2.3, while metabolomics involves the broader study of small-molecule metabolites within cells, tissues, or biofluids. Proton Transfer Reaction – Time-of-Flight Mass Spectrometry (PTR-TOF-MS) has emerged as a valuable tool in both domains due to its ability to detect a wide range of VOCs in real time with high sensitivity and resolution. In volatomics research, PTR-TOF-MS enables non-invasive monitoring of metabolic processes by identifying disease-related VOC patterns. As demonstrated in the work of Roquencourt et al. (2022), advanced signal processing and statistical techniques applied to PTR-TOF-MS data can enhance biomarker discovery by addressing challenges such as high dimensionality, overlapping peaks, and variability in breath composition. The integration of PTR-TOF-MS with robust analytical pipelines paves the way for personalized diagnostics and early disease detection based on breath metabolite signatures.



(a) (b)

Figure 2.3: Volatile Organic Compounds (VOCs) reaching the alveoli through bloodstream as a by products of metabolic process in cells. Images taken from Google.

Cappellin et al. (2011) emphasized the need for robust data processing frameworks for PTR-TOF-MS. The high-resolution datasets generated by PTR-TOF-MS require

advanced computational tools for peak identification, baseline correction, and multivariate analysis. Figure 2.4 demonstrates the typical data workflow, from raw spectra to data mining.



Figure 2.4: Workflow for PTR-TOF-MS data processing, including calibration, peak identification, and multivariate analysis and other instrument components.

2.2 Analyzing Breath Chemistry with PTR-MS

Breath contains a rich array of volatile organic compounds (VOCs) that reflect ongoing physiological and metabolic processes within the body. These chemical signatures are often present at trace levels, offering a non-invasive window into health and disease states of an individual. Analyzing this breath chemistry, has emerged as a promising approach in clinical diagnostics, environmental exposure assessment, and metabolic monitoring. Proton Transfer Reaction Mass Spectrometry (PTR-MS) is particularly well-suited for this task, as it provides real-time, sensitive, and selective detection of VOCs without the need for extensive sample preparation. This section explores how PTR-MS enables precise characterization of breath constituents, laying the foundation for identifying disease-related biomarkers.

2.2.1 Data collection and pre-processing

For this study, exhaled breath samples were analyzed using a Proton Transfer Reaction-Time of Flight-Mass Spectrometry (PTR-TOF-MS) system. Data collection was carried out at two primary locations: the University of Oklahoma (OU) Norman campus and the OU Health Sciences Center. The initial dataset included breath profiles from 190 Healthy individuals, 5 COVID-19 patients, 6 Respiratory Syncytial Virus (RSV) patients, and 5 Influenza (Flu) patients.

In addition to the locally collected data, we also incorporated breath samples acquired through a collaborative partnership with Breathonix, Singapore. This extended dataset included additional samples from Healthy individuals, COVID-19 patients, and Lung Cancer patients.

All datasets were merged and subjected to a unified pre-processing pipeline to ensure consistency across collection sources. Artifact and extraneous data removal procedures were applied to eliminate noise and instrument-related variability. Due to the limited number of RSV and Flu samples, these classes were excluded from the machine learning pipeline, as they would not provide sufficient statistical power for training or validation. The final dataset used for classification consisted of 190 Healthy samples, 55 COVID-19 samples, and 50 Lung Cancer samples.

2.2.2 Steps for Pre-processing of data

The following steps were undertaken to prepare the raw PTR-MS data for analysis and classification:

1. **Background Signal and Redundancy Removal:** After acquiring the raw data, background signals and molecular columns that contained no measurable

values or were deemed biologically irrelevant were removed. In addition, redundant instrument measurements and metadata not contributing to the analysis were excluded to reduce dimensionality and noise.

2. **Outlier Detection:** The dataset was then scanned for outliers, which could potentially distort statistical analysis and bias model training. These were identified using a combination of visual inspection (e.g., histograms and boxplots) and statistical thresholds (e.g., values beyond the 95% confidence interval) and subsequently removed.

3. **Data Standardization:** Finally, the data was standardized using z-score normalization, transforming each feature to have zero mean and unit variance. This step ensures uniform scaling across all VOCs and prevents features with larger numerical ranges from dominating the learning process.

2.3 Statistical methods

Statistical methods play a critical role in identifying potential biomarkers by providing rigorous, quantitative frameworks to detect meaningful differences in the biological data. In the context of breathomics and disease classification, statistical tools help distinguish between random variability and biologically relevant changes in VOC (volatile organic compound) concentrations across groups such as healthy individuals, COVID-19 patients, and lung cancer patients. Techniques such as hypothesis testing, effect size measurements (e.g., Cohen's d), and multivariate analysis enable researchers to evaluate the significance, magnitude, and consistency of observed differences. Unlike purely visual inspection or threshold-based filtering, statistical approaches offer reproducibility, scalability, and control over false discovery rates, which are the key elements when working with high-dimensional, noisy biological datasets. By highlighting the

most discriminative features, these methods form the foundation for selecting robust candidate biomarkers and improving the performance of downstream machine learning classifiers.

2.3.1 Log-normal distribution

Log-normal distribution plots are instrumental in revealing the underlying differences between classes for each volatile organic compound (VOC). In Figure 2.5, we show the distribution for a VOC, which allows us to visualize the variability and skewness inherent in VOC measurements, helping to distinguish potential biomarkers from irrelevant compounds, as investigated in more detail in Jia et al. (2024). In the context of lung cancer detection, certain VOCs display distinguishable patterns between patients with and without lung cancer, which we will discuss in more detail in section 3.2. However, due to the complex and highly variable nature of VOC distributions across classes, consistently identifying robust biomarkers remains a significant challenge.

To address this, we employ statistical techniques to quantify the separation between class distributions. Specifically, we use Cohen's d, a standardized effect size measure, to capture the mean difference between the two groups relative to their pooled standard deviation. VOCs with high Cohen's d values represent features with the most pronounced differences between classes and are strong candidates for biomarker selection.

These selected VOCs are subsequently used as input features for machine learning models, enabling more focused and reliable predictions of disease presence. By combining statistical filtering with machine learning, we aim to enhance the interpretability and predictive accuracy of the classification process.

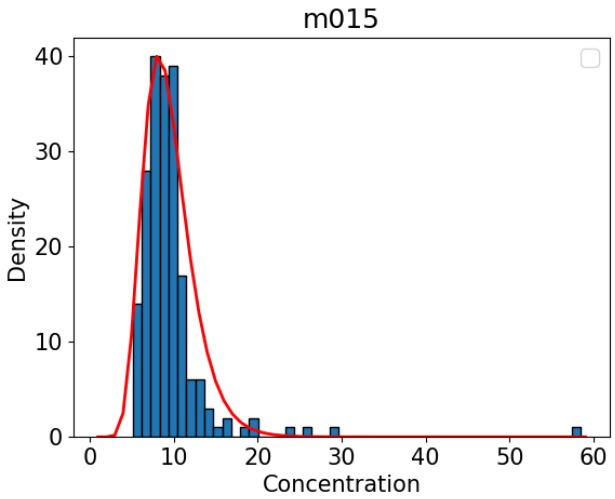


Figure 2.5: Distribution of m015 VOC among various samples fitted by log-normal distribution (red-line).

2.3.2 Cohen-D biomarkers

Cohen's d is a widely used statistical measure for quantifying the effect size between two groups by expressing the difference in their means relative to the pooled standard deviation [Cohen (2013)]. In the context of VOC analysis for biomarker discovery, Cohen's d provides a standardized and objective method to assess which compounds exhibit the most significant differences in concentration between diseased and healthy groups. This is particularly valuable when dealing with breath data, where VOC concentrations often follow a log-normal distribution, making visual or manual separation unreliable and prone to bias. While fitting log-normal distributions can highlight skewness and variability, it may overlook subtle but statistically meaningful shifts between groups. In contrast, Cohen's d captures both the magnitude and consistency of these shifts, allowing for more robust and scalable identification of potential biomarkers, especially in high-dimensional datasets common to metabolomic and volatomics studies [Broadhurst and Kell (2006)].

2.4 Visualization

In high-dimensional datasets such as those generated from VOC profiles in breathomics, visualization techniques are essential for uncovering hidden patterns and guiding classification tasks. When analyzing complex data across multiple disease states, such as healthy individuals, COVID-19 patients, and lung cancer patients, it becomes crucial to explore how samples cluster or separate in feature space. Dimensionality reduction techniques like Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and t-distributed Stochastic Neighbor Embedding (t-SNE) allow researchers to project high-dimensional data into two or three dimensions while preserving relevant structural information. These projections offer intuitive visual insights into potential group separations, identify outliers, and can guide downstream classification models by confirming the presence of meaningful biological variation.

2.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that transforms a dataset into a set of orthogonal components (principal components) that capture the maximum variance in the data. By projecting the original high-dimensional data onto the top few components, PCA reveals dominant trends and patterns, which can help in visually distinguishing between disease groups. In breathomics, PCA serves as a useful first-pass tool to assess whether VOC concentration differences between healthy, COVID-19, and lung cancer subjects manifest as distinct clusters, offering an interpretable way to explore data structure.

A basic way to reduce feature dimension is by mapping data into a lower dimensional feature space through linear projection. PCA finds projection directions that maximize data variance.

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be a set of instances, $w \in \mathbb{R}^p$ be a projection vector and $\tilde{x}_1 = w^T x_1, \ldots, \tilde{x}_n = w^T x_n$ be the projected instances. Variance of the projected instances is estimated as

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - \tilde{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - w^T \mu)^2, \tag{2.1}$$

where $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i$ is the projected sample mean and $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ is input sample mean.

PCA finds w by solving

$$\max_{w} J(w) \quad \text{s.t. } \|w\|^2 = 1. \tag{2.2}$$

The constraint helps to avoid trivial solution.

The above problem can be solved using the Lagrange multiplier technique, which shows

$$\Sigma w = \lambda w, \tag{2.3}$$

where $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$ is the sample covariance matrix. This suggests the optimal w is the eigenvector of Σ associated with the largest eigenvalue.

The second projection vector w_2 can be found in a similar way with an additional constraint $w^T w_2 = 0$. It is the eigenvector of Σ associated with the second largest eigenvalue. Practically, we can find k projection vectors by performing eigenvalue decomposition on Σ and retrieve the top k eigenvectors.

2.4.2 t-distributed Stochastic Neighbor Embedding (t-SNE)

t-distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction method designed to preserve local neighborhood relationships in high-dimensional data. It maps complex data into a lower-dimensional space in such a way that points that are close in the high-dimensional space remain close in the visualization. This makes t-SNE especially useful for identifying subtle, non-linear class boundaries and substructures that might be invisible to linear methods like PCA. For VOC data, t-SNE can reveal tightly grouped patient clusters and assist in visually validating disease-specific biomarker profiles.

t-SNE is a nonlinear dimensionality reduction technique that aims to preserve the local structure of high-dimensional data by modeling pairwise similarities.

Given a set of n high-dimensional data points $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$, t-SNE constructs a probability distribution P_{ij} over pairs of points in the high-dimensional space such that similar points have high probability:

$$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}, \quad P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n} \tag{2.4}$$

Here, σ_i is the bandwidth of a Gaussian centered at x_i, which is determined via binary search to match a predefined perplexity.

For low-dimensional embeddings $y_1, \ldots, y_n \in \mathbb{R}^d$ (typically $d = 2$ or 3), t-SNE defines a Student-t distribution with one degree of freedom (i.e., a Cauchy distribution) to measure similarities:

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{2.5}$$

The objective is to minimize the Kullback-Leibler divergence between the two distributions:

$$\mathcal{L}_{\text{t-SNE}} = \sum_{i \neq j} P_{ij} \log \left(\frac{P_{ij}}{Q_{ij}} \right) \tag{2.6}$$

This cost function is minimized using gradient descent, typically with momentum-based updates and early exaggeration phases to improve convergence.

2.4.3 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a more recent non-linear visualization technique that, like t-SNE, excels at capturing local relationships but also better preserves global structure in the data. It is built on manifold learning theory and is computationally more efficient than t-SNE, making it suitable for large and high-dimensional VOC datasets. UMAP is particularly effective in showing how different disease states like healthy, COVID-19, and lung cancer may form distinct but related manifolds, facilitating a clearer understanding of class separability and transition zones between conditions.

UMAP is a more recent dimensionality reduction technique grounded in manifold learning and topological data analysis. It builds a fuzzy topological representation of the data, then optimizes a low-dimensional embedding to preserve this structure.

Given a dataset $X = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^p$, UMAP starts by constructing a weighted k-nearest neighbor graph. For each data point x_i, it defines a local connectivity measure:

$$\mu_{ij} = \exp\left(-\frac{\|x_i - x_j\| - \rho_i}{\sigma_i} \right), \quad \text{for } x_j \in \text{kNN}(x_i) \tag{2.7}$$

Here, ρ_i is the distance to the nearest neighbor (to ensure local connectivity), and σ_i is chosen to normalize the number of neighbors (via a similar perplexity-like criterion). These local connectivities are symmetrized to form the fuzzy union:

$$P_{ij} = \mu_{ij} + \mu_{ji} - \mu_{ij} \cdot \mu_{ji} \tag{2.8}$$

For low-dimensional embeddings $y_1, \ldots, y_n \in \mathbb{R}^d$, UMAP uses a differentiable approximation to a cross-entropy loss between fuzzy sets:

$$Q_{ij} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1} \tag{2.9}$$

The embedding is optimized by minimizing the cross-entropy:

$$\mathcal{L}_{\text{UMAP}} = \sum_{i<j} \left[P_{ij} \log(Q_{ij}) + (1 - P_{ij}) \log(1 - Q_{ij})\right] \tag{2.10}$$

Here, a and b are hyperparameters chosen to control the shape of the embedding space (typically determined empirically).

2.5 Machine Learning Algorithms

Opportunities lie in integrating PTR-MS with machine learning and artificial intelligence for automated data interpretation. This approach could enhance the technology's utility in real-time monitoring and high-throughput environments.

We employ a diverse set of machine learning algorithms to the data. Specifically, we employ Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) classifiers. Each of these algorithms has distinct advantages in pattern recognition and classification tasks, making them

well-suited for analyzing complex biological data like VOC concentrations. Let us understand the working mechanism of each of the algorithms:

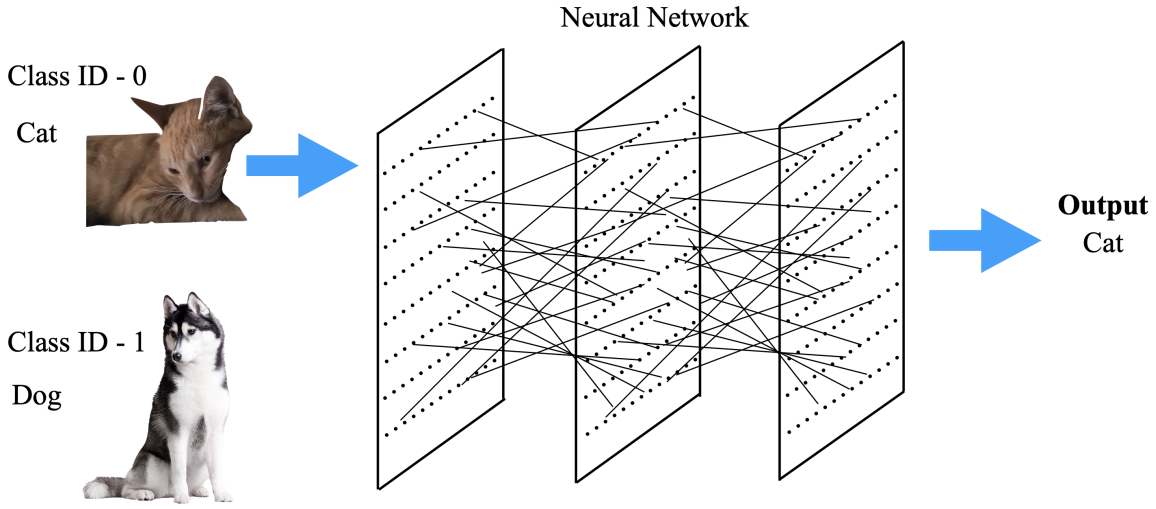


Figure 2.6: Classification of two classes: dog and cat using neural networks.

2.5.1 Artificial Neural Network

Artificial Neural Networks are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons) that transform input features through weighted connections and non-linear activation functions. During training, the network learns to minimize the prediction error by adjusting weights using algorithms like backpropagation. ANNs are capable of modeling complex non-linear relationships and are particularly useful in high-dimensional datasets like VOC profiles from breath. Their flexibility and scalability make them a powerful tool in classification and regression tasks [LeCun et al. (2015)].

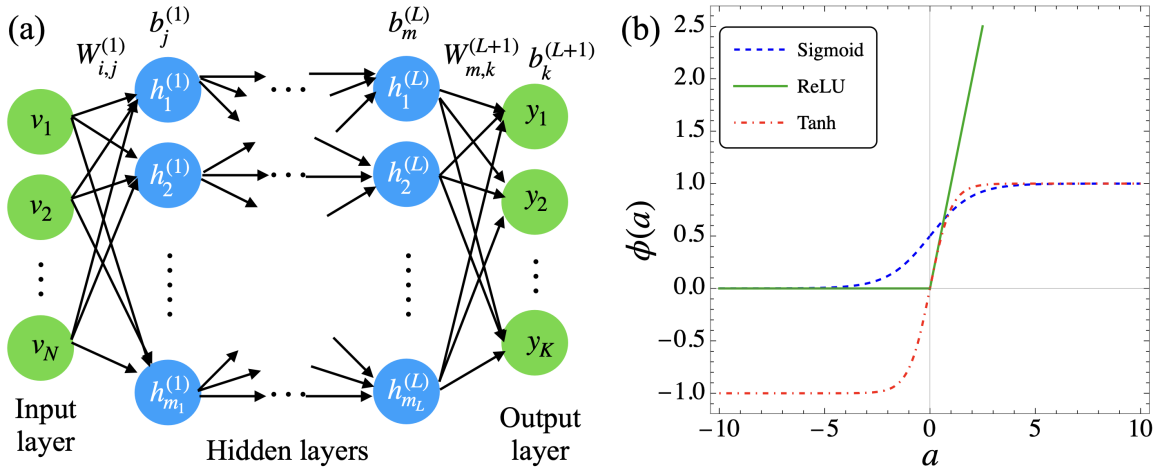


Figure 2.7: (a) Schematic representation of ANN with an input layer consisting of N input neurons $\mathbf{v}$, L hidden layers consisting of M_l hidden neurons $\mathbf{h}^{(l)}$ each and an output layer with K output neurons $\mathbf{y}$. Information is processed from left to right, as indicated by the arrows, where each neuron gets as input $a_j^{(l)}$ the value of each neuron $h_i^{(l-1)}$ of the previous layer multiplied by a weight $W_{i,j}^{(l)}$ including a bias $b_j^{(l)}$. (b) Examples of activation functions $\phi(a)$ as a function of neuron input a for (i) sigmoid represented by blue dashed (Eq. 2.13), (ii) rectified linear unit (ReLU) denoted by solid green (Eq. 2.14), and (iii) tanh shown with red dot-dashed (Eq. 2.15).

Artificial neural networks (ANNs) are one of the most widely used machine learning architectures. A typical ANN is depicted in Figure 2.7 (a), which primarily involves three sections: the first section is called the *input layer* (leftmost), where the input data or the information about the system is fed into the network, followed by a set of *hidden layers* (centre) where all the processing takes place before finally providing us with the desired information at the *output layer* (rightmost). Each layer consists of neurons, denoted by circles, which can take continuous real values. Here neurons within one layer are placed below each other, such that each vertical set depicts one layer.

In a general case of classification or regression, the output layer would consist of a single neuron, which for the example of cat or dog classification, would indicate whether or not a cat is present in the image, as illustrated in Figure 2.6. However, in

this chapter, we will extensively use ANN for multi-target regression, i.e. the output layer will consist of multiple neurons, each providing different information about the system under consideration. The neurons in the intermediate levels, known as hidden layers, do not have a clear interpretation yet. But, these layers are crucial for improving the performance of the algorithm, as it has been discovered that adding more hidden layers improves the achievable level of output prediction accuracy of the neural network.

Connections between the neurons in various layers are then used to define the network structure. For instance, the input layer with N input neurons, $\mathbf{v} = (v_1, v_2, ..., v_i, ..., v_N)$, is connected to each neuron $h_j^{(1)}$ of the first hidden layer, with superscript denoting the nth layer with M_1 hidden neurons, $\mathbf{h}^{(1)} = (h_1^{(1)}, ..., h_{M_1}^{(1)})$, via weights $W_{i,j}^{(1)}$. Each neuron $h_j^{(1)}$ of the hidden layers has an additional bias $b_j^{(1)}$. In the same manner, there are connections between every neuron in the first hidden layer $h_j^{(1)}$ and every neuron in the second hidden layer $h_k^{(2)}$ having M_2 neurons, and this goes on for every successive layer. The biases similarly exist for each neuron in all hidden layers. Finally, the output layer with K neurons $\mathbf{y} = (y_1, ..., y_k)$ is connected to the penultimate layer of the network, i.e. the final hidden layer. The values of the hidden variables are calculated via non-linear functions depending on the neurons in the previous layer and the weights connecting them, as well as the biases. Considering a neuron in the first hidden layer, $h_j^{(1)}$ it gets input,

$$a_j^{(1)} = \sum_{i=1}^{N} W_{i,j}^{(1)} v_i + b_j^{(1)}, \tag{2.11}$$

which is the sum over all incoming weights multiplied with the connected input variables plus the bias of the hidden neuron. The value of the neuron is then derived via applying an activation function $\phi(a)$ onto its input

$$h_j^{(1)} = \phi(a_j^1), \tag{2.12}$$

The most commonly used activation functions are the sigmoid function,

$$\phi_\sigma(a) = \frac{1}{1+\exp(-a)}, \tag{2.13}$$

the rectified linear unit (ReLU) function,

$$\phi_{ReLU}(a) = \begin{cases} 0, \text{ if } a < 0, \\ a, \text{ else,} \end{cases} \tag{2.14}$$

and Tanh function

$$\phi_{tanh}(a) = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{2.15}$$

These are plotted in Figure 2.7 (b), where it can be seen that they are zero or negative for $a < 0$, and greater than zero otherwise, similar to the Heaviside step function. Thus, the hidden neuron gets activated if its input crosses the zero threshold.

The hidden neurons in the following layer are activated by the activation functions depending on the connections to the previous layer, and so on for further layers. The values of the neurons in the output layer are finally given by the total network function

$$y_k = \phi\left[b_k^{(L)} + \sum_{j=1}^{M_L} W_{k,j}^{(L)} \phi\left(...\phi\left[b_m^{(1)} + \sum_{i=1}^{N} W_{m,i}^{(1)} v_i\right]\right)...\right], \tag{2.16}$$

where we consider a network with L hidden layers. The values of the hidden variables are plugged in iteratively as the activation function $\phi(a)$ of the input a until the dependence on the input layer is reached. Thus, information is processed through the network from the input to the output neurons via the hidden layers, which also gives the name feed-forward neural network to the ANNs. This is indicated by the arrows in Figure 2.7 (a).

The output neurons are highly non-linear functions depending on the input neurons, with the weights and biases being adjustable parameters. These free parameters are adapted such that the network provides the desired output, where the procedure of finding the right parameters is called the training or learning of the network. Several methods to perform this training exists and can be chosen depending on the considered task. It has been mathematically proven that an ANN can approximate any function arbitrarily well due to its high non-linearity, given enough hidden neurons. This can become inefficient and computationally expensive if too many hidden layers are needed, but in principle, a suitable network can be trained for any input data set.

The choice of weights to get a certain output is not unique, as many symmetries can be found in the weight space. For example, the hidden neurons can be interchanged, and the weights can be chosen such that the network still provides the same outcome. This is also true for changing the sign of the weights. Thus, many possible choices of weights provide the same network output, and it is hard to interpret the weights found during training.

We have introduced the ANN model in a general way with an arbitrary number of hidden layers and also arbitrary numbers of neurons per layer. These numbers, as well as the activation functions chosen for the individual layers, are ingredients which define the network structure. They build a set of hyper-parameters which can be adapted suitably to find a well-performing network but cannot be derived generally. Due to the

simple correspondence of the network structure to the underlying mathematical formula for the network function, the setup can straightforwardly be generalized further to more complex structures.

2.5.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a simple, non-parametric classification algorithm that makes predictions based on the similarity between input data points. When a new data point is introduced, the algorithm identifies the k closest training samples in the feature space, usually based on Euclidean distance, and assigns the class most common among those neighbors. Despite its simplicity, KNN is effective for problems with well-separated classes and has been widely used in biomedical applications due to its ease of interpretation and implementation [Altman (1992)]. However, its performance can degrade with high-dimensional data and imbalanced classes.

2.5.3 Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It introduces randomness through bootstrapped datasets and feature selection at each node, which helps reduce overfitting and improve generalization. RF is robust to noise and capable of handling both categorical and continuous variables, making it suitable for complex biological datasets like breathomics [Breiman (2001)]. It also provides feature importance rankings, which are valuable for biomarker discovery.

2.5.4 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models that aim to find the optimal hyperplane that maximizes the margin between two classes in a high-dimensional space. By using kernel functions, SVMs can efficiently perform non-linear classification by implicitly mapping input features to a higher-dimensional space. They are particularly effective in cases where the number of features exceeds the number of samples, a common scenario in metabolomics and volatomics data. SVMs are known for their robustness and high accuracy in complex classification tasks [Cortes and Vapnik (1995)].

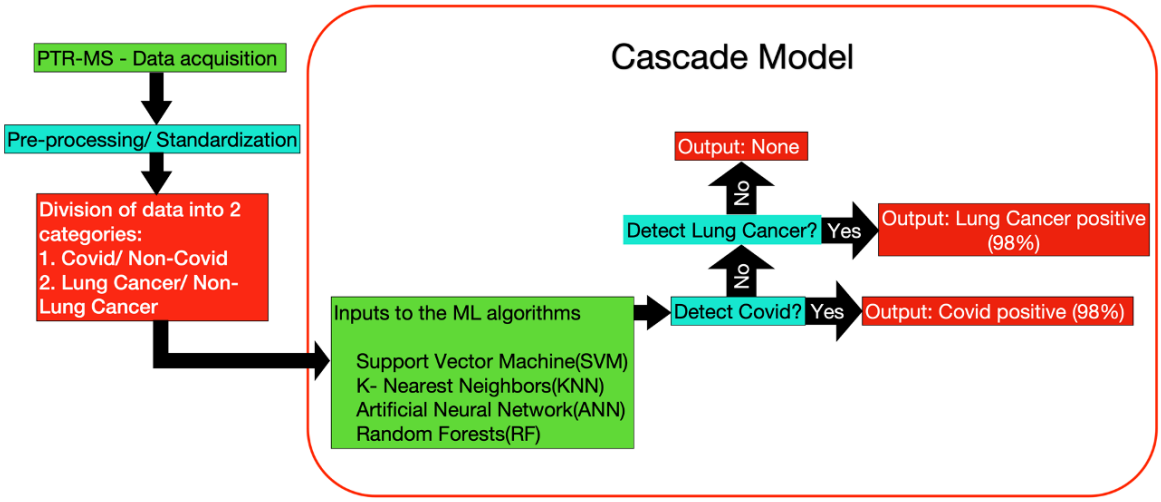


Figure 2.8: Schematic diagram of our multi-stage cascading (MSC) model, where on the left, we show the preliminary stages of the model including data acquisition and data preprocessing, while on the right, we show the core of the cascading model where the ML algorithms detect diseases at each stage and follow onto the next stage based on the output at the previous stage.

The models are trained on labeled data to recognize patterns associated with specific diseases, with each algorithm contributing to the overall prediction. To enhance the robustness and accuracy of our classification, we designed a multi-stage cascading (MSC) model shown in Fig. 2.8. In this approach, each machine learning algorithm is

deployed sequentially in stages. At each stage, the algorithm examines the breath sample for the presence of specific disease indicators. The output of each algorithm is then used to inform the next stage, creating a refined decision-making process. This cascading model ensures that the decision for positivity or negativity of a disease is based on the combined insights of multiple algorithms, reducing the risk of false positives or negatives. The final classification is achieved through a collective decision-making process, where the outputs from all algorithms are aggregated to provide a comprehensive and reliable diagnosis. This collective approach leverages the strengths of each model, which are independent of each other, to improve diagnostic accuracy and generalizability across diverse patient datasets.

Chapter 3

Results

In this chapter, we present and analyze the results obtained following the methodologies outlined in the preceding section. Each stage in the pipeline is designed to gradually build a better understanding of the underlying structure of the data and guide us toward reliable classification of the disease states based on VOC profiles.

We begin by examining the statistical distribution of individual volatile organic compounds (VOCs). Many of these compounds exhibit a log-normal-like distribution, which is typical for concentration-based measurements in biological systems. By studying these distributions, we aim to highlight key differences in VOC expression levels between different subject groups, such as healthy individuals, COVID-19 patients, and lung cancer patients. Through this comparative analysis, we identify candidate VOCs that could potentially serve as biomarkers for disease classification. These biomarkers are selected based on statistical significance and effect size (e.g., using Cohen's d), helping to narrow down the most discriminative features for downstream tasks.

Before applying machine learning algorithms for classification, we conduct several dimensionality reduction and visualization techniques to better understand the structure and separability of the data. In particular, we apply Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). These techniques allow us to project high-dimensional data into two or three dimensions for visualization while preserving important relationships between the samples.

The goal is to observe whether distinct clusters corresponding to different disease states naturally emerge in these reduced spaces. This step is essential, as it provides a qualitative assessment of how well the samples are separated based on the selected features. A lack of visible separation may indicate overlapping feature distributions, which could hinder classifier performance. On the other hand, well-separated clusters suggest the presence of meaningful patterns that machine learning models can exploit. These visualization tools therefore serve not only as exploratory analysis but also as a validation step for feature quality prior to classification.

In the following sections, we will delve deeper into each of these analyses, providing detailed plots and interpretations that build toward the final classification models and performance evaluation.

3.1 Visualization of Data for Classification: PCA, t-SNE, and UMAP

Before applying machine learning algorithms for disease classification, it is important to visually assess whether the available data exhibits separable patterns among the different classes. While statistical analysis (such as Cohen's d) can highlight potential biomarkers, they may not always provide sufficient discriminatory power on their own. This is especially critical in medical diagnosis, where a false positive can lead to unnecessary stress and interventions, and a false negative can result in missed diagnoses.

Therefore, as a preliminary step, we apply dimensionality reduction and visualization techniques to explore whether distinct clusters exist in the feature space.

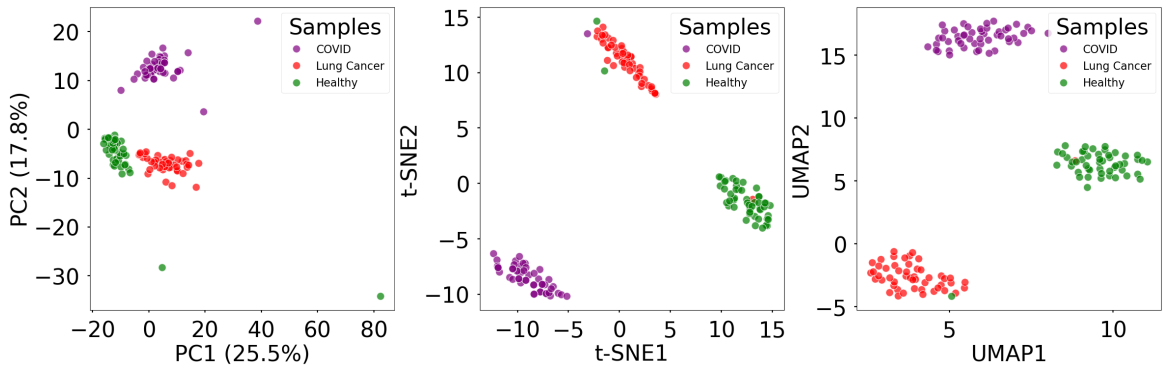


Figure 3.1: shows the 2D projections of the dataset using PCA (left), t-SNE (middle), and UMAP (right). Each point represents a subject, and the colors indicate the respective class: green for Healthy, red for Lung Cancer, and purple for COVID-19.

In this study, we employ three widely used unsupervised techniques for visualization: Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). These methods reduce the high-dimensional feature space (500+ VOCs) into two dimensions, enabling us to visually examine sample separability. The resulting plots are shown in Figure 3.1.

- **PCA Plot:** Principal Component Analysis was first applied to reduce the dimensionality of the data from 500 features to 50, preserving most of the global variance. The PCA projection retains a portion of the global variance in the first two components (25.5% and 17.8%, respectively). We observe some clustering behavior, particularly for the Healthy and COVID-19 classes. However, the Lung Cancer samples are partially overlapping with the Healthy cluster, suggesting that linear projection may not capture non-linear relationships among features effectively. This motivates the use of more sophisticated non-linear methods.

- **t-SNE Plot:** t-SNE was applied on the 50-dimensional PCA-reduced data using the following parameters: number of components = 2, perplexity = 30, learning

rate = 250, and number of iterations = 350. The resulting 2D projection reveals clearer separation among the three classes. Healthy, COVID-19, and Lung Cancer samples form distinct, tight clusters with minimal overlap. This supports the hypothesis that the dataset contains strong non-linear structure that can be leveraged for classification.

- **UMAP Plot:** UMAP was also applied on the 50-dimensional PCA-reduced data with the parameters: number of neighbors = 50, minimum distance = 0.5, number of components = 2. The UMAP projection shows well-separated clusters for each class. Notably, UMAP preserves more of the global structure of the dataset while maintaining local groupings, as reflected in the relative spacing and consistent cluster compactness. COVID-19 samples are well isolated, while Healthy and Lung Cancer samples form distinct but neighboring groups.

3.2 Bio-markers from Statistical Analysis

Following the preliminary visualizations using dimensionality reduction techniques, we now turn to a more quantitative approach to identify key biomarkers. Analysis revealed that specific volatile organic compounds (VOCs) exhibit significant discriminatory power between different disease classes. Using log-normal histogram plots, many VOCs were observed to follow approximately normal or log-normal distributions across subject groups such as Healthy, COVID-19, and Lung Cancer. This consistency justifies the application of statistical distance measures to systematically compare distributions.

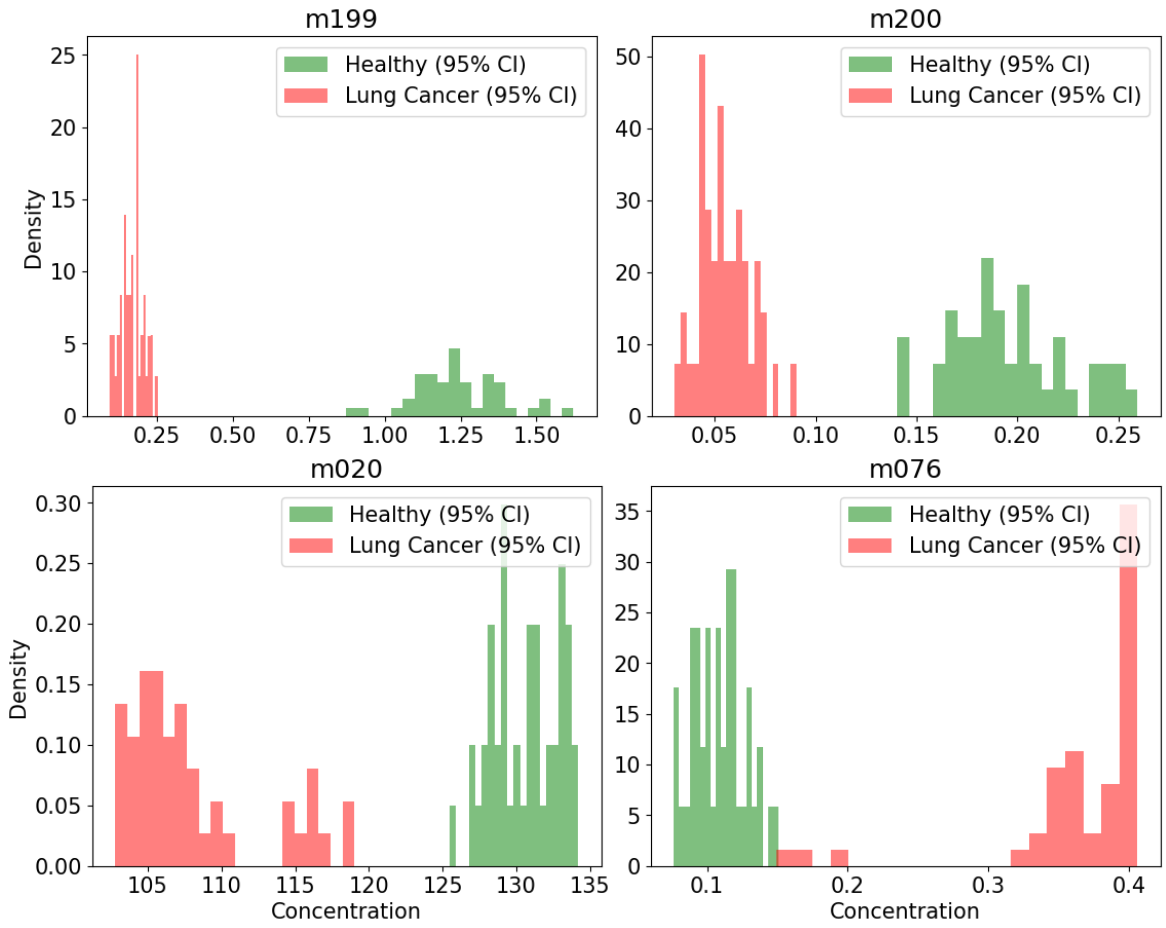


Figure 3.2: Distribution of VOCs across patients with lung cancer (legend markers).

To this end, we employ Cohen's d, a widely used effect size metric that quantifies the standardized difference between the means of two distributions. For any given VOC, Cohen's d provides a measure of how distinguishable that compound is between two classes, such as Healthy vs. Lung Cancer or Healthy vs. COVID-19. The metric is calculated as the difference in group means divided by the pooled standard deviation, and a higher Cohen's d indicates a greater separation between classes. Such compounds with high d values are considered strong candidates for biomarker selection.

This statistical analysis is particularly valuable given the high-dimensional nature of our dataset, which includes measurements for approximately 500 VOCs per sample. Directly feeding all 500 features into a machine learning model risks overfitting,

especially given the relatively modest sample size typical in clinical studies. Moreover, many of the features may be redundant or irrelevant for classification, which can degrade model performance and interpretability.

In addition to its role in feature reduction, the use of statistical effect size measures is particularly motivated by the *observed distributional overlap* between certain classes in our dataset. While the separation between Healthy and Lung Cancer samples is often visually evident in log-normal plots, this is not always the case for Healthy and COVID-19 samples. In many instances, the VOC distributions for these two groups appear *approximately overlapping*, making it difficult to visually identify discriminative features. As a result, applying a robust statistical method such as Cohen's d helps uncover *subtle but consistent differences* between the groups, enabling a more objective and reliable biomarker selection process even when class boundaries are not visually apparent.

By applying Cohen's d to each VOC and ranking them according to their discriminative power, we achieve a principled feature selection strategy. This allows us to retain only the most informative VOCs that show the largest statistical separation between disease classes while discarding noisy or less relevant features. Not only does this significantly reduce the dimensionality of the input space, but it also improves the signal-to-noise ratio for downstream machine learning tasks, enhancing both classification accuracy and model generalizability.

This approach has been effectively used in similar studies, including breath-based detection of COVID-19 and other respiratory illnesses [Canbay et al. (2024)]. In our case, the selected top-ranking VOCs based on Cohen's d serve as a refined and interpretable input for building classification models. In a later section, we apply this reduced feature set to various machine learning algorithms and analyze their performance in differentiating between disease states.

3.2.1 Cohen's *d* Biomarkers

The Table 3.1 and Table 3.2 list the top 20 volatile organic compounds (VOCs) identified as candidate biomarkers for distinguishing Lung Cancer and COVID-19 patients from healthy individuals. These VOCs were selected based on their Cohen's d values, a statistical effect size metric that quantifies the standardized difference between two group means. A higher Cohen's d indicates a greater separation between the healthy and diseased distributions, making such VOCs more relevant for diagnostic classification. This analysis serves as a crucial step in feature selection, especially given the high dimensionality of the original dataset, which contains over 500 VOC features.

Several VOCs such as m020, m022, m023, m024, and m199, appear in both Table 3.1 (Lung Cancer) and Table 3.2 (COVID-19), highlighting their potential role as non-specific markers of respiratory pathology. The presence of overlapping biomarkers suggests that there may be shared physiological or metabolic disruptions across these disease states, which manifest in altered exhaled breath profiles. By filtering the dataset to include only VOCs with the highest discriminative power, this statistical approach reduces noise and redundancy, thereby improving the efficiency and accuracy of subsequent machine learning models. These selected biomarkers will be utilized in the subsequent stages of this study for further analysis and classification using machine learning algorithms, enabling a more focused and interpretable modeling process.

Table 3.1: Top 20 VOCs Differentiating Lung Cancer from Healthy

S. No.	**VOCs**	**Cohen's *d***
1	m199	5.3922703321825
2	m076	4.53363504270905
3	m020	4.46276175634517

S. No.	VOCs	Cohen's *d*
4	m019.430	4.44674421967439
5	m022	4.36674265432898
6	m025	4.24993804998012
7	m024	4.16878366048228
8	m019.500	4.11261308031777
9	m200	4.06463985111664
10	m023	3.92748370369043
11	m040	3.87887639087529
12	m093.950	3.70079064146298
13	m095.950	3.68798519509946
14	m269	3.60590757491101
15	m145	3.58480495395544
16	m072.9500	3.56412426004183
17	m203	3.5614243658187
18	m047_Ethanol	3.26278557697531
19	m077.940	3.22922648055988
20	m261	3.18534265122926

Table 3.2: Top 20 VOCs Differentiating COVID-19 from Healthy (Based on Cohen's d values)

S. No.	**VOC**	**Cohen's *d***
1	m020	12.5932871
2	m023	10.501474
3	m022	10.4099077
4	m024	10.4815744
5	m016	9.66834248
6	m199	8.38636171
7	m025	7.82117517
8	m207.050	6.96777542
9	m183	5.84367249
10	m200	5.81029488
11	m019.430	5.3411914
12	m225	5.19249999
13	m145	4.75088929
14	m017.645	4.73318965
15	m016.750	4.68720197
16	m107.350	4.63397985
17	m215	4.39312044
18	m016.150	4.29289515
19	m015.280	4.28665812
20	m281	4.11600222

3.3 Classification with select biomarkers

To evaluate the diagnostic potential of statistically selected biomarkers, we conducted a classification study using a balanced dataset comprising 150 breath samples, 50 from Healthy individuals, 50 from COVID-19 patients, and 50 from Lung Cancer patients. These samples underwent standardized preprocessing, including removal of non-informative columns, clipping to within two standard deviations to minimize the influence of outliers, and normalization to ensure comparability across features.

Following preprocessing, we employed Cohen's d to identify VOCs that exhibited the highest discriminatory power between disease and non-disease classes. Two binary comparisons were made: Lung Cancer vs. Healthy, and COVID-19 vs. Healthy. The VOCs were ranked by their Cohen's d values, and the top 20 features for each disease were selected. Notably, some VOCs such as `m020`, `m022`, `m023`, `m024`, and `m199` appeared in the top lists for both COVID-19 and Lung Cancer, indicating their potential role as non-specific markers of respiratory pathology.

Using the selected top 20 biomarkers for Lung Cancer and COVID-19, we built classification models using two machine learning algorithms: Random Forest (RF) and K-Nearest Neighbors (KNN). The dataset was split into training, validation, and test sets using a 60%-20%-20% ratio. Two binary classification tasks were defined: (1) Lung Cancer vs. Healthy, and (2) COVID-19 vs. Healthy. Each model was trained on the training set, fine-tuned using the validation set, and evaluated on the held-out test set.

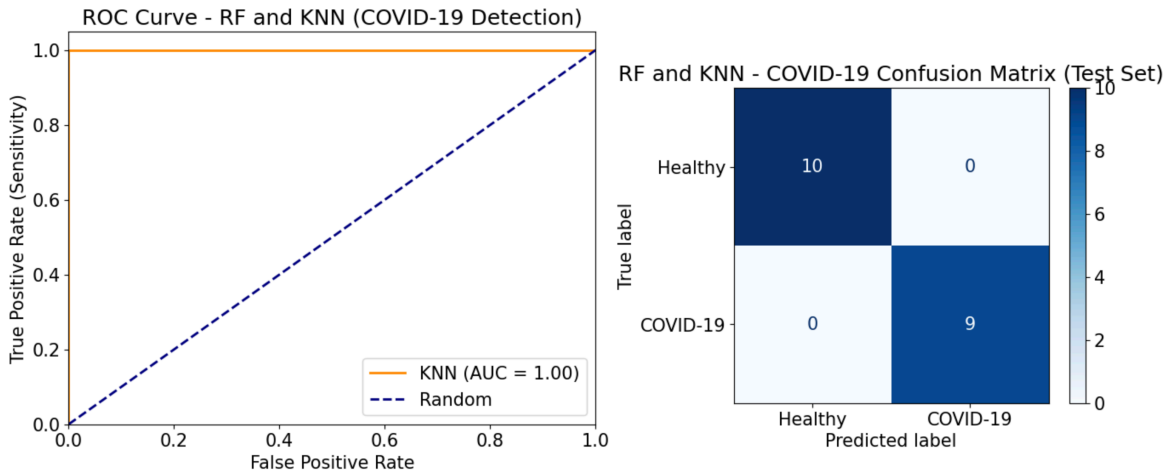


Figure 3.3: ROC curve and confusion matrix for COVID-19 classification using both Random Forest (RF) and K-Nearest Neighbors (KNN) on the test set. The left panel shows the ROC curve with an AUC of 1.00, reflecting perfect classification performance. The right panel presents the confusion matrix where both RF and KNN correctly classified all COVID-19 and Healthy samples, achieving 100% sensitivity and 100% specificity.

To assess the performance of the classifiers, we report three key evaluation metrics: sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). Sensitivity, also referred to as the true positive rate, quantifies the proportion of actual disease cases that are correctly identified by the model. Specificity measures the ability of the model to correctly identify negative cases, thereby reducing false positives. These two measures are particularly important in clinical contexts, where both missed detections and overdiagnosis can have serious consequences. The ROC curve illustrates the trade-off between sensitivity and the false positive rate (1 – specificity) across different classification thresholds. The AUC, calculated via trapezoidal integration, provides a single scalar value representing the model's discriminative ability, with values closer to 1.0 indicating better performance.

The results are summarized in Figures 3.3 and 3.4. Both classifiers — Random Forest (RF) and K-Nearest Neighbors (KNN) — demonstrated strong performance in detecting both Lung Cancer and COVID-19. For Lung Cancer classification, RF and

KNN achieved AUC scores of 0.96 and 0.95, respectively. Both models attained 90% sensitivity and 100% specificity, correctly identifying nearly all positive and negative cases. In the case of COVID-19 detection, both classifiers achieved perfect performance, with 100% sensitivity and 100% specificity on the test set. These results indicate that the selected VOC biomarkers, identified through Cohen's d statistic, effectively capture disease-specific patterns in exhaled breath and provide a robust foundation for accurate machine learning-based classification.

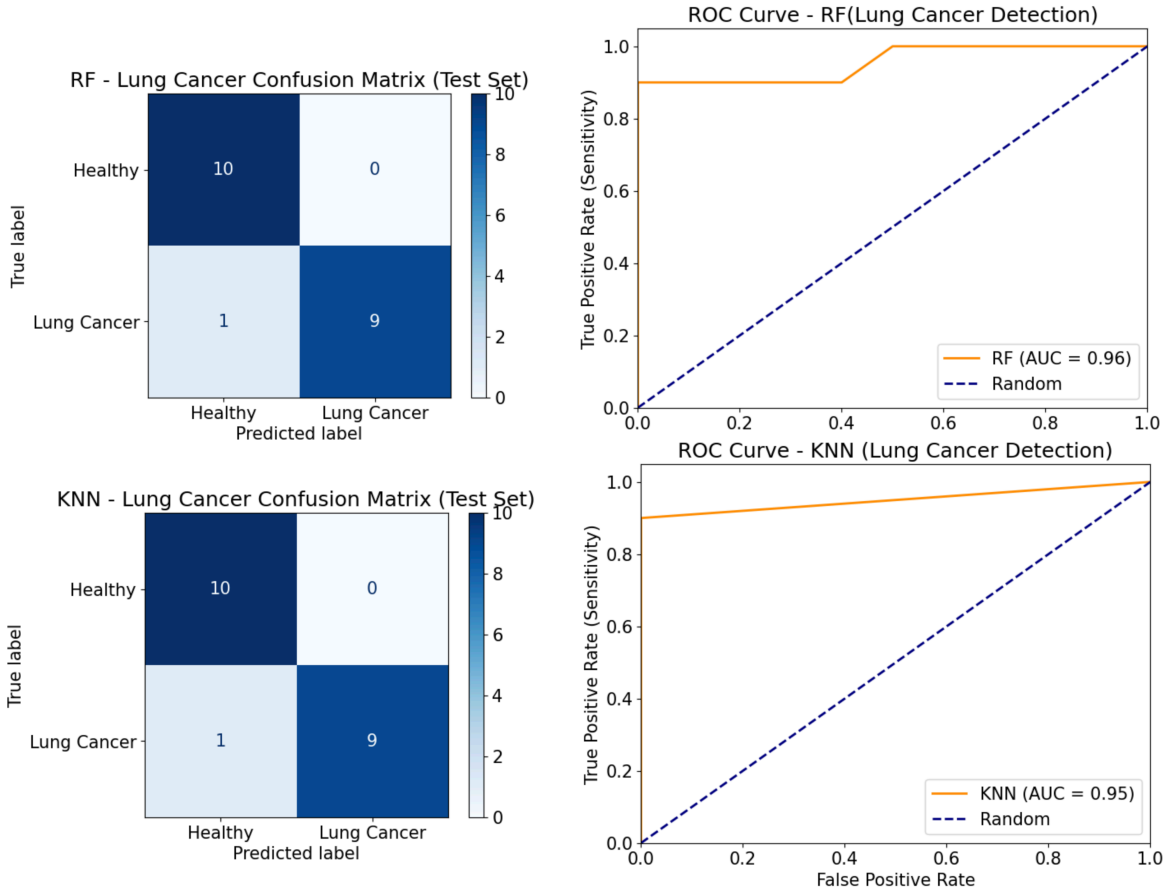


Figure 3.4: ROC curves and confusion matrices for Lung Cancer classification using Random Forest (RF) and K-Nearest Neighbors (KNN). Both models produced identical confusion matrices, achieving 90% sensitivity and 100% specificity. However, RF achieved a slightly higher AUC of 0.96 compared to 0.95 for KNN.

3.4 Classification with all features

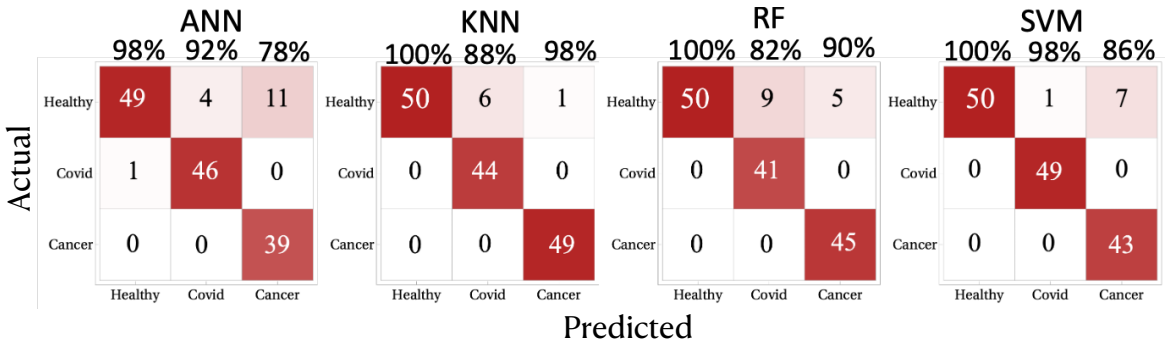


Figure 3.5: Confusion matrix indicating results from the different machine learning algorithms with (a) Artificial Neural network (ANN), (b) K-Nearest neighbor (KNN), (c) Random Forest (RF), and (d) Support vector machine (SVM).

In another approach, we explored a brute-force machine learning strategy, where all 500+ VOC features from the dataset were provided directly to supervised learning algorithms without any prior feature selection or dimensionality reduction. The goal was to evaluate whether disease-specific patterns could be effectively captured from the full high-dimensional feature space.

For this analysis, the dataset was partitioned into two binary classification tasks: Lung Cancer vs. Non-Lung Cancer and COVID-19 vs. Non-COVID-19. The Non-Lung Cancer group included samples from Healthy individuals and COVID-19 patients, while the Non-COVID-19 group consisted of Healthy and Lung Cancer samples. A range of machine learning models were trained using these labeled datasets in order to differentiate between disease groups according to their VOC signatures.

A multi-stage cascading classification strategy was employed during prediction. Each new test sample was first evaluated for COVID-19. If all algorithms in the ensemble predicted the sample as COVID-19 positive, it was labeled as "COVID-19 Positive." If not, the sample proceeded to the second stage, where it was evaluated for Lung Cancer. If all algorithms returned a positive prediction, the sample was labeled

"Lung Cancer Positive." If the sample was not positively classified in either stage, it was reported as "Negative for COVID-19 and Lung Cancer."

Figure 3.5 illustrates the classification performance of each algorithm. Among the evaluated models, **K-Nearest Neighbors (KNN)** achieved the highest classification accuracy at **96.0%**. **Support Vector Machine (SVM)** followed with **94.7%**, while **Random Forest (RF)** and **Artificial Neural Network (ANN)** yielded accuracies of **90.7%** and **89.3%**, respectively. These results are summarized in Table 3.3.

Table 3.3: Summary of classification performance for each model.

Model	**Net Accuracy**	**True Positives**	**True Negatives**	**False Positives**	**False Negatives**
ANN	89.3%	85	49	1	15
KNN	96.0%	93	50	0	7
RF	90.7%	86	50	0	14
SVM	94.7%	92	50	0	8

Chapter 4

Summary and Conclusions

4.1 Model Performance and Classification Results

This study demonstrates the potential of using volatile organic compound (VOC) profiles, collected through Proton Transfer Reaction-Mass Spectrometry (PTR-MS), for non-invasive disease classification. By integrating data from multiple sources and applying rigorous preprocessing and feature selection methods, we were able to effectively train machine learning (ML) models to distinguish between Healthy, COVID-19, and Lung Cancer subjects.

Two complementary approaches were examined. In the first, a brute-force strategy was implemented where all 500+ VOC features were provided directly to various ML algorithms. Using a multi-stage cascading classification architecture, test samples were sequentially evaluated for COVID-19 and Lung Cancer. This approach yielded strong binary classification performance, with K-Nearest Neighbors (KNN) achieving the highest accuracy (96.0%), followed by Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN).

In the second approach, we performed feature selection utilizing Cohen's d effect size to find statistically significant VOCs that may serve as potential biomarkers for each disease class. This method not only improved interpretability but also reduced model complexity, enabling effective classification using a smaller, more informative

feature set. Identifying disease-specific biomarkers with machine learning and statistical methods like Cohen's d opens the door for the creation of focused, portable diagnostic instruments.

Collectively, these approaches underscore the versatility and potential of VOC-based breath analysis when combined with customized machine learning techniques, either via comprehensive feature incorporation or biomarker-focused modeling.

4.2 Impact of Data Volume and Future Scalability

While the current models perform well, the robustness and generalizability of these results are still constrained by the available dataset size, particularly for underrepresented classes such as COVID-19. As more high-quality, labeled data becomes available, especially from diverse populations and clinical settings, the predictive accuracy of the models is expected to improve further. Future work would focus on expanding the dataset to enhance model confidence and to better account for inter-individual variability and potential confounding factors.

4.3 Proposal for On-site Implementation

To translate this research into a real-world clinical or screening environment, we propose the development of an automated machine learning (AutoML) pipeline integrated into an on-site computer system. Such a system would be capable of real-time VOC analysis, preprocessing, feature selection, and disease prediction without manual intervention. This would enable rapid, non-invasive, and scalable health screening at clinics, hospitals, or even public spaces, offering an important step forward for accessible early disease detection technologies.

Reference List

Altman, N. S., 1992: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46 (3)**, 175–185, https://doi.org/10.1080/00031305.1992.10475879.

Amann, A., B. d. L. Costello, W. Miekisch, J. Schubert, B. Buszewski, J. Pleil, N. Ratcliffe, and T. Risby, 2014: The human volatilome: volatile organic compounds (vocs) in exhaled breath, skin emanations, urine, feces and saliva. *Journal of Breath Research*, **8 (3)**, 034 001, https://doi.org/10.1088/1752-7155/8/3/034001.

Amor, R. E., M. K. Nakhleh, O. Barash, and H. Haick, 2019: Breath analysis of cancer in the present and the future. *European Respiratory Review*, **28 (152)**.

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, https://doi.org/10.1023/A:1010933404324.

Broadhurst, D. I., and D. B. Kell, 2006: Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, **2 (4)**, 171–196.

Canbay, H. S., and Coauthors, 2024: Identification of sars-cov-2 related volatile organic compounds from face masks. *Microchemical Journal*, **197**, 109 756.

Cappellin, L., F. Biasioli, A. Fabris, E. Schuhfried, C. Soukoulis, T. D. Märk, and F. Gasperi, 2011: On data analysis in ptr-tof-ms: from raw spectra to data mining. *Sensors and Actuators B: Chemical*, **155 (1)**, 183–190, https://doi.org/10.1016/j.snb.2010.11.084, URL https://www.sciencedirect.com/science/article/pii/S0925400510009135.

Cohen, J., 2013: *Statistical power analysis for the behavioral sciences*. routledge.

Cortes, C., and V. Vapnik, 1995: Support-vector networks. *Machine Learning*, **20 (3)**, 273–297, https://doi.org/10.1007/BF00994018.

Cusack, R. P., and Coauthors, 2024: Machine learning enabled detection of covid-19 pneumonia using exhaled breath analysis: a proof-of-concept study. *Journal of Breath Research*, **18 (2)**, 026 009.

Graus, M., M. Müller, and A. Hansel, 2010: High resolution ptr-tof: Quantification and formula confirmation of voc in real time. *Journal of the American Society for Mass Spectrometry*, **21**, 1037–1044, https://doi.org/10.1016/j.jasms.2010.02.006.

Hansel, A., A. Jordan, R. Holzinger, P. Prazeller, W. Vogel, and W. Lindinger, 1995: Proton transfer reaction mass spectrometry: On-line trace gas analysis at the ppb

level. *International Journal of Mass Spectrometry and Ion Processes*, **149-150**, 609–619, https://doi.org/10.1016/0168-1176(95)04294-U.

Hegen, O., J. I. Salazar Gómez, R. Schlögl, and H. Ruland, 2023: The potential of no+ and o2+• in switchable reagent ion proton transfer reaction time-of-flight mass spectrometry. *Mass Spectrometry Reviews*, **42 (5)**, 1688–1726.

Jia, Z., W. Q. Ong, F. Zhang, F. Du, V. Thavasi, and V. Thirumalai, 2024: A study of 9 common breath vocs in 504 healthy subjects using ptr-tof-ms. *Metabolomics*, **20 (4)**, 79.

Jordan, A., and Coauthors, 2009: A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (ptr-tof-ms). *International Journal of Mass Spectrometry*, **286**, 122–128, https://doi.org/10.1016/j.ijms.2009.07.005.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521 (7553)**, 436–444, https://doi.org/10.1038/nature14539.

Ligor, T., Ł. Pater, and B. Buszewski, 2015: Application of an artificial neural network model for selection of potential lung cancer biomarkers. *Journal of breath research*, **9 (2)**, 027 106.

Lindinger, W., A. Hansel, and A. Jordan, 1998: On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (ptr-ms)—medical applications, food control and environmental research. *International Journal of Mass Spectrometry and Ion Processes*, **173 (3)**, 191–241, https://doi.org/10.1016/S0168-1176(97)00281-4.

Roquencourt, C., S. Grassin-Delyle, and E. A. Thévenot, 2022: ptairms: real-time processing and analysis of ptr-tof-ms data for biomarker discovery in exhaled breath. *Bioinformatics*, **38 (7)**, 1930–1937, https://doi.org/10.1093/bioinformatics/btac031, URL https://academic.oup.com/bioinformatics/article/38/7/1930/6511435.