Laser Absorption Spectroscopy for the Detection of Lung Cancer Biomarkers in Exhaled Breath

by

Robyn Larracy

Bachelor of Science in Electrical Engineering, UNB, 2019

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Engineering

In the Graduate Academic Unit of Electrical and Computer Engineering

Supervisor:	Erik Scheme, PhD, Electrical and Computer Engineering
Examining Board:	Dawn MacIsaac, PhD, Electrical and Computer Engineering
	Kevin Englehart, PhD, Electrical and Computer Engineering
	Keith Brunt, PhD, Business, UNBSJ

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

October, 2021

Abstract

Early diagnosis of lung cancer greatly improves the likelihood of survival and remission, but limitations in existing technologies like low-dose computed tomography have prevented the implementation of widespread screening programs. Breath-based solutions that seek disease biomarkers in exhaled volatile organic compound (VOC) profiles show promise as affordable, accessible, and non-invasive alternatives to traditional imaging. In this thesis, a lung cancer detection framework using cavity ring-down spectroscopy (CRDS), an effective and practical laser absorption spectroscopy technique that has the ability to advance breath screening into clinical reality, is proposed. The main aims of this thesis were to 1) test the utility of infrared CRDS breath profiles for discriminating lung cancer patients from controls, 2) compare models with VOCs as predictors to those with patterns from the CRDS spectra (breathprints), and 3) present a robust approach for identifying relevant lung cancer biomarkers.

First, based on a proposed learning curve technique that estimated the limits of a model's performance at multiple sample sizes ($n = 10-158$), the CRDS-based models developed in this work were found to achieve classification performance comparable or superior to corresponding mass spectroscopy and sensor-based systems.

Second, using 158 collected samples (62 non-small cell lung cancer subjects and 96 controls), the accuracy range for the VOC-based model was $65.19\% - 85.44\%$ ($51.61\% - 66.13\%$ sensitivity and $73.96\% - 97.92\%$ specificity), depending on the

employed cross-validation technique. The model based on breathprint predictors generally performed better, with accuracy ranging from $71.52\% - 86.08\%$ ($58.06\% - 82.26\%$ sensitivity and $80.21\% - 88.54\%$ specificity).

Lastly, using a protocol based on consensus feature selection, three VOCs (isopropanol, dimethyl sulfide, and butyric acid) and two breathprint features (from a local binary pattern transformation of the spectra) were identified as possible lung cancer biomarkers.

This research demonstrates the potential of infrared CRDS breath profiles and the developed early-stage classification techniques for lung cancer biomarker detection and screening.

Acknowledgements

Most importantly, thank you to Dr. Erik Scheme and Dr. Angkoon Phinyomark for your guidance throughout this degree.

I would also like to thank Picomole Inc. for providing their raw breath sample data, their expertise and their support, especially Dr. Steve Graham, Dr. Gisia Beydaghyan, and Chris Purves, P.Eng. Additionally, I would like to acknowledge principal clinical investigators Dr. Tony Reiman, Dr. Luisa Galvis-Gomez, and Dr. Mahmoud Abdelsalam as well as the various other clinicians that contributed to sample collection at the Saint John Hospital, Dr. Everett Chalmers Hospital, and the Moncton Hospital.

This work was funded in part by the New Brunswick Innovation Foundation (NBIF) and the Mitacs Accelerate Fellowship in partnership with Picomole Inc.

Table of Contents

List of Tables

List of Figures

List of Symbols, Nomenclature or Abbreviations

ANN	Artificial Neural Network
AUC	Area Under (Receiver Operating Characteristic) Curve
CAT	Chronic Obstructive Pulmonary Disease Assessment Test
COPD	Chronic Obstructive Pulmonary Disease
CRDS	Cavity Ring-Down Spectroscopy
CV	Cross-Validation
ENT	Ears, Nose and Throat
LDA	Linear Discriminant Analysis
LC	Lung Cancer
LDCT	Low Dose Computed Tomography
LOOCV	Leave-One-Out Cross-Validation
mRMR	Minimum Redundancy Maximum Relevance
NSCLC	Non-Small Cell Lung Cancer
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares Discriminant Analysis
QDA	Quadratic Discriminant Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
SBS	Sequential Backward Selection
SCLC	Small Cell Lung Cancer
SFS	Sequential Forward Selection
SVM	Support Vector Machine
VOC	Volatile Organic Compound

Chapter 1

Introduction

As the deadliest and most prevalent form of cancer worldwide [1], lung cancer requires urgent and effective intervention. Unfortunately, the disease is most often discovered in its later stages when treatment options are limited and prognoses are poor. The Surveillance, Epidemiology, and End Results (SEER) Cancer Statistics Review reported that while the five year survival rate for lung cancer is 57.4% when tumors are localized to the lungs, fewer than 20% of cases are discovered at this early stage [2]. The majority of cases are diagnosed once the cancer has already metastasised and the five year survival rates drop to as low as 5.2% [2]. Early lung cancer detection is therefore crucial, yet widespread screening programs remain elusive.

While some health organizations recommend low-dose computed tomography (also called a low-dose CT scan, or LDCT) for this purpose, the technology's costs and tendency for over-diagnosis are significant hindrances [3]. The unmet need for an affordable and practical screening method has prompted a wave of research into breath-based solutions predicated on volatile organic compounds (VOCs). Thousands of VOCs have been identified in exhaled breath, ranging in concentrations from parts per trillion by volume (pptV) to parts per million by volume (ppmV) [4]. These may be exogenous, introduced into the breath through environmental exposure, or

endogenous, originating from metabolic processes within the lungs or elsewhere in the body. Because cellular metabolism is altered with disease, particular deviations in VOCs can reveal the presence of an anomaly like lung cancer [5]. If proven to be effective in the detection of such conditions, non-invasive breath collection may offer a viable path towards reliable and practical lung cancer screening.

Consequently, there has been considerable effort in recent years to establish a consistent set of breath biomarkers for lung cancer. Most commonly, breath samples are analyzed with gas chromatography-mass spectrometry (GC-MS) [6], which enables highly accurate VOC identification and quantification in a sample. Through statistical or machine learning techniques, works using GC-MS aim to characterize lung cancer by the presence, increase, or decrease of specific VOCs in the breath. Though these studies have reported encouraging results in distinguishing lung cancer subjects from non-cancer controls [7], there is very little agreement across studies in the VOCs proposed as biomarkers [8, 9]. A systematic literature review by Saalberg and Wolff [9] showed that the most frequently identified biomarkers across 52 research articles were found at most five times (i.e., 10% of the studies). The inconsistencies and contradictions across studies can be attributed, at least in part, to the absence of standardized methods for sampling, detection, and statistical analysis as well as the abundance of confounding factors in breath research [8]. Among others, these confounders include environmental conditions at the time of collection like ambient temperature, humidity, and exogeneous VOCs, as well as individual-specific differences like diet, smoking habits, gender, and comorbidities that overwhelm the weakly expressed lung cancer biomarkers [10]. Further, though very useful for laboratory research, technologies like GC-MS require time consuming, complex sample preparation and operation by experts, prohibiting scalable applications and limiting their suitability for routine clinical use [11, 12].

Motivated by the cost effectiveness, portability, and speed offered by sensor-based

technologies, another approach to breath analysis using electronic noses (e-noses) has recently emerged [7, 13]. A wide assortment of sensor types have been adopted for lung cancer breath research, including acoustic, quartz micro-balance, and metal oxide semi-conducting sensors [14, 15]. Typically unable to provide specific VOC concentrations, systems using e-nose technology rely on patterns in the sensor array's 'breathprint' for detecting disease. Therefore, unlike the traditional metabolomic (VOC-based) approach to breath analysis, machine learning techniques are often used to extract meaningful features directly from the sensor response. To date, e-noses have had promising results in capturing underlying disease signatures with this approach [7, 13]. Rather than targeting a limited number of specific VOCs, e-noses may be strengthened by their ability to provide a complete impression of the aggregate VOC profile [16]. Despite their advantages, e-nose sensors tend to be limited by low sensitivity, frequent calibration requirements, susceptibility to environmental factors like humidity and temperature, drifting, and memory effects [14, 17].

Though comparatively less common in breath research, laser absorption spectroscopy (LAS) is an attractive alternative to mass spectrometry and e-nose technologies. In recent years, advances in analyzer hardware and laser sources have progressed LAS techniques to a degree comparable to even GC-MS in sensitive, effective breath profiling [17], while maintaining low costs, quick analysis times and the ability to be operated by non-experts [18]. Like e-noses, these optical techniques can often be adapted for online analysis, eliminating the need for sample storage. However, LAS techniques also generally outperform e-noses in terms of sensitivity and robustness [18], and they permit the quantification of individual metabolites using a database of reference spectra. For all of these reasons, LAS breath tests are well-suited for real-world clinical applications and there are numerous commercially available, US Food and Drug Administration (FDA)-approved breath tests based on

portable optical technologies [18].

In this thesis, lung cancer detection systems were developed using mid-infrared breath profiles obtained through cavity ring-down spectroscopy (CRDS), an ultra-sensitive form of LAS. Using statistical and machine learning techniques, predictors from exhaled breath samples were used to discriminate 62 lung cancer patients from 96 non-cancer controls. The main aims of this thesis were to 1) evaluate the potential of CRDS for uncovering effective lung cancer breath biomarkers, 2) compare models using VOC concentrations as predictors to those with patterns from the CRDS absorption spectra (breathprints), and 3) present a robust approach for the identification of relevant lung cancer biomarkers, given the current lack of consensus in biomarkers across studies.

First, Chapter 2 provides background on breath analysis and the work performed in this field to date for the detection of lung cancer. Chapter 3 outlines the study design, sample collection and analysis techniques and provides a statistical comparison of cohort demographics and clinical factors. In Chapter 4, procedures for an exploratory statistical analysis as well as the development and validation of classification models based on VOC and breathprint features are introduced. Chapters 5 and 6 are the results and discussion, respectively, pertaining to the main experimental aims of this thesis. Finally, Chapter 7 summarizes the major findings of the studies and suggests areas for future work.

Chapter 2

Background and Literature Review

2.1 VOCs in Exhaled Breath

Each individual's exhaled breath contains a complex, unique mixture of compounds. The predominant among them are atmospheric molecules, such as water, carbon dioxide, nitrogen and oxygen, which constitute the vast majority of the breath volume [17]. In much smaller concentrations, at pptV to ppmV levels, are several hundred VOCs. These include hydrocarbons, alcohols, ketones, and aldehydes [19] produced either within the body (endogenous VOCs) or from external sources (exogenous VOCs). Endogenous VOCs originate through metabolic processes within the respiratory tract or elsewhere in the body, where they can dissolve into the blood stream and circulate to the lungs [5]. An example of an endogenous VOC is isoprene, which is thought to be related to cholesterol biosynthesis in the liver [20]. Exogenous breath VOCs, by contrast, may be introduced into the body through inhalation, skin contact, or even ingestion. After exposure, these VOCs may be immediately exhaled or they may remain in human tissue and continue presenting in the breath for short or long periods of time [5]. Breath levels of benzene and acetonitrile, for instance, have been found to be elevated for an hour and for one week, respectively, after

inhalation of cigarette smoke [8].

Although it is often difficult to distinguish these externally-produced VOCs from endogenous ones, given that many compounds have been shown to originate in both manners (such as acetone, one of the most abundant VOCs in breath [21]), both varieties hold valuable and distinct information regarding the state of an individual's health. Where exogenous VOCs may be useful markers of environmental exposure and the buildup of toxins in the body [5], those of endogenous origin have gained significant interest for their potential in detecting and monitoring disease. Altered metabolism in diseased cells, along with changes in biochemical pathways caused by the disease, affects the production of VOCs in the body. Among others, these changes may be associated with oxidative stress, overactivation of cytochrome P450 enzymes, altered lipid metabolism, and liver enzyme elevation that commonly accompany disease [22]. In turn, these alterations produce a distinct disease signature in exhaled breath, permitting the non-invasive assessment of health through breath analysis. This principle has been embraced for the study of countless conditions, including asthma, chronic obstructive pulmonary disease (COPD), sleep apnea, heart disease, neurological disorders, cystic fibrosis, and cancers [21].

2.2 Lung Cancer Biomarkers

Lung cancer detection has garnered particular attention in breath VOC analysis. Table 2.1 summarizes the techniques and participant groups considered in 24 representative works that developed classification models for discriminating lung cancer and non-lung cancer individuals in this field. Eleven applied mass spectroscopy for sample analysis, most commonly GC-MS for samples pre-concentrated with solid phase microextraction (SPME). Other varieties were proton transfer reaction mass spectroscopy (PTR-MS) and Fourier transform ion cyclotron resonance mass spec-

troscopy (FT-ICR-MS). Twelve studies were based on e-nose technologies, such as the commercially-available Cyranose 320 polymer sensor array (Sensigent, USA), as well as colorimetric, quartz microbalance (QMB), nanomaterial (gold and platinum nanoparticle), and metal oxide sensor arrays. One study [23] used ion mobility spectroscopy (IMS), which results in a chromatogram of 'VOC peaks' that can be analyzed on their own or matched to a library for VOC identification.

These works, however, varied widely in their study design. Sample sizes ranged from 17 to 484 (for lung cancer cases, 10 to 133; for controls, 5 to 361). While many studies included lung cancer (LC) patients of all stages, some narrowed their analysis to include only early-stage or late-stage cases. Similarly, many chose to exclude small-cell lung cancer (SCLC) patients, targeting instead only the more common non-small cell lung cancer (NSCLC) histologic subtypes. Several studies also assumed exclusions for patients that had undergone treatment for their cancer. The control cohorts differed across studies as well, with some consisting of only healthy volunteers and others including various lung conditions like COPD or benign pulmonary nodules. In one case [24], hospital staff at the collection site were recruited as control subjects to ensure similar exposure to environmental VOCs.

Table 2.1: Summary of related works in breath VOC lung cancer research from 1999 to 2020.

Reference		Technique	Cases		Controls
Phillips et al. [25]		GC-MS	60 Untreated LC	48	Benign Lung Abnormality
Machado et al. [26]		Cyranose 320	14 Untreated LC	62	Healthy & Non-Cancer Lung Diseases
Mazzone et al. [27]		Colorimetric Sensors	49 NSCLC	94	Healthy & Non-Cancer Lung Diseases
Wehinger et al. [24]		PTR-MS	17 Untreated LC	170	Hospital Staff & Healthy Volunteers
Dragonieri et al. [28]	(a)	Cyranose 320	10 NSCLC	10	Healthy
	(b)	Cyranose 320	10 NSCLC	10	COPD
Westhoff et al. [23]		IMS	32 Untreated LC	54	Healthy
Bajtarevic et al. [29]		SPME/GC-MS	65 LC	31	Healthy
D'Amico et al. [30]	(a)	QMB Sensors	28 LC	36	Healthy
	(b)	QMB Sensors	28 LC	28	Non-Cancer Lung Diseases
Wang et al. [31]		SPME/GC-MS	85 Untreated LC	158	Healthy & Benign Lung Disease
Peled et al. [32]		Nanomaterial Sensor Array	49 Untreated LC	19	Benign Lung Nodules
Broza et al. [33]		Nanomaterial Sensor Array	12 Early-Stage LC	5	Benign Lung Nodules
Bousamra et al. [34]		FT-ICR-MS	107 Early-Stage LC	40	Benign Lung Nodules
Hubers et al. [35]		Cyranose 320	38 LC	39	Non-Cancer
Ligor et al. [36]		SPME/GC-MS	123 Late-Stage LC	361	Healthy
Li et al. [37]	(a)	SPME/GC-MS	85 Untreated LC	40	Healthy Smokers
	(b)	SPME/GC-MS	85 Untreated LC	45	Healthy Non-Smokers

Table 2.1: Continued from previous page.

Reference		Technique		Cases		Controls
	(c)	SPME/GC-MS	85	Untreated LC	34	Benign Lung Nodules
Corradi et al. [38]		SPME/GC-MS	71	NSCLC	67	Benign Lung Nodules
Chang et al. [39]		Metal Oxide Sensors	37	NSCLC	48	Healthy
Shlomi et al. [40]		Nanomaterial Sensor Array	16	Untreated Early-Stage LC	30	Benign Lung Nodules
Sakumura et al. [41]		GC-MS	107	LC	29	Healthy
van de Goor et al. [42]		Metal Oxide Sensors	60	Untreated LC	107	Benign ENT Conditions
Tirzite et al. [43]	(a)	Cyranose 320	133	Untreated LC Non-Smokers	132	Non-Cancer Non-Smokers
	(b)	Cyranose 320	119	Untreated LC Smokers	91	Non-Cancer Smokers
Kononov et al. [44]		Metal Oxide Sensors	65	LC	53	Healthy
Rudnicka et al. [45]		SPME/GC-MS	108	LC	121	Healthy
Koureas et al. [46]		SPME/GC-MS	51	LC	53	Healthy

The cancer detection techniques employed in each of the studies, namely feature extraction, feature selection, and classification methods, are presented in Table 2.2. None of the studies employing mass spectroscopy analysis performed explicit feature extraction, as these works generally aimed to identify interpretable VOC biomarkers. Instead, these works passed VOC data directly to the feature selection and classification algorithms. Studies based on e-nose technologies, contrarily, typically relied on feature extraction techniques to enhance discriminative patterns from raw sensor readings. Many works focused on basic characteristics of the sensor signals, such as the peak resistance during the sensor's exposure to the sample or the integral of the sensor's resistance over time. Principal component analysis (PCA) was also employed in a few instances, which is a common dimensionality reduction technique that eliminates redundancy while maintaining the variability of the data, making it useful for high-dimensional sensor data. Similarly, in one work [42], a more advanced tensor decomposition technique was applied for dimensionality reduction.

While some studies bypassed feature selection and used all features for classification, many applied wrapper or filter techniques (or a combination of the two) for identifying discriminating features. Wrapper-style sequential forward selection (SFS) and sequential backward selection (SBS) were commonly employed, where optimal subsets of features are found through successive addition (SFS) or removal (SBS) of features from a candidate feature set and re-evaluation of a criterion. This criterion was most often classification error rate, although some works used stepwise discriminant analysis where F-values (representing a feature's significance to the classification model) were used as a criterion. Other works used simpler filter methods, based on t-tests, Wilcoxon rank sum tests, correlation coefficients, or the area under receiver operating characteristic (ROC) curves (AUC) to select useful features. Additionally, one work [36] employed a combination of PCA and hierarchical clustering to select relevant features, and another [41] performed an exhaustive search

of all possible combinations of features to find the set that maximized classification performance.

For classification, the most commonly employed learning algorithm was linear discriminant analysis (LDA), a probability-based method where class associations are dictated by a linear decision boundary. Others included quadratic discriminant analysis (QDA), support vector machines (SVM), random forests (RF), partial least squares discriminant analysis (PLS-DA), logistic regression models, and feedforward artificial neural networks (ANN). A few studies employed custom decision rules, using thresholding and majority vote strategies, for instance.

Although some models exhibit exceptional classification performance, even perfect or near-perfect accuracies, many are based on small sample sizes and lack proper validation [7]. Further, the VOC predictors used as features are highly inconsistent across studies [8], suggesting that results have been heavily influenced by noise from confounders. Beyond the analysis techniques and inclusion/exclusion criteria for the cohorts outlined in Table 2.1, studies varied in the volume of breath collected, the type of breath collected (alveolar, tidal or vital capacity breath), sampler technology, storage containers, duration of storage, and restrictions on smoking, eating and alcohol consumption prior to collection, to name a few. Further, during data analysis, some works considered room air measurements for correcting or informing breath data [24,25,27], while others explicitly did not [29]. Some even included clinical predictors like age and smoking history as additional classification features [24, 38, 43]. All of these factors may significantly impact the results of a study, making it difficult to draw consensus across the field.

As of yet, no clinically relevant biomarkers have emerged and no models have been sufficiently validated for clinical use. The demonstrated lack of accord in previous studies and the unavoidable presence of confounders points to a need for noise-robust machine learning and biomarker identification techniques. Feature extraction,

which has consisted mainly of basic sensing features and PCA to date, may be key to uncovering useful patterns among irrelevant signals from confounders. Further, rather than the single feature selection step typically performed for determining potential biomarkers, more reliable search methods are needed.

Table 2.2: Machine learning techniques (extraction: feature extraction, selection: feature selection, and classifiers) and classification performance estimates (acc.: classification accuracy, sen.: sensitivity, spe.: specificity, in %) in related breath VOC lung cancer works from 1999 to 2020.

Reference		**Extraction**	**Selection**	**Classifier**	**Acc.**	**Sen.**	**Spe.**
Phillips et al. [25]		n/a	SFS (F-Statistic)	LDA	69.4	71.7	66.7
Machado et al. [26]		Sensing Features (Resistance Changes)	n/a	SVM	88.2	71.4	91.9
Mazzone et al. [27]		Sensing Features (Color Changes)	n/a	RF	72.7	73.3	72.4
Wehinger et al. [24]		n/a	Wilcoxon Rank Sum Test Filtering	QDA	96	54	99
Dragonieri et al. [28]	(a)	PCA	n/a	LDA	90	n/a	n/a
	(b)	PCA	n/a	LDA	85	n/a	n/a
Westhoff et al. [23]		Peak Clustering	t-Test Filtering & SBS (Error Rate)	LDA	100	n/a	n/a
Bajtarevic et al. [29]	(a)	n/a	SFS (Error Rate)	Decision Rule	67.5	52	100
	(b)	n/a	SFS (Error Rate)	Decision Rule	80.4	71	100
	(c)	n/a	SFS (Error Rate)	Decision Rule	86.5	80	100
D'Amico et al. [30]	(a)	Sensing Features (Frequency Shift)	n/a	PLS-DA	93.8	85	100

Table 2.2: Continued from previous page.

Reference		Extraction	Selection	Classifier	Acc.	Sens.	Spec.
	(b)	Sensing Features (Frequency Shift)	n/a	PLS-DA	85.7	92.8	78.6
Wang et al. [31]		n/a	AUC Filtering	LDA	97.1	96.5	97.5
Peled et al. [32]		Sensing Features (Resistance Changes)	Wilcoxon Rank Sum	LDA	88	86	96
Broza et al. [33]		Sensing Features (Resistance Changes)	SFS (F-Statistic)	LDA	94.1	100	80
Bousamra et al. [34]		n/a	Wilcoxon Rank Sum	Decision Rule	85.0	88	77
Hubers et al. [35]		PCA	t-Test Filtering	Decision Rule	69.2	94.4	12.5
Ligor et al. [36]		n/a	PCA Selection and Clustering	ANN	70.1	63.5	72.4
Li et al. [37]	(a)	n/a	SBS (Error Rate)	Decision Rule	97	96	100
	(b)	n/a	SBS (Error Rate)	Decision Rule	95	100	86
	(c)	n/a	SBS (Error Rate)	Decision Rule	89	100	64
Corradi et al. [38]	(a)	n/a	Wilcoxon Rank Sum & Spearman Correlation Filtering	Logistic Regression	63.8	60.6	67.2
	(b)	n/a	Wilcoxon Rank Sum & Spearman Correlation Filtering	Logistic Regression	74.1	72.5	75.8
Chang et al. [39]		PCA	n/a	ANN	75	79	72
Shlomi et al. [40]		Sensing Features (Resistance Changes)	SFS (F-Statistic)	LDA	87.0	75	93.3

Table 2.2: Continued from previous page.

Reference		Extraction	Selection	Classifier	Acc.	Sens.	Spec.
Sakumura et al. [41]		n/a	Exhaustive Search (Error Rate)	SVM	89	92.5	75.9
van de Goor et al. [42]		Tensor Decomposition	n/a	ANN	86	88	86
Tirzite et al. [43]	(a)	Sensing Features (Maxima, Integral)	n/a	Logistic Regression	93.4	96.2	90.6
	(b)	Sensing Features (Maxima, Integral)	n/a	Logistic Regression	94.3	95.8	92.3
Kononov et al. [44]		n/a	n/a	Logistic Regression	97.2	95	100
Rudnicka et al. [45]		n/a	Wilcoxon Rank Sum Filtering, SFS (F-Statistic)	ANN	86.4	86.4	86.4
Koureas et al. [46]		n/a	n/a	RF	88.5	n/a	n/a

Chapter 3

Data

3.1 Equipment

Infrared breath profiles were measured by Picomole Inc. using CRDS. CRDS uses highly reflective mirrors (>99.8% in this work) to increase the effective path length of light trapped in an optical cavity. For a sample within the cavity, the decay rate (ring-down) of the trapped light is measured to determine the sample's light absorption. Two narrow linewidth (100kHz) CO_2 lasers with carbon isotopes ^{12}C and ^{13}C were tuned to a combined 73 lines in the mid-infrared region (see Table 3.1), a favourable spectral range for the detection of the small molecules that are of interest in breath analysis [11]. Specifically, the wavelengths spanned approximately 9.2-11.3 μm, where the decay time of light in the cavity is approximately one to two microseconds. At each wavelength, the average times from 500 ring-downs were measured for the breath sample (τ) and for a baseline nitrogen sample (τ_0). The absorption coefficients K comprising each spectrum were calculated from the average ring-down times according to (3.1), where c is the speed of light.

$$K = \frac{\tau_0 - \tau}{c \cdot \tau_0 \cdot \tau} \tag{3.1}$$

Table 3.1: Lines from the two CO_2 lasers used for the CRDS analysis.

Laser	**Branch**	**Line IDs**	**Wavelengths (μm)**
$^{12}CO_2$	9R	9R28, 9R26, 9R24, 9R22, 9R20, 9R18 9R16, 9R14, 9R12, 9R10	9.2295-9.3294
	9P	9P16, 9P20, 9P22, 9P24, 9P26, 9P28, 9P30	9.4731-9.6392
	10R	10R32, 10R30, 10R28, 10R26, 10R24, 10R22, 10R20, 10R18, 10R16, 10R14, 10R12, 10R10, 10R08	10.1703-10.3337
	10P	10P08, 10P10,10P12,10P14, 10P16, 10P18, 10P20, 10P22, 10P24, 10P26, 10P28, 10P30, 10P32	10.4762-10.7186
$^{13}CO_2$	10R	10R36, 10R34, 10R32, 10R30, 10R28, 10R26, 10R24, 10R22, 10R20, 10R18, 10R16, 10R14, 10R12, 10R10, 10R08, 10R06	10.6522-10.8844
	10P	10P08, 10P10, 10P12, 10P14, 10P16, 10P18, 10P20, 10P22, 10P24, 10P26, 10P28, 10P30, 10P32, 10P34	11.0247-11.3099

The analysis was performed four times for each sample, at desorption temperatures of 75°C, 150°C, 225°C, and 300°C, yielding four different spectra for a subject. Certain volatiles require higher temperatures than others to be fully released from the Tenax TA adsorbent material, so the four spectra may characterize different assortments of VOCs. Fig. 3.1 depicts the four spectra obtained for one random subject.

3.2 Sample Size

Due to the expense associated with sample collection involving human subjects, the sample size required to develop the early-stage screening system was considered prior to collection. There is no gold standard approach to sample size estimation for machine learning applications, and the number of samples needed to effectively train and validate a model depends largely on the complexity of the problem. Across 55 lung cancer breath studies from 1999 to 2019 [7,47], the average number of recruited

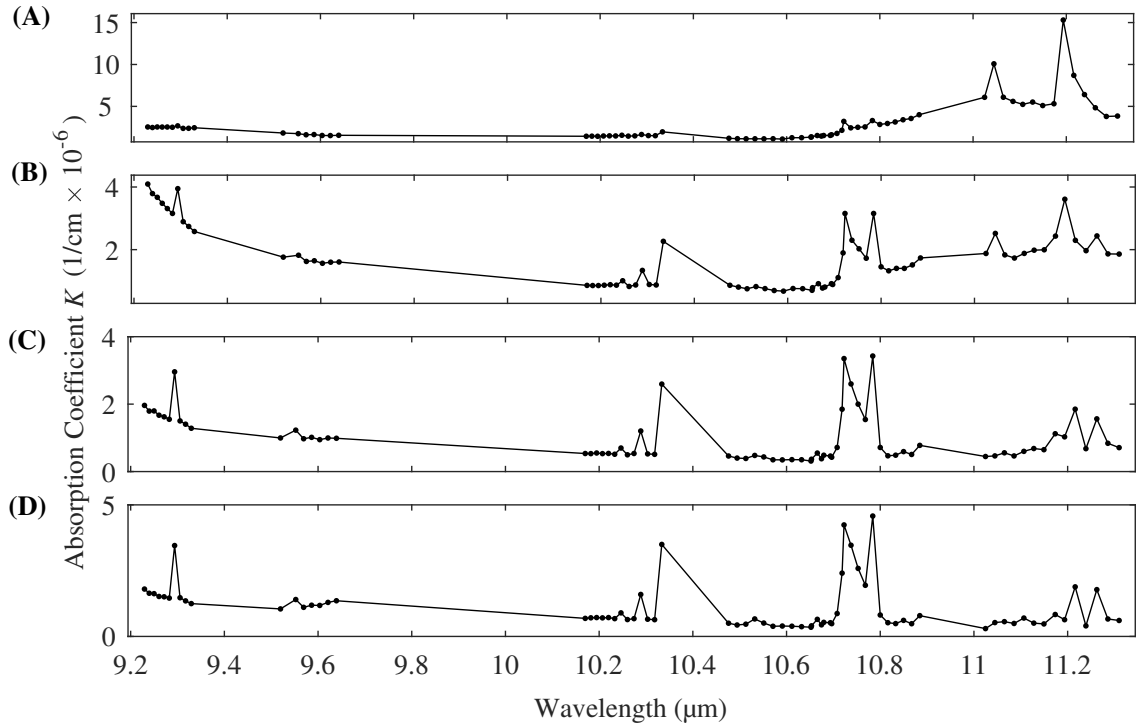

Figure 3.1: Example spectra for one sample, desorbed at four different temperatures: A) 75°C, B) 150°C, C) 225°C, D) 300°C.

lung cancer participants was 58 (median 37, range 5 to 233) and the average number of non-cancer controls was 71 (median 42, range 5 to 259). However, it is important to note that low sample sizes likely contributed to the inconsistent results across studies, and many studies did not budget samples for proper classifier validation [7].

Hence, the sample size requirements were estimated through analytical means. First, power calculations were performed using G*Power V3.1.9.4 [48]. Assuming a two-tailed 5% type I error rate and a power of 95%, a sample size of 42 subjects per group (84 in total) was found to be necessary for detecting a difference of effect size $d = 0.8$. This value reflects the median number of samples collected in prior works. Given that statistical significance does not necessarily translate to satisfactory classification ability and, further, the power calculation does not account for the high dimensionality of the data (292 measurements per subject; 73 for each spectrum), it was anticipated that 84 subjects may be too few for the model development.

Consequently, the required sample size was estimated from a classification perspective using simulated learning curves. Balanced two-class datasets spanning sample sizes 10 through 600 were synthesized, each consisting of 292 features randomly drawn from unit variance Gaussian distributions. While most features were drawn from the same zero-mean distribution for both classes, representing irrelevant features (or noise), ten percent of the features (29) were made to be differentiable between the two classes by increasing the mean of the positive class distribution to 0.8. This represents an effect size of 0.8 for each of the 29 discriminable features. For each dataset, a nested leave-one-out cross-validation framework was implemented to develop and validate a classification model. In each cross-validation fold, the top 10% of features were selected using t-tests and class associations were learned using a linear support vector machine (SVM). In addition to classification accuracy, the percentage of discriminable features that were correctly identified through the t-test procedure was recorded. This was repeated 100 times and recorded performance metrics were averaged to create the learning curves. An additional dataset of 100,000 samples was simulated for estimating the true classification ability afforded by the distributions of the 29 discriminative features, with 50% randomly selected for training and 50% for testing an SVM classifier.

As evidenced in Fig. 3.2, even assuming an optimistic case with 10% of the measured features exhibiting a large $d = 0.8$ effect, 84 samples were not sufficient. In fact, in this example, 200 samples were necessary to consistently identify the relevant features and to converge within 2% of the true accuracy of the problem (98.04%). It should be noted that this learning curve approach cannot definitively predict sample requirements, as feature distributions are unknown prior to collection. Presumably, real-world measurements will exhibit various distributions and effect sizes, and correlated measurements should be expected. Nevertheless, this illustrative learning curve example along with the range of sample sizes utilized in previous works

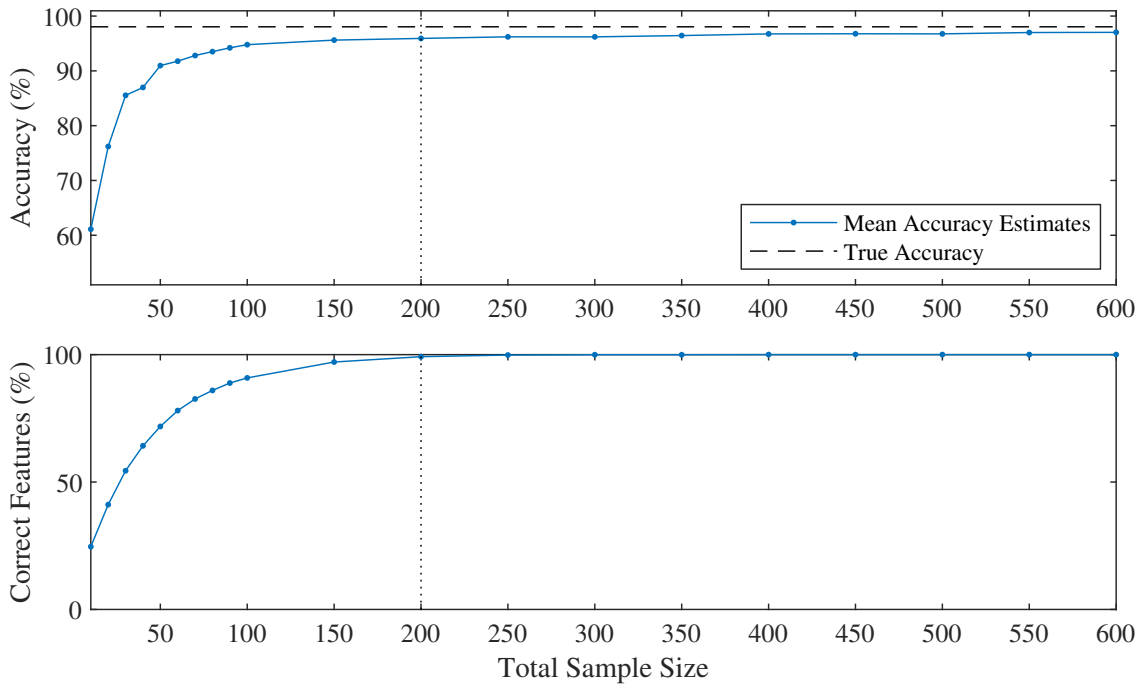

Figure 3.2: Classification accuracy and the proportion of correctly selected features for the simulated datasets as sample size was increased.

suggested that approximately 200 samples (100 per group) may be a reasonable starting point for a prototype model.

3.3 Sample Collection

One hundred biopsy-confirmed lung cancer patients and 98 non-cancer control subjects were enrolled in the study to provide breath samples. All subjects provided informed consent as per the Horizon Health Network's Research Ethics Board (#100099), and analyses of the collected data were conducted as approved by the University of New Brunswick's Research Ethics Board (#2019-068).

Collection was performed at three hospitals using an exhaled breath sampler developed by Picomole Inc. [49], which tracks CO_2 levels to collect alveolar breath into Tenax TA sorbent tubes (Figure 3.3). Before their shipment to sample collection sites, all sorbent tubes were conditioned and batch tested to ensure low background

levels. Subjects were also asked to abstain from smoking for 4 hours and drinking alcohol for 8 hours prior to collection, where they were instructed to breathe deeply and exhale into a single-use filter on the sampler's mouthpiece until 10-litre (10L) samples were amassed. Apart from shipment periods, collected samples were maintained at -20°C until analysis.

Figure 3.3: Usage of the exhaled breath sampler developed by Picomole Inc. [49].

Post-collection, the inclusion criteria for lung cancer subjects were amended to exclude eleven otherwise eligible patients with ambiguous or small-cell histologic subtypes. Given the limited representation in the dataset for these subtypes, the analysis was focused on non-small cell lung cancer (NSCLC) patients which constitute approximately 85% of the lung cancer population. Also disqualifying subjects that had missing data (for example, if they were unable to provide the full 10L sample) and patients that had undergone any form of lung cancer treatment, the

remaining 62 pre-treatment NSCLC patients and 96 control subjects were included in the analysis.

A comparison of demographics and clinical factors for the two cohorts is provided in Table 3.2. These include sex, smoking habits, comorbid lung conditions, age, self-assessed breathing symptom scores from the COPD Assessment Test (CAT; scores are 0 at best and 5 at worst [50]), overall CAT score (sum of all breathing symptom scores, 0 at best and 40 at worst), and histologic subtypes for the NSCLC patients. Other lung conditions found in the two cohorts were COPD, asthma, pneumonia and bronchitis. Data for these factors were available for all subjects, except for smoking habits which were known for 152 of 158 subjects. The presented group proportions are corrected for this missing data.

Table 3.2: Demographics and clinical factors for the lung cancer and control cohorts.

Factor	**Control** $(\mu \pm \sigma)$	**Lung Cancer** $(\mu \pm \sigma)$	***p*-value**
Sample Size	96	62	-
Sex			
Female	53.1 %	50%	0.75
Male	46.9 %	50%	
Smoking			
Current smokers	6.7%	19.4%	
Former smokers	48.9%	75.8%	$< 0.0001^{\dagger}$
Never smokers	44.4%	4.8%	
Other Lung Conditions	36 (37.5%)	44 (71.0%)	$< 0.0001^{\dagger}$
Age (yrs)	63.3 ± 13.5	69.7 ± 8.8	0.004^{*}
Cough	1.2 ± 1.0	2.5 ± 1.2	$< 0.0001^{*}$
Chest Mucus	0.5 ± 0.9	1.9 ± 1.5	$< 0.0001^{*}$
Chest Tightness	0.2 ± 0.6	1.0 ± 1.4	$< 0.0001^{*}$
Breathlessness	0.9 ± 1.2	2.4 ± 1.8	$< 0.0001^{*}$
Confinement to Home	0.2 ± 0.6	1.7 ± 1.7	$< 0.0001^{*}$
Concern Leaving Home	0.1 ± 0.2	0.5 ± 1.2	0.0007^{*}
Problems Sleeping	0.6 ± 1.0	1.3 ± 1.6	0.0004^{*}
Low Energy	0.9 ± 1.1	2.5 ± 1.5	$< 0.0001^{*}$
CAT Score	4.5 ± 4.1	13.7 ± 8.3	$< 0.0001^{*}$
Diagnosis			
Adenocarcinoma	-	58.1%	
Squamous cell carcinoma	-	37.1%	-
Unspecified NSCLC	-	4.8%	

* t-test indicated a significant difference between groups ($p < 0.05$)
† Fisher's exact test indicated a significant difference between groups ($p < 0.05$)

Chapter 4

Methods

4.1 Exploratory Data Analysis

As an initial exploration of the dataset, to add interpretability to prospective classification models and inform future sample collection, statistical analysis was performed for the CRDS measurements. Two representations of the data were considered: the raw absorption coefficients at each wavelength in units of 1/cm (73 variables per desorb temperature), and the concentrations of the compounds in ppbV identified in each spectrum through a spectral regression procedure (described in Section 4.2, 150 variables per desorb temperature). First, each of the variables were compared between the lung cancer and control cohorts to assess their potential for classification. Second, various subgroups were compared to gauge the effects of confounding factors on the measurements. These factors included sex (female vs. male), smoking habits (never-smokers vs. former smokers, non-smokers vs. smokers), age (under 55 vs. 55 and above), and CAT score (scores below 6 vs. scores 6 and higher).

Through a visual inspection, the absorption coefficients were found to follow approximately log-normal distributions, as do most biological measurements [51]. For this reason, log-transformed coefficients were used for all comparisons. The

choice of test for each comparison was determined by a Lilliefors test for normality: if the test indicated a non-normal distribution for either group's log-transformed coefficients at a significance level of 0.05, a non-parametric Wilcoxon rank sum test was performed. Otherwise, a t-test was performed for that variable.

For the compound concentrations, approximately log-normal but zero-inflated distributions were observed due to the absence or inability to detect many of the examined 150 compounds in the spectra. Hence, a two-part testing procedure was implemented as described by Gleiss et al. [52]. Two tests were performed: 1) a χ^2 test to compare the proportions of zero-concentrations between groups, and 2) a t-test or a Wilcoxon rank sum test to compare log-transformed, non-zero concentrations between groups. As with the absorption coefficients, the choice of t-test or Wilcoxon rank sum test was determined by the result of a Lilliefors test. The two-part test statistic $\chi^2_{(2)}$, which follows a χ^2 distribution with two degrees of freedom, was given by (4.1) where $\chi^2_{(1)}$ is the continuity-corrected test statistic from the χ^2 test and U is the continuity-corrected and normalized test statistic from the Wilcoxon rank sum test [53].

$$\chi^2_{(2)} = \chi^2_{(1)} + U^2 \tag{4.1}$$

In cases where one group consisted of only zero values, p-values were derived from a χ^2 distribution with one degree of freedom using only the $\chi^2_{(1)}$ test statistic.

As the recommended descriptors for log-normal data [51], geometric mean (μ_g) and geometric standard deviation (σ_g) were reported as summary statistics. These are given by equations (4.2) and (4.3), respectively, where x is the variable of interest and n is the number of subjects with a non-missing value (in the case of absorption coefficients) or non-zero value (in the case of VOC concentrations) for the variable.

$$\mu_g = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{4.2}$$

$$\sigma_g = exp\left(\sqrt{\frac{\sum_{i=1}^{n}\left(\ln\frac{x_i}{\mu_g}\right)^2}{n}}\right) \tag{4.3}$$

4.2 VOC Features

The first classification model developed was based on the traditional metabolite approach to breath analysis, which requires knowledge of the individual VOC constituents in a sample. Quantitative estimation of the compounds present in each spectrum was performed with a stepwise linear fitting algorithm [54]. Using a reference library of 152 absorption cross-sections for commonly-occurring compounds in human breath (see Appendix A) from the Quantitative Infrared Database provided by Pacific Northwest National Laboratory [55, 56] and the HITRAN 2016 database [57], compounds were iteratively added or removed from the regression model based on the Akaike Information Criterion (AIC) for each potential fit given by (4.4). RSS represents the sum of squares of the fit residual, N_y is the number of measurements in the spectrum, N_x is the number of compounds in the fit, and AIC_k is a parameter used to control the balance between the goodness of fit and entropy, fixed at 2 in this work.

$$AIC = N_y \log\left(\frac{RSS}{N_y}\right) + AIC_k N_x \tag{4.4}$$

The stepwise procedure was stopped when the AIC could not be further improved by adding or removing a compound from the fit. In case of missing absorption measurements (on average, 1.8 missing per spectrum), these wavelengths were ignored.

The fitting procedure resulted in a vector of compound concentrations in ppbV

for each spectrum, which were log-transformed to create the final VOC-based feature matrix for classification. Two compounds, water and carbon dioxide, were included in the reference library for the fitting procedure but were disregarded for further analysis as they were mostly flushed from the tubes prior to measurement and are not considered to be viable biomarkers. In total, there were 600 VOC-based features: 150 compounds for each of the four desorption temperatures, where compounds that were not found in a spectrum were assigned a concentration of zero ppbV.

4.3 Spectral Breathprint Features

For the second approach, spectral breathprint features were derived directly from the CRDS absorption profiles through pre-processing and feature extraction methods. Due to the high variability in breathprints observed across subjects, in this study, the crafted features emphasized the shape of the spectra rather than individual absorption values.

First, a piecewise cubic spline interpolant was fit to the available wavelengths and any missing absorption values were inferred. More details about data pre-processing of spectral breath breathprints can be found in Appendix B. Next, the first order and the second order spectral derivatives were extracted to highlight peaks and troughs in the spectra. Along with the raw spectra, these derivatives were then transformed using a one-dimensional local binary pattern (1D-LBP) feature extraction technique [58], a modification of the two-dimensional method commonly used for texture analysis in image processing applications. The scale-invariant LBP features describe the structure of an input sequence by capturing relationships between neighboring points. Briefly, a nine-point moving window was used to create a series of eight-bit binary codes for each input spectrum or derivative, where ones represent points in the window that are greater than its center value and zeros represent

points that are smaller (see Figure 4.1). For example, a window with a peak at its center, surrounded by four smaller values on either side, would be assigned a code of 00000000. The LBP series were further transformed by counting the frequencies of the 256 possible patterns to produce a set of histogram features. In this thesis, only the histogram features corresponding to the 58 'uniform' eight-bit patterns, those which contain at most two bitwise transitions from zero to one or the reverse, were included in the final feature matrix. These uniform patterns are less likely to capture random processes than the more complex, less frequent non-uniform patterns, and have been shown to contain the most discriminative information [59]. These 58 uniform patterns are provided in Table 4.1. For the remainder of this thesis, breathprint features are denoted by the spectrum type (raw or derivative: 1^{st} or 2^{nd}) and the 8-bit LBP code in its decimal representation.

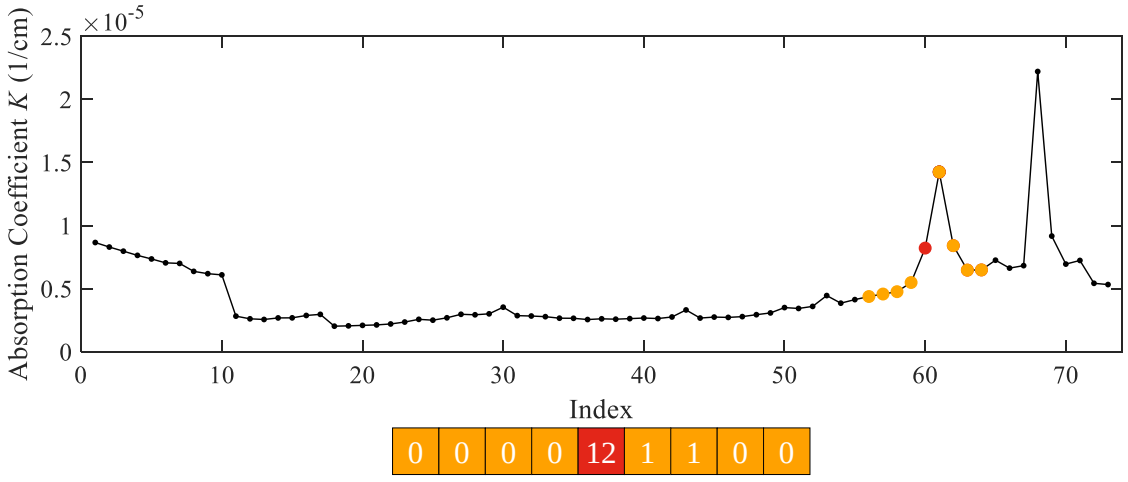

Figure 4.1: Illustration of the 1D-LBP procedure for a raw absorption spectrum. For the point at the center of the moving window (red), the LBP code describes its value in relation to the other points in the window (yellow).

4.4 Classification Models

For both VOC- and spectral breathprint-based feature matrices, the same feature selection and classification techniques were used in developing the models. For feature

Table 4.1: Eight-bit uniform binary patterns, which have at most two bitwise transitions from 0 to 1 or vice versa.

Pattern		Pattern		Pattern	
Binary	**Decimal**	**Binary**	**Decimal**	**Binary**	**Decimal**
0000 0000	0	0011 1110	62	1100 1111	207
0000 0001	1	0011 1111	63	1101 1111	223
0000 0010	2	0100 0000	64	1110 0000	224
0000 0011	3	0110 0000	96	1110 0001	225
0000 0100	4	0111 0000	112	1110 0011	227
0000 0110	6	0111 1000	120	1110 0111	231
0000 0111	7	0111 1100	124	1110 1111	239
0000 1000	8	0111 1110	126	1111 0000	240
0000 1100	12	0111 1111	127	1111 0001	241
0000 1110	14	1000 0000	128	1111 0011	243
0000 1111	15	1000 0001	129	1111 0111	247
0001 0000	16	1000 0011	131	1111 1000	248
0001 1000	24	1000 0111	135	1111 1001	249
0001 1100	28	1000 1111	143	1111 1011	251
0001 1110	30	1001 1111	159	1111 1100	252
0001 1111	31	1011 1111	191	1111 1101	253
0010 0000	32	1100 0000	192	1111 1110	254
0011 0000	48	1100 0001	193	1111 1111	255
0011 1000	56	1100 0011	195	-	-
0011 1100	60	1100 0111	199	-	-

selection, the minimum redundancy maximum relevance (mRMR) algorithm was employed [60], where features are sequentially selected based on their correspondence to the class membership labels and distinction from other features. Specifically, in this work, the ranking criterion was based on the features' Pearson correlations with the class labels and more highly ranked features. Briefly, the mRMR selection procedure is as follows:

1. Calculate the relevance measure for each feature, which is the absolute value of the Pearson correlation between the feature and the class labels. Select the feature with the highest relevance to be the top ranked feature.

2. For each remaining (not yet ranked) feature, calculate the absolute value of

its correlation with the top ranked feature. This is the feature's redundancy measure.

3. Subtract each feature's redundancy measure from its relevance measure. Select the feature with the largest difference as the second highest ranked feature.

4. For each remaining feature, calculate the absolute value of its correlation with the second highest ranked feature. Take the mean of this value with the feature's previous redundancy measure from Step 2 to find its new redundancy measure.

5. As in Step 3, select the feature with the largest relevancy minus redundancy difference as the next ranked feature.

6. Continue recalculating the redundancy measures and selecting the top features until all features have been ranked.

The optimal number of selected features was then determined using classification performance. Beginning with a dataset consisting of only the top feature, the ranked features were added one at a time to the dataset and its performance was re-evaluated. To limit model complexity, only feature set sizes up to 79, half the total number of subjects, were assessed. The best performing feature set for a given set of samples was that which provided the best balance between sensitivity (true positive rate) and specificity (true negative rate), based on the point closest-to-(0,1) corner in a receiver operating characteristic (ROC) curve [61] for the samples' leave-one-out cross-validation (LOOCV) predictions. This distance D is given by (4.5).

$$D = \sqrt{(1 - specificity)^2 + (1 - sensitivity)^2} \tag{4.5}$$

Both intermediate and culminating classification models (those used for tuning feature number and those used for finding final performance estimates, respectively)

used a linear SVM algorithm for learning class associations. Using features from a labeled set of training samples, the SVM algorithm constructs a hyperplane in the feature space to best separate the two classes. The algorithm attempts to maximize the margin around the hyperplane (i.e., its distance from samples in either class) which then acts as a decision boundary for classifying future samples.

4.5 Performance Evaluation

Due to the limited sample size, a learning curve approach was used to provide a contextualized look at each model's classification ability. A learning curve depicts a model's classification performance at several incremental sample sizes, as it is able to increasingly learn the relevant patterns from the features. Starting with a subset of ten subjects, random progressive sampling was used to create a series of datasets ranging in size up to the maximum sample size (158 subjects) in increments of ten. Equal class sizes were maintained in each subset until sample sizes exceeded 124, where only control subjects remained to be added. For each sample size, the models were redeveloped and validated to obtain a series of empirical performance estimates. This was repeated over ten iterations for both the VOC- and breathprint-based models and the learning curves were averaged.

Because models trained with small sample sizes can be prone to overfitting, where noise in the data is learned in addition to or rather than relevant patterns, two different approaches were used for obtaining the performance estimates: 1) LOOCV and 2) nested LOOCV (see Figure 4.2). More details about model validation for studies with small sample sizes can be found in Appendix C. LOOCV is a data-efficient resampling method for estimating the generalization performance of a model. In short, one sample is used for testing the classifier while all others are considered training exemplars, used for learning classification parameters. This is repeated for multiple

'folds', with each sample used once as the test exemplar, and the final performance estimate is found by averaging the classifier's performance over all test subjects. In this way, the estimate reflects the classifier's efficacy for subjects unseen during training. The nested LOOCV framework takes this one step further by incorporating model development steps like feature ranking and feature number optimization into the cross-validation (CV) procedure as well, requiring that these steps use only training subjects. Unlike the traditional non-nested framework, where a single optimal feature set is found using all samples and then fixed during validation, the nested framework finds the optimal features for each intermediate training set. This eliminates the potential for information leakage between the training and test sets at every stage of the model's development.

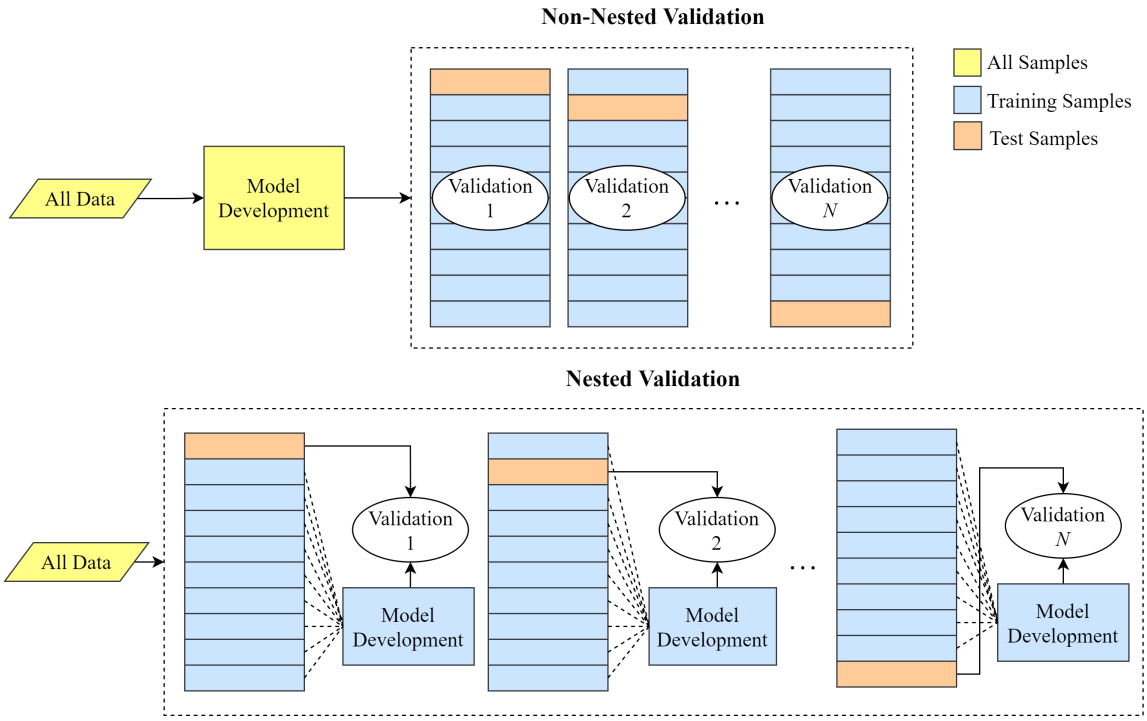

Figure 4.2: Non-nested and nested CV procedures. The non-nested framework uses all samples for model development (feature ranking and feature number optimization) while the nested framework uses only the training samples in each fold for model development.

Though neither framework is ideal alone, when combined, these non-nested and

nested CV approaches provide an estimate of the upper and lower bounds of a model's performance [62]. Overfitting is a problem with small sample sizes because strong noise-based differences between groups may appear by chance, and often perfect class separation may be achieved by exploiting these spurious patterns. High variance, overly complex models that do not generalize well to new samples are therefore a concern. With the non-nested approach, because model development steps include information from test subjects, this overfitting can inflate performance estimates for the test set. On the other hand, with the nested approach, overfitting to training samples will reduce performance on the test set, resulting in more conservative (even pessimistic) performance estimates. Hence, the two approaches are useful in tandem for observing the range of estimates.

4.6 Robust Biomarker Identification

For the final experimental aim, a robust consensus-based procedure was used to identify possible lung cancer biomarkers from the VOC and breathprint feature sets. Inspired by the consensus nested cross-validation (cnCV) method [63], the predictors that were selected most frequently during the learning curve procedure were considered 'consensus features'. In brief, the identification of consensus features, embedded into the learning curve performance estimation, was performed as follows:

1. Select a random subset of N samples from the dataset.

2. For the given subset, apply the CV and nested CV frameworks to compute optimal feature sets and assess their classification performance. For the non-nested framework, there will be a single feature set. For the nested framework, there will be N feature sets.

3. Identify the features that were selected in at least $T\%$ of the nested CV feature sets, creating a single feature set representing the consensus across CV folds.

For the given subset of samples, there will now be a single 'fold-consensus' nested CV feature set and a single non-nested CV feature set.

4. Repeat Step 2 and Step 3 for multiple sample sizes N by progressively adding random samples to the subset.

5. Repeat Steps 2, 3, and 4 for multiple iterations with different random subsets of samples.

6. For each sample size, identify the features that appeared in at least $T\%$ of iterations from the non-nested CV and fold-consensus nested CV feature sets. This will create two 'iteration-consensus' feature sets.

7. Identify the features from the iteration-consensus feature sets that were selected for at least $T\%$ of sample sizes to establish the final sets of potential biomarkers.

In this thesis, a threshold T of 68% was selected, as the empirical rule states that 68% of the data observed following a normal distribution lies within one standard deviation of the mean. Further, because high discord was expected for selected feature sets in the smallest tested sample sizes, sample sizes lower than 40 were disregarded when determining the sample size consensus in Step 7. This same procedure was used for both the VOC and spectral breathprint feature sets.

The potential biomarkers were lastly assessed through statistical means, to 1) compare each consensus feature between the lung cancer and control groups, and 2) examine each feature's association with confounding factors: sex, smoking habits, age, and CAT score. As in Section 4.1, due to the zero-inflated distributions for the log-transformed VOC features, the two-part testing procedure was again implemented for these comparisons. For the breathprint features, which were count variables, standard Wilcoxon rank sum tests were used.

Chapter 5

Results

5.1 Exploratory Data Analysis

Table 5.1 presents the results of the absorption coefficient comparisons between the lung cancer and control groups that were statistically significant at a level of $\alpha = 0.05$. The majority of significant differences were identified in the 75°C and 150°C spectra, which exhibited 28 and 24 significant lines, respectively. Notably, in all cases aside from the 10P12 and 10P14 $^{13}CO_2$ lines of the 300°C spectra, the indicated absorption coefficients were found to be increased in the lung cancer group.

Table 5.1: Absorption coefficients that exhibited significant differences between the lung cancer and control groups. Coefficients are labeled by their corresponding laser line and laser type (denoted 12 for the $^{12}CO_2$ laser or 13 for the $^{13}CO_2$ laser). The reported parameter μ_g is the geometric mean and σ_g is the geometric standard deviation.

Temp.	**Line ID**	**Wavelength** μm	**Control** $\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	**Lung Cancer** $\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
75°C	10R28 (12)	10.19	1.77 (1.82)	1.93 (1.54)	0.04
	10R26 (12)	10.21	1.79 (1.79)	1.95 (1.53)	0.03
	10R24 (12)	10.22	1.87 (1.98)	1.98 (1.53)	0.04
	10R22 (12)	10.23	1.90 (1.94)	2.02 (1.52)	0.03
	10R20 (12)	10.25	2.20 (2.14)	2.39 (1.68)	0.01
	10R18 (12)	10.26	1.89 (1.85)	2.10 (1.52)	0.007
	10R16 (12)	10.27	1.97 (1.85)	2.23 (1.51)	0.003
	10R14 (12)	10.29	2.44 (1.82)	2.74 (1.54)	0.01
	10R12 (12)	10.30	2.06 (1.87)	2.40 (1.53)	0.001
	10R10 (12)	10.32	2.05 (1.80)	2.42 (1.51)	0.001
	10R08 (12)	10.33	3.33 (1.87)	3.76 (1.64)	0.02
	10P08 (12)	10.48	1.71 (1.89)	2.18 (1.58)	0.0001
	10P10 (12)	10.49	1.66 (1.92)	2.14 (1.58)	< 0.0001
	10P12 (12)	10.51	1.64 (1.90)	2.11 (1.56)	0.0001
	10P14 (12)	10.53	1.66 (1.89)	2.08 (1.56)	0.0002
	10P16 (12)	10.55	1.65 (1.88)	2.05 (1.54)	0.0002
	10P18 (12)	10.57	1.56 (1.91)	1.97 (1.52)	0.0001
	10P20 (12)	10.59	1.59 (1.87)	1.96 (1.52)	0.0002
	10P22 (12)	10.61	1.63 (1.84)	1.98 (1.51)	0.0002
	10P24 (12)	10.63	1.66 (1.83)	1.98 (1.51)	0.0005
	10P26 (12)	10.65	1.73 (1.77)	2.03 (1.50)	0.001
	10P28 (12)	10.67	1.83 (1.73)	2.11 (1.49)	0.002
	10P30 (12)	10.70	1.94 (1.71)	2.17 (1.50)	0.01
	10R36 (13)	10.65	1.76 (1.78)	2.04 (1.50)	0.003
	10R34 (13)	10.67	1.87 (1.72)	2.19 (1.48)	0.001
	10R32 (13)	10.68	1.87 (1.69)	2.16 (1.48)	0.003
	10R30 (13)	10.69	1.97 (1.67)	2.21 (1.46)	0.01
	10R28 (13)	10.71	2.28 (1.64)	2.51 (1.47)	0.02
150°C	10R26 (12)	10.21	1.01 (1.58)	1.17 (1.65)	0.03
	10R22 (12)	10.23	1.04 (1.59)	1.19 (1.66)	0.03
	10R20 (12)	10.25	1.36 (1.74)	1.49 (1.59)	0.03
	10R18 (12)	10.26	1.02 (1.60)	1.20 (1.68)	0.02
	10R16 (12)	10.27	1.09 (1.59)	1.28 (1.64)	0.008

Table 5.1: Continued from previous page.

Temp.	**Line ID**	**Wavelength** μm	**Control** $\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	**Lung Cancer** $\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
	10R12 (12)	10.30	1.16 (1.60)	1.36 (1.63)	0.008
	10R10 (12)	10.32	1.17 (1.58)	1.41 (1.65)	0.004
	10P08 (12)	10.48	1.07 (1.58)	1.31 (1.69)	0.003
	10P10 (12)	10.49	1.02 (1.60)	1.27 (1.70)	0.003
	10P12 (12)	10.51	1.08 (1.59)	1.32 (1.66)	0.003
	10P14 (12)	10.53	1.12 (1.59)	1.34 (1.64)	0.004
	10P16 (12)	10.55	1.06 (1.59)	1.28 (1.62)	0.001
	10P18 (12)	10.57	0.97 (1.61)	1.20 (1.64)	0.001
	10P20 (12)	10.59	0.95 (1.62)	1.16 (1.64)	0.002
	10P22 (12)	10.61	0.95 (1.61)	1.16 (1.64)	0.002
	10P24 (12)	10.63	0.95 (1.63)	1.15 (1.64)	0.005
	10P26 (12)	10.65	1.00 (1.61)	1.20 (1.61)	0.003
	10P28 (12)	10.67	1.03 (1.59)	1.21 (1.62)	0.01
	10P30 (12)	10.70	1.08 (1.77)	1.32 (1.60)	0.006
	10R36 (13)	10.65	0.98 (1.62)	1.17 (1.64)	0.005
	10R34 (13)	10.67	1.17 (1.59)	1.35 (1.55)	0.004
	10R32 (13)	10.68	1.05 (1.60)	1.26 (1.59)	0.003
	10R30 (13)	10.69	1.15 (1.60)	1.36 (1.59)	0.004
	10R28 (13)	10.71	1.57 (1.61)	1.73 (1.50)	0.02
225°C	10P16 (12)	10.55	0.65 (1.91)	0.70 (1.65)	0.04
300°C	10P12 (13)	11.07	0.75 (2.48)	0.51 (2.18)	0.01
	10P14 (13)	11.09	0.78 (2.59)	0.52 (2.07)	0.03

Table 5.2: VOCs that exhibited significant differences in concentrations between the lung cancer and control groups. N_{SJ} denotes the number of subjects for which the concentrations were non-zero (i.e. the compound was present in the sample and able to be detected), and the reported geometric mean (μ_g) and geometric standard deviation (σ_g) values include only those non-zero concentrations.

Temp.	**VOC**	**Control**		**Lung Cancer**		p
		N_{SJ}	$\mu_g(\sigma_g)$ ppbV	N_{SJ}	$\mu_g(\sigma_g)$ ppbV	
75°C	3-Carene	0/96	0 (0)	5/62	1.54 (1.50)	0.02
	Cumene	16/96	2.83 (1.76)	5/62	4.43 (1.25)	0.02
	Dimethyl sulfide	3/96	0.76 (1.06)	10/62	1.74 (1.51)	0.002
	Isopropanol	33/96	2.18 (2.21)	36/62	2.57 (1.75)	0.01
	Perfluoroisobutylene	15/96	0.12 (2.25)	7/62	0.60 (2.00)	< 0.0001
150°C	Butyric acid	10/96	1.49 (2.36)	0/62	0 (0)	0.02
	D-Limonene	13/96	1.86 (2.09)	4/62	0.77 (1.20)	0.0006
	Menthol	0/96	0 (0)	5/62	0.98 (2.95)	0.02
	Methyl propyl ketone	3/96	1.50 (1.16)	6/62	7.41 (2.04)	0.03
300°C	Furfural	2/96	0.62 (1.54)	9/62	1.05 (1.92)	0.02

The results of the control and lung cancer VOC comparisons are shown in Table 5.2. Half of the significant results were consonant differences, where the group with the larger proportion of non-zero concentrations for the compound also had higher concentrations on average. Two compounds (cumene and perfluroisobutylene from the 75°C desorb) exhibited dissonant differences, with fewer instances but higher concentrations in the lung cancer group. Additionally, in three cases (3-carene from the 75°C desorb, butyric acid and menthol from the 150°C desorb), the compound only appeared in one group. For these compounds, the χ^2 test indicated significant differences in the frequency of the compound's presence.

Significant results from the absorption coefficient subgroup comparisons are presented in Table 5.3. Interestingly, all of the identified differences between subgroups were unanimous increases for one subgroup: 16 measurements were higher in females compared to males, 11 were higher in former smokers compared to never smokers, 19 were higher in smokers compared to never smokers, and 10 were higher in younger subjects compared to older subjects. No significant differences in absorption coeffi-

cients were found in association with CAT breathing symptom scores.

Table 5.3: Absorption coefficients that exhibited significant differences due to confounding factors. Coefficients are labeled by their corresponding laser line and laser type (denoted 12 for the $^{12}CO_2$ laser or 13 for the $^{13}CO_2$ laser). The reported parameter μ_g is the geometric mean and σ_g is the geometric standard deviation.

Temp.	**Line ID**	**Wavelength** μm	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
			Females	**Males**	
			$n = 82$	$n = 76$	
75°C	9P28 (12)	9.62	2.87 (1.87)	2.31 (1.89)	0.03
150°C	9R16 (12)	9.29	6.12 (1.66)	5.03 (1.64)	0.02
225°C	9P26 (12)	9.60	1.69 (1.94)	1.41 (1.66)	0.04
	9P28 (12)	9.62	1.84 (1.94)	1.53 (1.68)	0.03
	9P30 (12)	9.64	1.89 (1.99)	1.55 (1.74)	0.02
	10R34 (13)	10.67	0.85 (1.76)	0.72 (1.61)	0.04
	10P28 (13)	11.24	1.30 (2.22)	0.85 (3.18)	0.04
300°C	9P20 (12)	9.55	2.36 (1.80)	1.95 (1.52)	0.03
	9P22 (12)	9.57	1.99 (1.89)	1.60 (1.59)	0.03
	9P24 (12)	9.59	2.18 (1.92)	1.76 (1.65)	0.03
	9P26 (12)	9.60	2.45 (2.14)	1.85 (1.77)	0.03
	9P28 (12)	9.62	2.70 (2.12)	2.09 (1.82)	0.03
	9P30 (12)	9.64	2.96 (2.29)	2.22 (1.88)	0.03
	10P20 (12)	10.59	0.71 (2.40)	0.56 (1.95)	0.03
	10R34 (13)	10.67	0.97 (1.95)	0.78 (1.67)	0.02
	10P32 (13)	11.29	1.02 (2.90)	0.58 (2.35)	0.002
			Never-Smokers	**Former Smokers**	
			$n = 43$	$n = 91$	
75°C	10P08 (12)	10.48	1.71 (1.85)	1.98 (1.75)	0.04
	10P10 (12)	10.49	1.66 (1.89)	1.95 (1.76)	0.03
	10P12 (12)	10.51	1.67 (1.86)	1.92 (1.75)	0.04
	10P14 (12)	10.53	1.66 (1.86)	1.92 (1.74)	0.02
	10P16 (12)	10.55	1.66 (1.84)	1.89 (1.72)	0.04
	10P18 (12)	10.57	1.59 (1.83)	1.80 (1.70)	0.04
	10P22 (12)	10.61	1.63 (1.81)	1.83 (1.68)	0.04
150°C	10R08 (12)	10.33	3.63 (1.65)	4.44 (1.71)	0.02
	10P32 (12)	10.72	2.83 (1.65)	3.41 (1.69)	0.04
	10R22 (13)	10.75	3.02 (1.68)	3.64 (1.69)	0.03
	10R20 (13)	10.77	2.59 (1.68)	3.11 (1.63)	0.04
			Non-Smokers	**Smokers**	
			$n = 140$	$n = 18$	

Table 5.3: Continued from previous page.

Temp.	**Line ID**	**Wavelength** μm	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
150°C	10R32 (12)	10.17	1.01 (1.60)	1.34 (1.74)	0.03
	10R30 (12)	10.18	1.03 (1.59)	1.37 (1.72)	0.03
	10R28 (12)	10.19	1.01 (1.60)	1.35 (1.73)	0.02
	10R26 (12)	10.21	1.04 (1.59)	1.38 (1.72)	0.03
	10R24 (12)	10.22	1.04 (1.59)	1.36 (1.73)	0.03
	10R22 (12)	10.23	1.06 (1.60)	1.39 (1.73)	0.03
	10R18 (12)	10.26	1.05 (1.60)	1.44 (1.76)	0.02
	10R16 (12)	10.27	1.12 (1.59)	1.48 (1.71)	0.02
	10R12 (12)	10.30	1.20 (1.60)	1.54 (1.69)	0.04
	10R10 (12)	10.32	1.22 (1.60)	1.56 (1.67)	0.03
	10P18 (12)	10.57	1.02 (1.62)	1.33 (1.70)	0.03
	10P10 (13)	11.04	2.47 (1.45)	3.05 (1.74)	0.04
	10P14 (13)	11.09	1.70 (1.54)	2.16 (1.74)	0.04
	10P26 (13)	11.22	2.05 (1.56)	2.96 (1.87)	0.003
	10P32 (13)	11.29	1.55 (1.78)	2.87 (2.28)	0.0006
225°C	9P20 (12)	9.55	1.80 (1.70)	2.14 (1.62)	0.04
	10R18 (12)	10.26	0.65 (1.71)	0.83 (1.74)	0.03
	10P32 (13)	11.29	0.79 (2.43)	1.34 (2.67)	0.02
300°C	10P32 (13)	11.29	0.73 (2.76)	1.29 (2.33)	0.04
			Younger Subjects $n = 26$	**Older Subjects** $n = 132$	
75°C	9R24 (12)	9.25	5.27 (2.09)	3.92 (1.79)	0.03
	9R14 (12)	9.31	4.98 (2.00)	3.54 (1.76)	0.02
150°C	10P34 (13)	11.31	2.35 (1.86)	1.76 (1.92)	0.03
225°C	9P28 (12)	9.62	1.95 (1.90)	1.64 (1.81)	0.03
300°C	9P22 (12)	9.57	2.04 (1.58)	1.76 (1.80)	0.03
	9P24 (12)	9.59	2.34 (1.63)	1.91 (1.83)	0.02
	9P26 (12)	9.60	2.60 (1.72)	2.07 (2.03)	0.02
	9P28 (12)	9.62	3.01 (1.75)	2.28 (2.03)	0.01
	9P30 (12)	9.64	3.28 (1.83)	2.47 (2.16)	0.01
	10P28 (13)	11.24	1.30 (1.82)	0.94 (2.47)	0.03

The results of the subgroup comparisons of VOC concentrations are shown in Table 5.4. A few VOCs showed an association with multiple factors, such as 1,1,1-Trichloroethane (sex, age, CAT score) and cineole (smoking, age, CAT score). The differences were less uniform than those identified for the absorption coefficients, with varied increases and decreases in compounds between subgroups.

Table 5.4: VOCs that exhibited significant differences in concentrations due to confounding factors. N_{SJ} denotes the number of subjects for which the concentrations were non-zero (i.e. the compound was present in the sample and able to be detected), and the reported geometric mean (μ_g) and geometric standard deviation (σ_g) values include only those non-zero concentrations.

Temp.	**VOC**	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
			Females $n = 82$		**Males** $n = 76$	
75°C	1,1,1-Trichloroethane	17/82	1.14 (2.21)	12/76	0.53 (1.82)	0.02
	Benzaldehyde	0/82	0 (0)	6/76	9.68 (4.00)	0.03
	Methanol	19/82	0.36 (2.46)	4/76	0.34 (1.72)	0.01
	Propyleneimine	6/82	5.74 (1.77)	4/76	15.62 (1.46)	0.01
150°C	Ethene	37/82	0.04 (2.04)	32/76	0.03 (1.64)	0.04
	Hexafluoropropene	22/82	0.19 (2.25)	9/76	0.11 (1.96)	0.02
	Trifluoronitrosomethane	13/82	1.10 (2.76)	3/76	0.38 (1.14)	0.01
	Propylene sulfide	5/82	3.92 (1.76)	4/76	1.49 (1.54)	0.04
	m-Cresol	2/82	0.33 (1.09)	11/76	0.40 (1.70)	0.04
	n-Nonane	7/82	8.18 (1.30)	8/76	20.41 (1.94)	0.004
225°C	Thioglycol	11/82	19.40 (1.62)	11/76	11.10 (1.47)	0.02
300°C	2,3-Dimethylbutane	0/82	0 (0)	7/76	10.46 (1.57)	0.02
	3-Carene	5/82	7.29 (1.95)	4/76	2.67 (1.13)	0.01
	3-Methylfuran	29/82	1.93 (1.71)	25/76	1.16 (1.77)	0.004
	Benzene	56/82	3.76 (2.88)	42/76	2.33 (2.44)	0.02
	Methylamine	2/82	1.49 (1.07)	11/76	1.41 (1.54)	0.04
	tert-Butyl alcohol	7/82	0.40 (1.33)	0/76	0 (0)	0.03
			Never-Smokers $n = 43$		**Former Smokers** $n = 91$	
75°C	2-Methyl-1-propanal	6/43	2.38 (1.15)	1/91	1.85 (1.00)	0.007
	Hexafluoropropene	22/43	0.16 (1.78)	25/91	0.18 (2.63)	0.04
	Perfluoroisobutylene	9/43	0.11 (2.13)	12/91	0.33 (2.89)	0.02
150°C	Acetol	4/43	1.18 (1.28)	16/91	2.06 (1.94)	0.03
	Furan	5/43	0.76 (2.09)	2/91	0.31 (1.10)	0.04

Table 5.4: Continued from previous page.

Temp.	**VOC**	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
	Isobutene	7/43	1.21 (2.01)	6/91	0.52 (1.78)	0.03
	Perfluoroisobutylene	15/43	0.19 (2.14)	17/91	0.36 (3.23)	0.04
	tert-Butyl alcohol	8/43	0.33 (1.26)	21/91	0.48 (1.68)	0.03
	tert-Butyl methyl ether	9/43	1.44 (2.30)	13/91	4.43 (1.94)	0.005
225°C	Cineole	6/43	0.13 (1.56)	13/91	0.30 (1.80)	0.03
	Ethyl benzene	5/43	4.94 (1.12)	1/91	3.66 (1.00)	0.02
	Methanol	13/43	0.17 (1.46)	18/91	0.44 (2.64)	0.0005
300°C	Crotonaldehyde	28/43	0.52 (1.79)	35/91	0.50 (1.66)	0.02
	Octafluoropropane	12/43	9.51 (1.54)	23/91	5.54 (1.90)	0.02
		Non-Smokers $n = 140$		**Smokers** $n = 18$		
75°C	Propylene sulfide	28/140	1.00 (1.82)	4/18	3.56 (1.76)	0.001
150°C	Ethanol	4/140	0.86 (2.48)	4/18	3.34 (5.06)	0.006
225°C	Diethylether	3/140	5.63 (1.94)	3/18	2.09 (1.33)	0.03
	Isobutanol	0/140	0 (0)	2/18	0.23 (1.75)	0.004
300°C	Cumene	2/140	1.99 (1.02)	3/18	3.00 (1.59)	0.008
		Younger Subjects $n = 26$		**Older Subjects** $n = 132$		
75°C	1,1,1-Trichloroethane	4/26	2.31 (2.11)	25/132	0.71 (1.99)	0.03
	Ethyl butyrate	2/26	3.43 (1.81)	42/132	1.36 (2.04)	0.02
	Isopropanol	7/26	3.90 (1.80)	62/132	2.25 (1.96)	0.02
	Perfluoroisobutylene	4/26	0.06 (1.63)	18/132	0.26 (2.66)	0.0004
	o-Xylene	10/26	5.65 (4.37)	18/132	2.22 (2.22)	0.005
150°C	Ethene	5/26	0.06 (2.03)	64/132	0.03 (1.86)	0.01
	Methyl mercaptan	2/26	9.72 (1.05)	34/132	4.39 (1.59)	0.03
	Propyleneimine	3/26	10.16 (2.16)	1/132	9.44 (1.00)	0.01
	Octafluoropropane	5/26	9.31 (1.17)	24/132	5.99 (2.01)	0.03
225°C	2-Hexanol	6/26	1.87 (3.75)	8/132	1.10 (2.63)	0.03
	2-Methyl-1-pentene	2/26	1.80 (3.12)	0/132	0 (0)	0.02
	Cineole	4/26	0.13 (1.40)	16/132	0.26 (1.91)	0.03
300°C	Hexanal	5/26	2.75 (1.23)	1/132	9.44 (1.00)	< 0.0001
	Isoamyl alcohol	2/26	0.58 (1.52)	0/132	0 (0)	0.02
	tert-Butyl alcohol	4/26	0.35 (1.16)	3/132	0.47 (1.39)	0.03
		Low CAT Score $n = 83$		**High CAT Score** $n = 75$		
75°C	1,1,1-Trichloroethane	17/83	0.61 (2.07)	12/75	1.30 (2.02)	0.04
	2-Nonanone	9/83	9.78 (1.87)	9/75	4.53 (1.36)	0.008

Table 5.4: Continued from previous page.

Temp.	VOC	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	N_{SJ}	$\mu_g(\sigma_g)$ $1/cm \times 10^{-6}$	p
	Nitromethane	14/83	0.28 (2.38)	12/75	0.13 (1.48)	0.01
150°C	2-Methylfuran	8/83	2.11 (2.41)	0/75	0 (0)	0.02
	Diethylether	4/83	3.33 (1.56)	5/75	1.40 (1.57)	0.04
225°C	Acetic acid	5/83	1.53 (1.99)	10/75	0.50 (1.91)	0.01
	Cineole	5/83	0.18 (1.71)	15/75	0.25 (1.96)	0.04
300°C	Ethyl butyrate	4/83	0.87 (1.62)	12/75	1.59 (1.62)	0.02
	tert-Butyl methyl ether	5/83	1.30 (1.76)	18/75	1.80 (2.09)	0.007

5.2 Classification Models

The learning curves (based on sample sizes N=10, 20, ..., 150, 158) for the VOCs and the spectral breathprints are shown in Fig. 5.1. The VOC-based model performance estimated at the maximum sample size (N=158) was 85.44% accuracy (66.13% sensitivity, 97.92% specificity) using the non-nested framework, and 65.19% accuracy (51.61% sensitivity, 73.96% specificity) using the nested framework. The non-nested model was based on a fixed feature set of 19 VOC predictors while the nested models used an average of around 27 predictors (17-50) across CV folds. For the spectral breathprint-based models, the final performance estimates were 86.08% accuracy (82.26% sensitivity, 88.54% specificity) with the non-nested framework and 71.52% accuracy (58.06% sensitivity, 80.21% specificity) with the nested framework. The non-nested model used 24 breathprint predictors and the nested model used an average of around 28 (18-49) predictors. The receiver operating characteristic (ROC) curves for these final models are shown in Figure 5.2. It should be noted that the 95% confidence intervals naturally narrow as sample size increases as there is more overlap in subjects across the ten iterations. At the maximum sample size, only one 'subset' is possible and therefore only one iteration was performed.

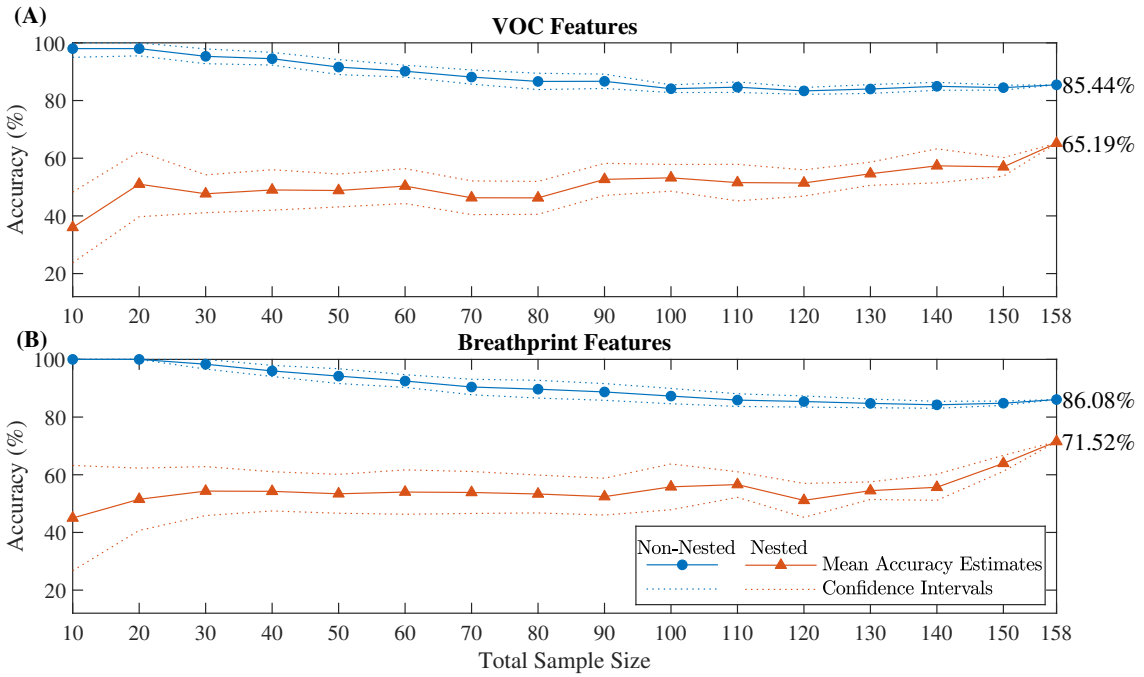

Figure 5.1: Average learning curves and their 95% confidence intervals from ten iterations of non-nested and nested CV estimates for models built using (A) log-transformed VOC concentrations and (B) spectral breathprint features.

The potential biomarkers identified during the consensus procedure for each feature type (VOC-based or spectral breathprint-based) and validation framework (non-nested or nested CV) are shown in Fig. 5.3. Using the VOC approach, the features from the non-nested CV procedure that fulfilled the consensus criterion were dimethyl sulfide (identified at a desorption temperature of 75°C), isopropanol (75°C), and butyric acid (150°C). The nested CV results corroborated both the dimethyl sulfide (75°C) and isopropanol (75°C) findings. For the breathprint approach, both nested and non-nested techniques yielded the same two consensus features: the frequency of LBP 31 (in binary, 00011111) in the raw 75°C spectra and the frequency of LBP 28 (00011100) in the raw 225°C spectra.

Table 5.5 presents the results of the comparisons between the lung cancer and control cohorts for each of these potential biomarkers. Two-part tests (based on a combination of a χ^2 test and either a t-test or Wilcoxon rank sum test, given the result of a Lilliefors test for normality) were used for the VOC features and stan-

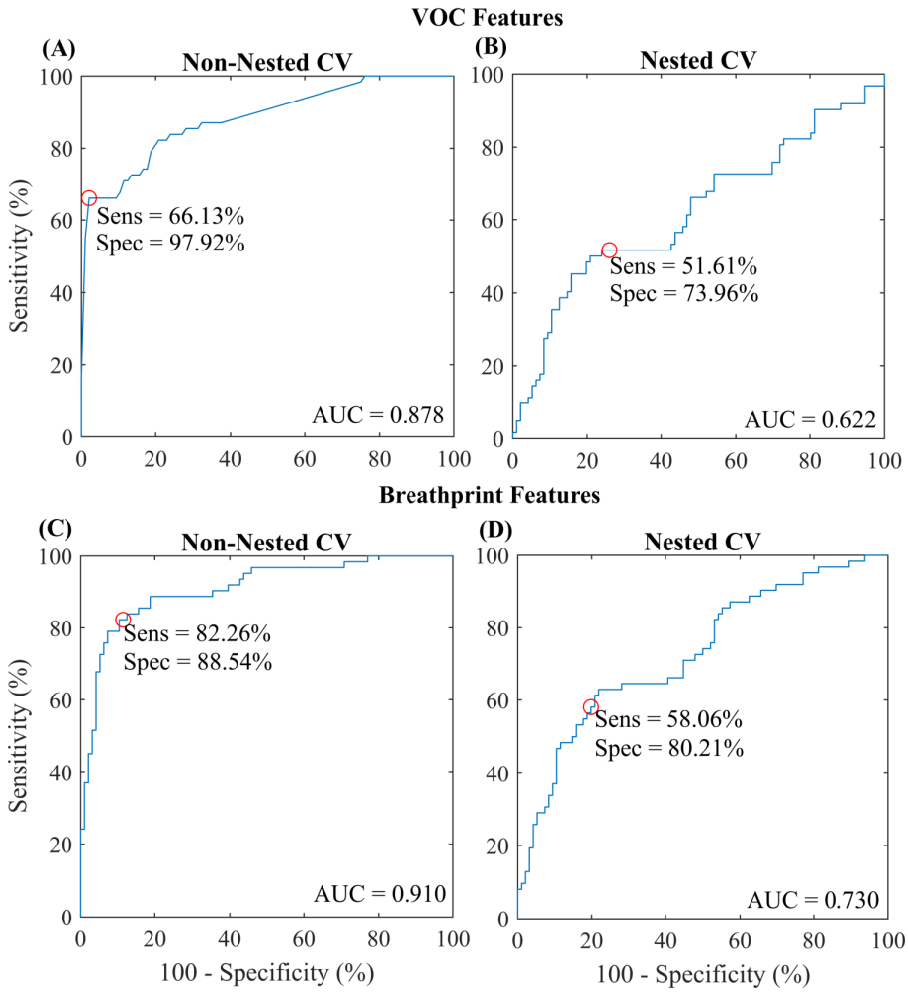

Figure 5.2: ROC curves for models developed using all 158 subjects: VOC-based models (A and B) and breathprint models (C and D) evaluated with non-nested CV (A and C) and nested CV (B and D).

dard Wilcoxon rank sum tests were used for the breathprint features. In each case, statistically significant differences were found ($p < 0.05$). For the VOC features, the two-part tests indicated consonant increases of dimethyl sulfide (75°C) and isopropanol (75°C) in the lung cancer group. Butyric acid (150°C), which was only present in the control group, was found to be significantly decreased in proportion for the lung cancer group. For the breathprint features, the frequency of raw LBP 31 (75°C) was found to be significantly increased and the frequency of raw LBP 28 (225°C) was found to be significantly decreased in the lung cancer cohort.

Comparisons were also performed for various subgroups in Table 5.6 to test

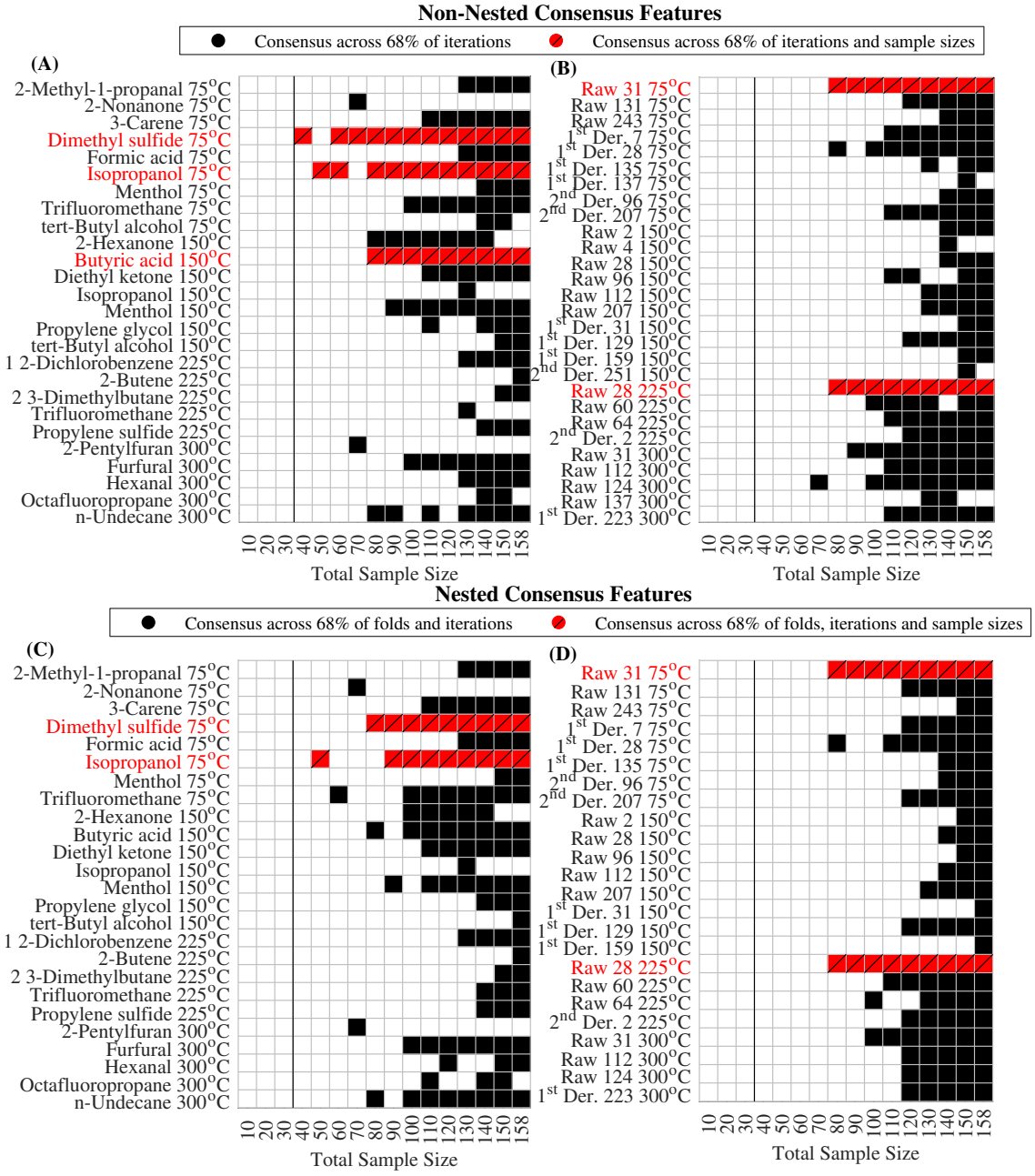

Figure 5.3: Consensus VOC and breathprint features identified for (A, B) the non-nested and (C, D) the nested CV learning curves.

each feature's association with possible confounders. Few significant differences were found between subgroups. A dissonant difference was found for isopropanol (75°C) due to age, with an increase in instances but lower concentrations in the older group compared to the younger group ($p = 0.04$). For the spectral breathprint features,

the raw LBP 28 (225°C) feature was found to be decreased ($p = 0.003$) in former smokers compared to never-smokers as well as in subjects with high CAT scores compared to low CAT scores ($p = 0.02$), and the frequency of raw LBP 31 (75°C) was increased for active smokers compared to non-smokers ($p = 0.01$).

Table 5.5: Tests of significance for the VOC features (the logarithm of the linear fitting coefficients) and spectral features (the LBP frequencies) identified through the 68% consensus procedure. Note: N_{SJ} represents the number of subjects with a non-zero value for that feature, and the mean (μ) and standard deviation (σ) include only those non-zero values. Features with zero values represent the absence of a compound in a spectrum (for the VOC approach) or the absence of a specific LBP in a spectrum or spectral derivative (for the breathprint approach).

Feature	**Control**		**Lung Cancer**		***p*-value**
	N_{SJ}	$\mu \pm \sigma$	N_{SJ}	$\mu \pm \sigma$	
Dimethyl Sulfide 75°C	3/96	4.33 ± 0.06	10/62	5.16 ± 0.41	0.003^{*}
Isopropanol 75° C	33/96	5.39 ± 0.79	36/62	5.55 ± 0.56	0.01^{*}
Butyric Acid 150°C	10/96	5.01 ± 0.86	0/62	0.00	0.02^{*}
Raw LBP 31 75°C	74/96	1.77 ± 0.99	58/62	2.21 ± 1.16	$0.0004^{\dagger}$
Raw LBP 28 225°C	15/96	1.17 ± 0.25	1/62	1.00 ± 0.00	$0.005^{\dagger}$

* Two-part test indicated a significant difference between groups ($p < 0.05$)
† Wilcoxon rank sum test indicated a significant difference between groups ($p < 0.05$)

Table 5.6: Subgroup tests of significance for each of the potential VOC biomarkers (the logarithm of the linear fitting coefficients) and breathprint biomarkers (the LBP frequencies). Note: N_{SJ} represents the number of subjects with a non-zero value for the feature, and the reported mean (μ) and standard deviation (σ) include only those non-zero values.

Feature	N_{SJ}	$\mu \pm \sigma$	N_{SJ}	$\mu \pm \sigma$	***p*-value**
	Females		**Males**		
Dimethyl Sulfide 75°C	10/82	4.97 ± 0.58	3/76	4.97 ± 0.38	0.28
Isopropanol 75°C	36/82	5.56 ± 0.66	33/76	5.38 ± 0.72	0.59
Butyric Acid 150°C	4/82	4.34 ± 0.74	6/76	5.45 ± 0.74	0.25
Raw LBP 31 75°C	70/82	2.03 ± 1.09	62/76	1.89 ± 1.10	0.28
Raw LBP 28 225°C	9/82	1.11 ± 0.33	7/76	1.00 ± 0.00	0.70
	Never-Smokers		**Former Smokers**		
Dimethyl Sulfide 75°C	2/43	4.35 ± 0.10	7/91	5.07 ± 0.58	0.40
Isopropanol 75°C	14/43	5.57 ± 0.78	45/91	5.48 ± 0.72	0.26
Butyric Acid 150°C	2/43	4.35 ± 1.15	8/91	5.17 ± 0.84	0.58
Raw LBP 31 75°C	33/43	1.88 ± 0.96	76/91	1.93 ± 1.12	0.52
Raw LBP 28 225°C	9/43	1.00 ± 0.00	4/91	1.25 ± 0.50	$0.003^{\dagger}$

	Never or Former Smokers		**Active Smokers**		
Dimethyl Sulfide 75°C	9/140	4.91 ± 0.59	4/18	5.12 ± 0.34	0.16
Isopropanol 75°C	60/140	5.49 ± 0.72	9/18	5.34 ± 0.45	0.70
Butyric Acid 150°C	10/140	5.01 ± 0.90	0/18	0.00	0.51
Raw LBP 31 75°C	114/140	1.90 ± 1.06	18/18	2.33 ± 1.24	$0.01^{\dagger}$
Raw LBP 28 225°C	14/140	1.07 ± 0.27	2/18	1.00 ± 0.00	0.90
	Younger Subjects (<55)		**Older Subjects (≥ 55)**		
Dimethyl Sulfide 75°C	1/26	4.42 ± 0.00	12/132	5.02 ± 0.52	0.71
Isopropanol 75°C	7/26	5.97 ± 0.64	62/132	5.42 ± 0.68	0.04^{*}
Butyric Acid 150°C	1/26	6.23 ± 0.00	9/132	4.87 ± 0.84	0.38
Raw LBP 31 75°C	20/26	2.10 ± 1.33	112/132	1.94 ± 1.05	0.63
Raw LBP 28 225°C	2/26	1.00 ± 0.00	14/132	1.07 ± 0.27	0.65
	Low CAT Score (<6)		**High CAT Score (≥ 6)**		
Dimethyl Sulfide 75°C	4/83	4.61 ± 0.53	9/75	5.13 ± 0.46	0.14
Isopropanol 75°C	30/83	5.40 ± 0.77	39/75	5.53 ± 0.63	0.13
Butyric Acid 150°C	8/83	5.10 ± 0.96	2/75	4.65 ± 0.73	0.27
Raw LBP 31 75°C	65/83	1.94 ± 1.03	67/75	1.99 ± 1.16	0.21
Raw LBP 28 225°C	13/83	1.08 ± 0.28	3/75	1.00 ± 0.00	$0.02^{\dagger}$

* Two-part test indicated a significant difference between groups ($p < 0.05$)

† Wilcoxon rank sum test indicated a significant difference between groups ($p < 0.05$)

Chapter 6

Discussion

6.1 Lung Cancer Screening Using CRDS

The first goal of this thesis was to determine the utility of CRDS, an LAS technique, for detecting lung cancer through the analysis of exhaled breath profiles. The non-nested and nested CV accuracy estimates were 85.44% and 65.19% for the VOC features and 86.08% and 71.52% for the breathprint features, respectively. The disparity in non-nested and nested estimates can be attributed mainly to overfitting effects, although some pessimism is inherent for the nested estimates due to the reduced information during feature selection [64]. By using all samples in model development, however, as with the non-nested framework, the models were able to find the most convenient noise-based features in addition to truly relevant patterns. This was especially evident in the smallest sample sizes ($N < 40$) where perfect or near-perfect accuracies were achieved with this method. Eventually, with sufficient sample size, the two types of estimates should converge to a value within the indicated performance range (Fig. 5.1), reflecting the true underlying quality of the features.

Based on the performance estimate ranges for the current classification models, the discrimination ability of CRDS breath profiles is at least on par with many other

screening technologies. In a large population screening study, chest CT detected lung cancer with 55% sensitivity and 95% specificity [65]. For breath-based systems, studies employing a variety of spectrometry and e-nose technologies from 1999 to 2019 have reported accuracies ranging from 60% to 100% (sensitivities from 50% to 100% and specificities from 12% to 100%) in discriminating lung cancer subjects from non-cancer controls [7, 47]. It is important, however, to interpret and compare these reported classification performance estimates with caution. In addition to study designs, sample collection protocols, and measurement techniques, which vary greatly across studies, the employed statistical and machine learning schemes are critical to understanding the performance estimates. In fact, many studies reporting very high accuracies employed non-nested CV or other validation frameworks that can yield overly-optimistic estimates [7]. Lack of correction for overoptimism is, indeed, a key methodological shortcoming in many breath-based lung cancer detection studies. For instance, Westhoff et al. [23] reported perfect LOOCV accuracy in distinguishing 32 lung cancer patients from 54 healthy controls using ion mobility spectroscopy (IMS) peaks, but they used information from both training and test subjects for tuning the optimal feature set. Bajtarevic et al. [29] similarly used all available subjects for selecting VOC predictors to develop their decision rule, which achieved 80% sensitivity and 100% specificity for a dataset of 65 lung cancer and 31 control subjects. Contrarily, few studies employ nested CV or other techniques that prevent information leakage, such as hold-out validation. This was the case for Mazzone et al. [27], who used an independent 30% subset of their 49 NSCLC and 94 non-cancer control subjects for validation, achieving 73.3% sensitivity and 72.4% specificity with colorimetric sensor array measurements. Also considering sample size and model complexity (associated with the adopted classification parameters and the number of selected features), which can exacerbate the contrast between estimates from different validation frameworks, direct comparisons across reported

performance estimates are not possible.

Despite the abundance of research in exhaled breath analysis for lung cancer detection using spectroscopic and sensor-based techniques, few studies have used LAS for this purpose. Skeldon et al. [66] used tunable diode laser absorption spectroscopy (TDLAS) for measuring ethane, an accepted marker of oxidative stress, in the exhaled breath of 12 lung cancer patients and 12 matched controls. Notably, they did not find a significant difference between the two cohorts, concluding that a singular non-specific marker such as ethane is unlikely to provide sufficient evidence of a particular pathological condition such as lung cancer. Mitrayana et al. [67] used laser photoacoustic spectroscopy (LPAS) for comparing acetone in the breath of 11 lung cancer patients and two control populations, consisting of 10 healthy volunteers and 9 patients with other lung diseases. They found a significant increase in acetone for lung cancer patients compared to the reference groups, but recommended further analysis because measurements for all populations fell within the expected concentration range of acetone in normal breath [68]. Both studies were limited by very low sample sizes and targeted approaches: since biomarkers may be associated with many diseases, and moreover, a disease may be characterized by several biomarkers, an individual's entire composite breath profile may be necessary for lung cancer detection.

The CRDS system presented in this thesis is novel to the field of lung cancer detection. As a LAS technique, it provides low-cost, quick, and accurate breath profiling that can be performed by non-experts. However, unlike other LAS techniques like TDLAS and LPAS, CRDS is essentially calibration-free and is immune to fluctuations in laser intensity [69]. Further, due to the considerable path length in the cavity (approximately one kilometer in the current study), CRDS affords higher sensitivity than other forms of LAS [70]. For a breath sample, CRDS therefore provides an ultra-sensitive, highly reproducible set of measurements. Moreover,

unlike Skeldon et al. [66] and Mitrayana et al. [67], the CRDS system in this work measured absorptions for a wide range of infrared wavelengths. By broadening the investigation to include the entire VOC composition of a sample, rather than targeting individual presumed biomarkers, the present CRDS analysis was able to better capture distinguishing signals.

6.2 VOC Concentrations vs. Spectral Breathprints

The second goal of this thesis was to compare models trained with VOC features to models trained with spectral breathprint features. While the non-nested CV performance estimates were very similar for both sets of features (within 3% difference in mean accuracy at all sample sizes), the more conservative estimates from the nested framework were generally higher for the breathprint features (Fig. 5.1). In fact, using all 158 samples, an improvement in nested LOOCV accuracy of over 6% was observed compared to the VOC-based model.

The conventional VOC-based approach to classification is complementary to spectrometry techniques that can quantify VOCs in a sample with high accuracy (such as GC-MS). Through statistical and machine learning methods, the most relevant VOCs are identified as lung cancer biomarkers and used as predictors for detecting the disease in future samples. The main advantage of this approach is its interpretability, and it permits investigation into the metabolic processes that produced the specific VOC biomarkers. Further, if definite VOC biomarkers were identified, then they would be able to translate to several different platforms. However, as the search for these biomarkers in recent years has yet to yield any VOCs of clinical relevance, the breathprinting approach is a promising alternative.

The breathprinting technique, most often associated with cross-reactive sensor arrays that cannot discern specific VOC constituents, aims to identify patterns asso-

ciated with disease from the sensor response. The flexibility of this approach offers some advantages over standard VOC identification. It is not reasonable to expect a homogeneous breath profile across all individuals with lung cancer; in addition to environmental confounders and individual-specific differences, different lung cancer cell mutations will ensure that all breath profiles are unique to a degree [71]. As the complex relationships, origins and metabolic pathways for VOCs in exhaled breath are still not well understood [72], fixating on specific VOCs may be a limiting approach to detecting disease. With breathprints, however, machine learning techniques can be harnessed to uncover subtle differences in breath profiles and learn different manifestations of the lung cancer. Advanced feature extraction and transformation techniques may be able to uncover complex patterns that the VOC identification approach cannot. Additionally, classification does not require knowledge of the specific VOCs in a sample.

Based on the improved performance observed with the spectral breathprinting technique, this approach for lung cancer detection is recommended. Though more abstract than the VOC features, the histogram features from the 1D-LBP representations of the spectra and their derivatives were able to more effectively characterize the health of the subjects than the VOC concentrations. Further, because both sets of features are different representations of the same original data, and fitting VOCs to the spectra is not a trivial undertaking, it is computationally expedient to bypass the VOC identification step and focus on drawing out the most useful patterns from the breathprints themselves. An advantage of CRDS is that it permits VOC extraction when necessary, so further analysis and interpretation can be performed in addition to the breathprinting approach to examine the samples from a biological perspective. With non-selective e-nose technologies, this type of investigation would require a secondary technology such as GC-MS [26].

6.3 Robust Biomarker Identification

The third goal of this thesis was to propose a robust breath biomarker identification method for VOCs and spectral breathprints. Based on the proposed method, there were two VOCs that were identified by the consensus criteria as lung cancer breath biomarkers from both nested and non-nested CV learning curves: (1) isopropanol and (2) dimethyl sulfide (Fig. 5.3).

Isopropanol (also called isopropyl alcohol or 2-propanol) is a common ingredient in regular household items such as disinfectants and hand sanitizers. The results of the present investigation are in support of previous studies [12, 24, 73–75] which suggest that exhaled isopropanol concentrations are significantly higher ($p < 0.05$) in patients with lung cancer than in healthy controls (Table 5.5). While some suggest that this may be caused by the use of disinfectants in hospital rooms [76], isopropanol has been repeatedly identified as a potential lung cancer biomarker in previous investigations [12, 24, 26, 73–75]. Among these, the impact of isopropanol on discrimination has been identified as very high. For example, using SPME for pre-concentration and gas chromatography-time of flight-mass spectrometry (GC-TOF/MS), isopropanol had the highest discriminant ability among 20 potential VOC breath biomarkers, determined by LDA [73]. Using similar SPME-GC-MS technology, but different multivariate analysis (i.e., PCA), the first three principal components (PC1-PC3) showed significant differences between 31 lung cancer patients, 31 smokers, and 31 healthy controls, and isopropanol and 1-propanol were the most positively correlated substances on PC3 [74]. With a combination of the predictors isopropanol, formaldehyde, and age for QDA classification, an accuracy of 96% (sensitivity 54%, specificity 99%) was achieved in distinguishing 17 lung cancer patients and 170 healthy controls using proton transfer reaction mass-spectrometry (PRT-MS) [24]. Review studies [8,9] also determined propanol as the most frequently emerging biomarker of lung cancer, and in the human body, propanol is believed to

be mostly isopropanol (or 2-propanol) [68].

Dimethyl sulfide (DMS) in breath is most often associated with halitosis [77]. The results of the present investigation indicate an increase in dimethyl sulfide ($p < 0.05$) in the exhaled breath of lung cancer patients as compared to healthy controls (Table 5.5), which is in support of previous findings [75, 76]. While Kischkel et al. [74] reported that the concentration of dimethyl sulfide was lowest in lung cancer patients, they posited that their finding may be related to dental status rather than to cancer specific effects. Despite this, and similar to isopropanol, dimethyl sulfide has been identified as a key VOC breath biomarker for discrimination between lung cancer patients and healthy controls using decision tree classification [75].

There are some interesting considerations regarding isopropanol and dimethyl sulfide. First, both compounds have been identified as breath biomarkers for patients with cystic fibrosis, a progressive, serious genetic disease that causes breathing abnormalities including lung infections, a persistent cough, and shortness of breath [68,78]. Second, both compounds have shown some level of association with smoking. A previous work found that dimethyl sulfide was one of the most important compounds for discriminating healthy subjects with different smoking habits [76]. Additionally, in the breath of active smokers, one study [76] found that the concentration of isopropanol was higher while another [74] found that the concentration was lower when compared to non-smoking controls and lung cancer patients. However, it should be noted that neither isopropanol nor dimethyl sulfide were found to differ significantly due to smoking habits in this work, and neither have emerged as established smoking-related VOCs in the literature [8].

From the non-nested CV learning curve, one additional VOC biomarker was identified, butyric acid, which was found in the non-cancer control subjects but not in lung cancer patients (Table 5.5). To our knowledge, this VOC has not been previously identified as a potential lung cancer biomarker [8, 9]. It has, however,

been found to be decreased in oesophagogastric cancer patients compared to healthy subjects [79]. It is interesting to note that this fatty acid can also increase in breath after ingestion of a meal in healthy subjects [80]. In the present investigation, butyric acid was found only in subjects that had eaten less than 14 hours prior to data collection (most more recently than 6 hours). The finding may therefore be at least partially related to food intake in healthy subjects. It should also be reiterated that butyric acid was not identified as a biomarker using the proposed consensus method with the nested CV framework. It is possible that the stricter protocol for nested CV, which imposes a 68% agreement across CV folds in addition to the consensus across iterations and sample sizes, was able to filter out butyric acid as a noise-based feature. The nested CV protocol is especially effective at minimizing the effects of outlier subjects: after all, butyric acid was only found in a total of ten subjects.

In addition to smoking history and food/diet, age and gender have been recognized as important confounding factors in previous studies on the concentration variation of VOCs [8, 81]. In this work, no differences were found between females and males for any of the identified consensus biomarkers, and only isopropanol was found to show some association with age, as shown in Table 5.6. Although the subgroup analysis did not have strong evidence to indicate that confounders had a significant impact on classification, it is possible that systemic differences may have biased performance due to unmatched factors for the cancer and control cohorts. Importantly, some mismatch was tolerated to mitigate the risk of undiagnosed lung cancer in the control group. Some factors, including comorbidities, diet, and medication use, further, were intentionally unconstrained for the subjects to ensure a realistic representation of the lung cancer population. Regardless, confounding factors (e.g. smoking, age/gender, food/diet, alcohol, medication/drugs, exercise, and other diseases) are a complex issue that should be incorporated in future studies to gain a greater understanding of their effects on VOC concentrations and spectral

breathprints.

For the spectral breathprints (VOC patterns), two consensus features were identified (Table 5.5). Both are features based on the 1D-LBP of raw spectra. This feature extraction method has shown its discriminative power in many biomedical signal and image processing applications with the ability to analyze data in real-time applications (due to its computational simplicity) [58]. To the best of the author's knowledge, however, this study is the first to apply 1D-LBP for extracting useful information from breathprints.

Ultimately, the present learning curves (Fig. 5.1) suggest that more samples are needed to fully exploit the potential of this approach, as the curves do not appear to have reached convergence by 158 samples. Due to the high complexity and variability associated with exhaled breath, a large sample size is necessary to represent the wide array of profiles that would be encountered during screening. The addition of samples would also permit the implementation of more advanced machine learning techniques, which should further improve classification performance. Many feature extraction methods could be potentially useful for breathomics analysis, such as PCA and barcoding [82, 83]. While the present investigation was restricted to the 1D-LBP features to limit overfitting to the available samples, future research should consider exploring the discrimination performance of additional feature extraction methods. Similarly, deep learning algorithms can take advantage of larger sample sizes to learn entirely new feature mappings.

Moreover, although the present investigation employed a library of 152 common VOCs in exhaled breath, it has been acknowledged that more than 3,000 different VOCs can be observed in human breath samples [25]. Future studies may therefore incorporate a larger spectral library. Further, the spectral regression procedure is limited by the number of measured wavelengths: with 73 measurements, a maximum of 73 VOCs can be fitted. Ongoing hardware advances may soon expand the range

and resolution of the CRDS spectra, enabling the VOC information to be expanded from the source data.

Chapter 7

Conclusions

7.1 Summary

The overarching purpose of this thesis was to develop an effective and practical exhaled breath screening tool for lung cancer. To this end, the three main components of this work were 1) evaluating the efficacy of infrared CRDS breath profiles for characterizing the lung cancer's signature, 2) comparing the performance of traditional VOC-based classification models to those based on patterns in the spectra, and lastly, 3) presenting a consensus-based approach for identifying relevant biomarkers.

Chapters 1 and 2 demonstrated the need for effective lung cancer screening, the basis for breath analysis, and the work that has been done to date in detecting lung cancer through exhaled breath. As lung cancer is most often detected too late for effective treatment, and current LDCT screening is prohibitively costly and has a tendency for overdiagnosis, a considerable number of works have explored solutions based on breath VOCs. The endogenous VOCs of interest, associated with metabolic changes accompanying lung cancer, have been investigated primarily using mass spectroscopy and e-nose sensor array technologies to date. Limitations in these technologies were briefly discussed, such as the complex and labor-intensive

procedures required for mass spectroscopy analysis and the vulnerability to temperature and humidity that burdens many e-nose systems. Methodological limitations that have lead to inconsistent results across studies, including small sample sizes and a lack of emphasis on proper model and biomarker validation, were also examined. Hence, this thesis focused on practical, reliable CRDS technology and robust data analysis strategies to address these limitations.

In Chapter 3, the CRDS technology was introduced and the study design and sample collection protocols were described. A multi-temperature, two-laser procedure was detailed for obtaining an information dense dataset of 292 CRDS absorption coefficients in the mid-infrared region for each sample. The sample size requirements were estimated prior to sample collection using a classification learning curve for a simulated dataset. Given optimistic conditions, where 10% of simulated features exhibited a large difference ($d = 0.8$) between classes, at least 200 samples (100 per class) were found to be necessary to consistently detect features of interest and converge to the true accuracy of the problem. Therefore, 100 lung cancer patients and 98 controls were enrolled in the study for sample collection, though 62 pre-treatment NSCLC patients and 96 controls were included in the analysis after exclusions. In a comparison of the two cohorts, significant differences were found in age and breath symptom scores, with higher values for lung cancer group on average, as well as in smoking and other lung conditions, with higher instances in the lung cancer group on average.

Chapters 4, 5 and 6 detailed the methodology, results and discussion, respectively, for the main experimental aims of the thesis. Two contrasting approaches for modeling the lung cancer signature were considered: 1) using VOC predictors, similar to mass spectroscopy studies, and 2) using patterns from spectral breathprints, similar to e-nose studies. First, a preliminary exploratory statistical analysis was performed, in which raw VOC concentrations and absorption coefficients were

compared between the lung cancer and control cohorts and between various subgroups. Although many variables showed significant associations with factors like age and smoking, reinforcing what was already known about confounders from previous works, there were several significant differences between the lung cancer patients and controls, indicating potential for the screening model.

For developing the classification models, these raw VOC and absorption coefficient variables were transformed into meaningful features. While the VOC features were simply log-transformed fitting coefficients from a stepwise spectral fitting algorithm, the breathprint features were based on scale-invariant 1D-LBP representations of the raw, first derivative and second derivative spectra. Using mRMR for feature selection and linear SVM learning algorithm for classification, thorough evaluation was performed through a learning curve approach based on two forms of cross-validation. These two cross-validation techniques, non-nested LOOCV and nested LOOCV, provided a demonstrative range of performance estimates at each learning curve sample size. The most relevant features were identified during the learning curve procedure through consensus in selected features across cross-validation folds, iterations, and sample sizes.

From the learning curve results, the classification models developed in this work were shown to be on par with corresponding mass spectroscopy and e-nose systems in discrimination ability. With all 158 samples, the accuracy estimates for the VOC-based model were $65.19\% - 85.44\%$ and the estimates for the breathprint-based model were $71.52\% - 86.08\%$. Hence, this thesis realized its first objective in demonstrating the feasibility of the novel CRDS breath profiling system for discriminating non-small cell lung cancer patients from non-cancer controls.

For the second aim, the spectral breathprinting approach for classification was endorsed as it provided improved discrimination ability over the traditional VOC-based approach (over 6% using nested LOOCV). With more samples, the flexibility

of the breathprinting approach also permits further improvement in classification performance through additional feature extraction, which can augment the information from the breath profiles to better capture underlying disease signatures. As CRDS technology permits the quantification of VOCs in a sample through spectral regression, a supplemental VOC-based analysis can be performed alongside the breathprint classification for interpretation when desired.

Lastly, for the third aim, through the proposed consensus-based biomarker identification procedure, isopropanol, dimethyl sulfide, and butyric acid were found as potential VOC lung cancer biomarkers along with two 1D-LBP spectral breathprint biomarkers. This rigorous process for feature assessment may aid in filtering out noise-based features and reduce the disagreement in identified biomarkers across studies.

Overall, the objectives set out and achieved in this thesis serve as an early-stage validation of infrared CRDS combined with machine learning for lung cancer screening.

7.2 Limitations and Future Work

As machine learning models depend so heavily on the quality of the data used in their development, future works should primarily aim to improve study design. In this work, the screening system was designed for the simplified case in which all subjects fall cleanly into one of two classes: NSCLC patients or non-cancer controls. Further, although many of the controls presented with other lung conditions, such as COPD, the cohort included several healthy, non-symptomatic individuals that would likely not be the focus of a screening program. With a real-world, unconstrained screening population, which would include individuals with various other conditions and cancers, the identified biomarkers may not offer the same level of discriminability.

Future works should therefore aim to recruit mainly high-risk individuals (long-time smokers above 55, for instance), including many with conditions that present similarly to lung cancer. Additionally, rather than the binary lung cancer/no lung cancer problem considered in this work, the detection framework could be reframed as a regression model that yields predictions about the disease's progression, which would account for potential changes in the disease signature across stages.

Given the complexity of the breath profiles and the results observed from the learning curve procedure, it was noted that the sample size used in this pilot work was insufficient to fully capture the relevant lung cancer patterns. Likewise, due to differences observed between the two cohorts in factors like smoking and age, it was suggested that systemic differences could have biased the developed classification models. Hence, future works should consider larger sample sizes and take care to limit or control for confounding effects during classification as much as possible through subject matching. Post-collection, subgroup-specific models (for example, one model for smokers, one for non-smokers) may be a similarly useful tool for addressing confounders when sample size permits. To mitigate the influence of environmental VOCs, future works should also consider comparing breath spectra to spectra for the surrounding room air. This could help to enhance the endogenous compounds of interest and reduce irrelevant differences related to collection site.

Regarding data analysis, there are multiple techniques for pre-processing and feature engineering that can be further explored, especially with a larger sample size. For the VOC-based approach, this may include alternative transformations for the concentrations, such as n^{th} root scaling for example. As previously noted, the VOC fitting procedure may also be improved with a more comprehensive spectral library or through modifications to the stepwise algorithm (by incorporating prior probabilistic knowledge, for instance). With the breathprinting approach, future works should explore more shape-emphasizing feature extraction techniques such as

the aforementioned barcoding method [83] as well as deep learning features. Combinations of pre-processing techniques (such as scaling and detrending) and feature extraction techniques should also be considered for optimizing performance.

Additionally, for both VOC and breathprint approaches, alternative feature selection and classification techniques may provide better results. For feature selection, this may include metaheuristics like genetic algorithms and particle swarm optimization, or stepwise selection algorithms like sequential floating forward selection. For classification, in addition to deep learning techniques, learning algorithms designed for sparse or missing data may improve performance and, for models based on absorption coefficients, may also eliminate the need for imputation.

Bibliography

[1] F. Bray *et al.*, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, pp. 394–424, 2018.

[2] N. Howlader, A. Noone, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. Cronin, "SEER cancer statistics review, 1975-2016," Bethesda, MD, 2018.

[3] E. Patz Jr, P. Pinsky, C. Gatsonis, J. Sicks, B. Kramer, M. Tammemägi, C. Chiles, W. Black, and D. Aberle, "Overdiagnosis in low-dose computed tomography screening for lung cancer," *JAMA Intern Med*, vol. 174, no. 2, pp. 269–274, 2014.

[4] W. Li, H. Y. Liu, Z. R. Jia, P. P. Qiao, X. T. Pi, J. Chen, and L. H. Deng, "Advances in the early detection of lung cancer using analysis of volatile organic compounds: From imaging to sensors," *Asian Pac J Cancer Prev*, vol. 15, no. 11, pp. 4377–4384, 2014.

[5] T. Paff, M. P. van der Schee, P. Brinkman, W. M. van Aalderen, E. G. Haarman, and P. J. Sterk, "Breathomics in Lung Disease," *Chest*, vol. 147, no. 1, pp. 224–231, 2015.

[6] G. Pennazza and M. Santonico, *Breath Analysis*. Elsevier Science, 2018.

[7] A. Krilaviciute, J. A. Heiss, M. Leja, J. Kupcinskas, H. Haick, and H. Brenner, "Detection of cancer through exhaled breath: A systematic review," *Oncotarget*, vol. 6, no. 36, pp. 38 643–38 657, 2015.

[8] Z. Jia, A. Patra, V. K. Kutty, and T. Venkatesan, "Critical review of volatile organic compound analysis in breath and in vitro cell culture for detection of lung cancer," *Metabolites*, vol. 9, no. 52, 2019.

[9] Y. Saalberg and M. Wolff, "VOC breath biomarkers in lung cancer," *Clin Chim Acta*, vol. 459, pp. 5–9, 2016.

[10] W. Miekisch, J. Herbig, and J. K. Schubert, "Data interpretation in breath biomarker research: pitfalls and directions," *J Breath Res*, vol. 6, no. 3, p. 036007, 2012.

[11] T. Stacewicz, Z. Bielecki, J. Wojtas, P. Magryta, J. Mikolajczyk, and D. Szabra, "Detection of disease markers in human breath with laser absorption spectroscopy," *Opto-Electron Rev*, vol. 24, no. 2, pp. 82–94, 2016.

[12] B. Buszewski, T. Ligor, T. Jezierski, A. Wenda-Piesik, M. Walczak, and J. Rudnicka, "Identification of volatile lung cancer markers by gas chromatography–mass spectrometry: comparison with discrimination by canines," *Anal Bioanal Chem*, vol. 404, pp. 141–146, 2012.

[13] B. Behera, R. Joshi, G. K. Anil Vishnu, S. Bhalerao, and H. J. Pandya, "Electronic nose: A non-invasive technology for breath analysis of diabetes and lung cancer patients," *J Breath Res*, vol. 13, no. 2, 2019.

[14] A. D. Wilson and M. Baietto, "Applications and advances in electronic-nose technologies," *Sensors*, vol. 9, pp. 5099–5148, 2009.

[15] R. Capuano, A. Catini, R. Paolesse, and C. Di Natale, "Sensors for lung cancer diagnosis," *J Clin Med*, vol. 8, no. 2, p. 235, 2019.

[16] A. D. Wilson, "Biomarker metabolite signatures pave the way for electronic-nose applications in early clinical disease diagnoses," *Curr Metabolomics*, vol. 5, no. 2, pp. 90–101, 2017.

[17] C. Wang and P. Sahay, "Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits," *Sensors*, vol. 9, no. 10, pp. 8230–8262, 2009.

[18] B. Henderson, A. Khodabakhsh, M. Metsälä, I. Ventrillard, F. M. Schmidt, D. Romanini, G. A. Ritchie, S. te Lintel Hekkert, R. Briot, T. Risby, N. Marczin, F. J. Harren, and S. M. Cristescu, "Laser spectroscopy for breath analysis: Towards clinical implementation," *Appl Phys B: Lasers Opt*, vol. 124, no. 8, pp. 1–21, 2018.

[19] J. D. Fenske and S. E. Paulson, "Human breath emissions of VOCs," *J Air Waste Manag Assoc*, vol. 49, no. 5, pp. 594–598, 1999.

[20] R. Salerno-Kennedy and K. D. Cashman, "Potential applications of breath isoprene as a biomarker in modern medicine: a concise overview," *Wien Klin Wochenschr*, vol. 117, no. 5, pp. 180–186, 2005.

[21] S. Das, S. Pal, and M. Mitra, "Significance of exhaled breath test in clinical diagnosis: a special focus on the detection of diabetes mellitus," *J Med Biol Eng*, vol. 36, no. 5, pp. 605–624, 2016.

[22] R. E. Amor, M. K. Nakhleh, O. Barash, and H. Haick, "Breath analysis of cancer in the present and the future," *Eur Respir Rev*, vol. 28, no. 152, 2019.

[23] M. Westhoff, P. Litterst, L. Freitag, W. Urfer, S. Bader, and J. I. Baumbach, "Ion mobility spectrometry for the detection of volatile organic compounds in exhaled breath of patients with lung cancer: Results of a pilot study," *Thorax*, vol. 64, pp. 744–748, 2009.

[24] A. Wehinger, A. Schmid, S. Mechtcheriakov, M. Ledochowski, C. Grabmer, G. A. Gastl, and A. Amann, "Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas," *Int J Mass Spectrom*, vol. 265, no. 1, pp. 49–59, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1387380607002382

[25] M. Phillips, J. Herrera, S. Krishnan, M. Zain, J. Greenberg, and R. N. Cataneo, "Variation in volatile organic compounds in the breath of normal humans," *J Chromatogr B Biomed Appl*, vol. 729, no. 1, pp. 75–88, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378434799001279

[26] R. F. Machado, D. Laskowski, O. Deffenderfer, T. Burch, S. Zheng, P. J. Mazzone, T. Mekhail, C. Jennings, J. K. Stoller, J. Pyle, J. Duncan, R. A. Dweik, and S. C. Erzurum, "Detection of lung cancer by sensor array analyses of exhaled breath," *Am J Respir Crit Care Med*, vol. 171, no. 11, pp. 1286–1291, 2005. [Online]. Available: https://doi.org/10.1164/rccm.200409-1184OC

[27] P. J. Mazzone, J. Hammel, R. Dweik, J. Na, C. Czich, D. Laskowski, and T. Mekhail, "Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array," *Thorax*, vol. 62, no. 7, pp. 565–568, 2007.

[28] S. Dragonieri, J. T. Annema, R. Schot, M. P. van der Schee, A. Spanevello, P. Carratú, O. Resta, K. F. Rabe, and P. J. Sterk, "An electronic nose in the discrimination of patients with non-small cell lung cancer and copd," *Lung Cancer*, vol. 64, no. 2, pp. 166–170, 2009.

[29] A. Bajtarevic, C. Ager, M. Pienz, M. Klieber, K. Schwarz, M. Ligor, T. Ligor, W. Filipiak, H. Denz, M. Fiegl, W. Hilbe, W. Weiss, P. Lukas, H. Jamnig, M. Hackl, A. Haidenberger, B. Buszewski, W. Miekisch, J. Schubert, and A. Amann, "Noninvasive detection of lung cancer by analysis of exhaled breath," *BMC Cancer*, vol. 9, p. 348, 2009.

[30] A. D'Amico, G. Pennazza, M. Santonico, E. Martinelli, C. Roscioni, G. Galluccio, R. Paolesse, and C. Di Natale, "An investigation on electronic nose diagnosis of lung cancer," *Lung Cancer*, vol. 68, no. 2, pp. 170–176, 2010.

[31] Y. Wang, Y. Hu, D. Wang, K. Yu, L. Wang, Y. Zou, C. Zhao, X. Zhang, P. Wang, and K. Ying, "The analysis of volatile organic compounds biomarkers for lung cancer in exhaled breath, tissues and cell lines," *Cancer Biomark*, vol. 11, no. 4, pp. 129–137, 2012.

[32] N. Peled, M. Hakim, P. A. Bunn Jr, Y. E. Miller, T. C. Kennedy, J. Mattei, J. D. Mitchell, F. R. Hirsch, and H. Haick, "Non-invasive breath analysis of pulmonary nodules," *J Thorac Oncol*, vol. 7, no. 10, pp. 1528–1533, 2012.

[33] Y. Y. Broza, R. Kremer, U. Tisch, A. Gevorkyan, A. Shiban, L. A. Best, and H. Haick, "A nanomaterial-based breath test for short-term follow-up after lung tumor resection," *Nanomed: Nanotech Biol Med*, vol. 9, no. 1, pp. 15–21, 2013.

[34] M. Bousamra II, E. Schumer, M. Li, R. J. Knipp, M. H. Nantz, V. Van Berkel, and X.-A. Fu, "Quantitative analysis of exhaled carbonyl compounds distinguishes benign from malignant pulmonary disease," *J Thorac Cardiovasc Surg*, vol. 148, no. 3, pp. 1074–1081, 2014.

[35] A. J. Hubers, P. Brinkman, R. J. Boksem, R. J. Rhodius, B. I. Witte, A. H. Zwinderman, D. A. Heideman, S. Duin, R. Koning, R. D. Steenbergen *et al.*, "Combined sputum hypermethylation and enose analysis for lung cancer diagnosis," *J Clin Pathol*, vol. 67, no. 8, pp. 707–711, 2014.

[36] T. Ligor, Ł. Pater, and B. Buszewski, "Application of an artificial neural network model for selection of potential lung cancer biomarkers," *J Breath Res*, vol. 9, no. 2, 2015.

[37] M. Li, D. Yang, G. Brock, R. J. Knipp, M. Bousamra, M. H. Nantz, and X.-A. Fu, "Breath carbonyl compounds as biomarkers of lung cancer," *Lung Cancer*, vol. 90, no. 1, pp. 92–97, 2015.

[38] M. Corradi, D. Poli, I. Banda, S. Bonini, P. Mozzoni, S. Pinelli, R. Alinovi, R. Andreoli, L. Ampollini, A. Casalini *et al.*, "Exhaled breath analysis in suspected cases of non-small-cell lung cancer: a cross-sectional study," *J Breath Res*, vol. 9, no. 2, p. 027101, 2015.

[39] J.-E. Chang, D.-S. Lee, S.-W. Ban, J. Oh, M. Y. Jung, S.-H. Kim, S. Park, K. Persaud, and S. Jheon, "Analysis of volatile organic compounds in exhaled breath for lung cancer diagnosis using a sensor system," *Sens Actuators B Chem*, vol. 255, pp. 800–807, 2018.

[40] D. Shlomi, M. Abud, O. Liran, J. Bar, N. Gai-Mor, M. Ilouze, A. Onn, A. Ben-Nun, H. Haick, and N. Peled, "Detection of lung cancer and egfr mutation by electronic nose system," *J Thorac Oncol*, vol. 12, no. 10, pp. 1544–1551, 2017.

[41] Y. Sakumura, Y. Koyama, H. Tokutake, T. Hida, K. Sato, T. Itoh, T. Akamatsu, and W. Shin, "Diagnosis by volatile organic compounds in exhaled breath from lung cancer patients using support vector machine algorithm," *Sensors*, vol. 17, no. 2, p. 287, 2017.

[42] R. Van de Goor, M. van Hooren, A.-M. Dingemans, B. Kremer, and K. Kross, "Training and validating a portable electronic nose for lung cancer screening," *J Thorac Oncol*, vol. 13, no. 5, pp. 676–681, 2018.

[43] M. Tirzïte, M. Bukovskis, G. Strazda, N. Jurka, and I. Taivans, "Detection of lung cancer with electronic nose and logistic regression analysis," *J Breath Res*, vol. 13, no. 1, p. 016006, 2018.

[44] A. Kononov, B. Korotetsky, I. Jahatspanian, A. Gubal, A. Vasiliev, A. Arsenjev, A. Nefedov, A. Barchuk, I. Gorbunov, K. Kozyrev *et al.*, "Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer," *J Breath Res*, vol. 14, no. 1, p. 016004, 2019.

[45] J. Rudnicka, T. Kowalkowski, and B. Buszewski, "Searching for selected vocs in human breath samples as potential markers of lung cancer," *Lung Cancer*, vol. 135, pp. 123–129, 2019.

[46] M. Koureas, P. Kirgou, G. Amoutzias, C. Hadjichristodoulou, K. Gourgoulianis, and A. Tsakalof, "Target analysis of volatile organic compounds in exhaled breath for lung cancer discrimination from other pulmonary diseases and healthy persons," *Metabolites*, vol. 10, no. 8, p. 317, 2020.

[47] I. A. Ratiu, T. Ligor, V. Bocos-Bintintan, C. A. Mayhew, and B. Buszewski, "Volatile Organic Compounds in Exhaled Breath as Fingerprints of Lung Cancer, Asthma and COPD," *J Clin Med*, vol. 10, no. 1, p. 32, 2020.

[48] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav Res Methods*, vol. 39, no. 2, pp. 175–191, 2007.

[49] Picomole Inc., "Breath sampler," https://www.picomole.com/breath-sampler, 2020.

[50] P. Jones, G. Harding, I. Wiklund, P. Berry, and N. Leidy, "Improving the process and outcome of care in copd: development of a standardised assessment tool," *Prim Care Respir Med*, vol. 18, no. 3, pp. 208–215, 2009.

[51] J. D. Pleil, J. R. Sobus, M. A. Stiegel, D. Hu, K. D. Oliver, C. Olenick, M. Strynar, M. Clark, M. C. Madden, and W. E. Funk, "Estimating common parameters of lognormally distributed environmental and biomonitoring data: Harmo-

nizing disparate statistics from publications," *J Toxicol Environ Health B Crit Rev*, vol. 17, no. 6, pp. 341–368, 2014.

[52] A. Gleiss, M. Dakna, H. Mischak, and G. Heinze, "Two-group comparisons of zero-inflated intensity values: The choice of test statistic matters," *Bioinformatics*, vol. 31, no. 14, pp. 2310–2317, 2015.

[53] S. Taylor and K. Pollard, "Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values," *Stat Appl Genet Mol Biol*, vol. 8, no. 1, 2009.

[54] J. Cormier and D. Dufour, "Apparatus and method for rapid and accurate quantification of an unknown, complex mixture," [Online]. Available from: https://patents.google.com/patent/WO2008074142A1, 2008, patent WO2008074142A1.

[55] S. Sharpe, T. Johnson, R. Sams, P. Chu, G. Roderick, and P. Johnson, "Gas-phase databases for quantitative infrared spectroscopy," *Appl Spectrosc*, vol. 58, no. 12, 2004.

[56] T. Johnson, L. Profeta, R. Sams, D. Griffith, and R. Yokelson, "An infrared spectral database for detection of gases emitted by biomass burning," *Vib Spectrosc*, vol. 53, pp. 97–102, 2010.

[57] I. Gordon, L. Rothman, C. Hill, R. Kochanov, Y. Tan, P. Bernath, M. Birk, V. Boudon, A. Campargue, K. Chance, B. Drouin, J.-M. Flaud, R. Gamache, J. Hodges, D. Jacquemart, V. Perevalov, A. Perrin, K. Shine, M.-A. Smith, J. Tennyson, G. Toon, H. Tran, V. Tyuterev, A. Barbe, A. Császár, V. Devi, T. Furtenbacher, J. Harrison, J.-M. Hartmann, A. Jolly, T. Johnson, T. Karman, I. Kleiner, A. Kyuberis, J. Loos, O. Lyulin, S. Massie, S. Mikhailenko, N. Moazzen-Ahmadi, H. Müller, O. Naumenko, A. Nikitin,

O. Polyansky, M. Rey, M. Rotger, S. Sharpe, K. Sung, E. Starikova, S. Tashkun, J. V. Auwera, G. Wagner, J. Wilzewski, P. Wcisło, S. Yu, and E. Zak, "The HITRAN2016 molecular spectroscopic database," *J Quant Spectrosc Radiat Transf*, vol. 203, pp. 3–69, 2017, HITRAN2016 Special Issue. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022407317301073

[58] Y. Kaya, M. Uyar, R. Tekin, and S. Yıldırım, "1d-local binary pattern based feature extraction for classification of epileptic eeg signals," *Appl Math Comput*, vol. 243, pp. 209–219, 2014.

[59] O. Lahdenoja, J. Poikonen, and M. Laiho, "Towards understanding the formation of uniform local binary patterns," *Int Sch Res Notices*, vol. 2013, 2013.

[60] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[61] M. Rota and L. Antolini, "Finding the optimal cut-point for gaussian and gamma distributed biomarkers," *Comput Stat Data Anal*, vol. 69, pp. 1–14, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947313002600

[62] R. Larracy, A. Phinyomark, and E. Scheme, "Machine learning model validation for early stage studies with small sample sizes," 2021, accepted to the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

[63] S. Parvandeh, H.-W. Yeh, M. P. Paulus, and B. A. McKinney, "Consensus features nested cross-validation," *Bioinformatics*, vol. 36, no. 10, pp. 3093–3098, 01 2020.

[64] L. I. Kuncheva and J. J. Rodríguez, "On feature selection protocols for very low-sample-size data," *Pattern Recognit*, vol. 81, pp. 660–673, 2018.

[65] S. Sone, F. Li, Z.-G. Yang, T. Honda, Y. Maruyama, S. Takashima, M. Hasegawa, S. Kawakami, K. Kubo, M. Haniuda, and T. Yamanda, "Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner," *Br J Cancer*, vol. 84, pp. 25–32, 2001. [Online]. Available: https://doi.org/10.1054/bjoc.2000.1531

[66] K. Skeldon, L. McMillan, C. Wyse, S. Monk, G. Gibson, C. Patterson, T. France, C. Longbottom, and M. Padgett, "Application of laser spectroscopy for measurement of exhaled ethane in patients with lung cancer," *Respir Med*, vol. 100, no. 2, pp. 300–306, 2006.

[67] L. Mitrayana, D. K. Apriyanto, and M. Satriawan, "Co2 laser photoacoustic spectrometer for measuring acetone in the breath of lung cancer patients," *Biosensors*, vol. 10, no. 55, 2020.

[68] C. Turner, P. Španěl, and D. Smith, "A longitudinal study of ammonia, acetone and propanol in the exhaled breath of 30 subjects using selected ion flow tube mass spectrometry, SIFT-MS," *Physiol Meas*, vol. 27, no. 4, pp. 321–337, feb 2006. [Online]. Available: https://doi.org/10.1088/0967-3334/27/4/001

[69] O. Vaittinen, F. M. Schmidt, M. Metsälä, and L. Halonen, "Exhaled breath biomonitoring using laser spectroscopy," *Current Anal Chem*, vol. 9, no. 3, pp. 463 – 475, 2013.

[70] K. K. Chow, M. Short, and H. Zeng, "A comparison of spectroscopic techniques for human breath analysis," *Biomed Spectrosc Imaging*, vol. 1, no. 4, pp. 339–353, 2012.

[71] N. Peled, O. Barash, U. Tisch, R. Ionescu, Y. Y. Broza, M. Ilouze, J. Mattei, P. A. Bunn, F. R. Hirsch, and H. Haick, "Volatile fingerprints of cancer specific genetic mutations," *Nanomed Nanotech Biol Med*, vol. 9, no. 6, pp. 758–766, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1549963413000142

[72] A. Dent, T. Sutedja, and P. Zimmerman, "Exhaled breath analysis for lung cancer," *J Thorac Dis*, vol. 5, no. S5, pp. S540–S550, 2013.

[73] J. Rudnicka, T. Kowalkowski, T. Ligor, and B. Buszewski, "Determination of volatile organic compounds as biomarkers of lung cancer by spme–gc–tof/ms and chemometrics," *J Chromatogr B Biomed Appl*, vol. 879, no. 30, pp. 3360–3366, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570023211005812

[74] S. Kischkel, W. Miekisch, A. Sawacki, E. M. Straker, P. Trefz, A. Amann, and J. K. Schubert, "Breath biomarkers for lung cancer detection and assessment of smoking related effects — confounding variables, influence of normalization and statistical algorithms," *Clin Chim Acta*, vol. 411, no. 21, pp. 1637–1644, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009898110003840

[75] J. Rudnicka, M. Walczak, T. Kowalkowski, T. Jezierski, and B. Buszewski, "Determination of volatile organic compounds as potential markers of lung cancer by gas chromatography–mass spectrometry versus trained dogs," *Sens Actuators B Chem*, vol. 202, pp. 615–621, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925400514006947

[76] A. Ulanowska, T. Kowalkowski, E. Trawińska, and B. Buszewski, "The application of statistical methods using VOCs to identify patients with lung

cancer," *J Breath Res*, vol. 5, no. 4, p. 046008, 2011. [Online]. Available: https://doi.org/10.1088/1752-7155/5/4/046008

[77] C. N. Harvey-Woodworth, "Dimethylsulphidemia: the significance of dimethyl sulphide in extra-oral, blood borne halitosis," *Br Dent J*, vol. 214, no. 7, p. E20, 2013.

[78] M. Barker, M. Hengst, J. Schmid, H.-J. Buers, B. Mittermaier, D. Klemp, and R. Koppmann, "Volatile organic compounds in the exhaled breath of young patients with cystic fibrosis," *Eur Respir J*, vol. 27, no. 5, pp. 929–936, 2006. [Online]. Available: https://erj.ersjournals.com/content/27/5/929

[79] S. R. Markar, T. Wiggins, S. Antonowicz, S.-T. Chin, A. Romano, K. Nikolic, B. Evans, D. Cunningham, M. Mughal, J. Lagergren, and G. B. Hanna, "Assessment of a noninvasive exhaled breath test for the diagnosis of oesophagogastric cancer," *JAMA Oncol*, vol. 4, no. 7, pp. 970–976, 2018.

[80] K. J. Raninen, J. E. Lappi, M. L. Mukkala, T.-P. Tuomainen, H. M. Mykkänen, K. S. Poutanen, and O. J. Raatikainen, "Fiber content of diet affects exhaled breath volatiles in fasting and postprandial state in a pilot crossover study," *Nutr Res*, vol. 36, no. 6, pp. 612–619, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027153171600049X

[81] I. Kushch, B. Arendacká, S. Štolc, P. Mochalski, W. Filipiak, K. Schwarz, L. Schwentner, A. Schmid, A. Dzien, M. Lechleitner, V. Witkovský, W. Miekisch, J. Schubert, K. Unterkofler, and A. Amann, "Breath isoprene – aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study," *Clin Chem Lab Med*, vol. 46, no. 7, pp. 1011–1018, 2008.

[82] A. Smolinska, A.-C. Hauschild, R. R. R. Fijten, J. W. Dallinga, J. Baumbach, and F. J. van Schooten, "Current breathomics—a review on data preprocessing techniques and machine learning in metabolomics breath analysis," *J Breath Res*, vol. 8, no. 2, p. 027105, apr 2014. [Online]. Available: https://doi.org/10.1088/1752-7155/8/2/027105

[83] I. S. Patel, W. R. Premasiri, D. T. Moir, and L. D. Ziegler, "Barcoding bacterial cells: a sers-based methodology for pathogen identification," *J Raman Spectrosc*, vol. 39, no. 11, pp. 1660–1672, 2008. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jrs.2064

[84] R. E. Roberts, J. E. A. Selby, and L. M. Biberman, "Infrared continuum absorption by atmospheric water vapor in the 8–12-μm window," *Appl Opt*, vol. 15, no. 9, pp. 2085–2090, Sep 1976.

Appendix A

Compound Library

A total of 152 compounds were included in the reference library for the stepwise spectral fitting procedure. The compounds' absorption coefficients at each wavelength were calculated from the estimated concentration values (in ppbV) from the Quantitative IR database [55,56], with the exception of ammonia, carbon dioxide and water which were based on the HITRAN 2016 database [57]. The water cross-section was modified with added continuum absorption according to RSB (Roberts-Selby-Biberman) approximation [84].

Table A.1: Compounds included in stepwise fitting procedure.

Compound Name	**Compound Name**
1-Butene	Ethanol
1-Decene	Ethene
1-Heptene	Ethyl acetate
1-Hexanoic acid	Ethyl benzene
1-Hexene	Ethyl butyrate
1-Octene	Ethyl mercaptan
1-Pentanol	Ethyl tert-butyl ether
1-Pentene	Formaldehyde
1-Propanol	Formic acid
1-Undecene	Furan
1,1,1-Trichloroethane	Furfural
1,2-Dichlorobenzene	Glycolaldehyde

Table A.1: Continued from previous page.

Compound Name	Compound Name
1,2,3,5-Tetramethylbenzene	Heptane
1,3-Butadiene	Hexafluoropropene
2-Butene	Hexanal
2-Butoxyethanol	Hexane
2-Ethyl-1-hexanol	Hydrogen peroxide
2-Ethyltoluene	Hydrogen sulfide
2-Hexanol	Isoamyl alcohol
2-Hexanone	Isobutane
2-Methyl-1-pentene	Isobutanol
2-Methyl-1-propanal	Isobutene
2-Methyl-2-butene	Isocumene
2-Methylfuran	Isopentane
2-Methylpentane	Isoprene
2-Nonanone	Isopropanol
2-Pentylfuran	Isopropylamine
2,3-Butanedione	Isovaleraldehyde
2,3-Dimethylbutane	Menthol
2,4-Dimethylpentane	Methanesulfonyl chloride
2,4,4-Trimethyl-1-pentene	Methanol
2,5-Dimethylfuran	Methyl acetate
3-Carene	Methyl ethyl ketone
3-Methyl-1-butene	Methyl mercaptan
3-Methylfuran	Methyl propyl ketone
3-Methylhexane	Methyl vinyl ketone
3-Methylpentane	Methylamine
4-Picoline	Methylisobutyl ketone
Acetaldehyde	Trifluoronitrosomethane
Acetic acid	Trifluoromethane
Acetol	Trifluormethyl sulfurpentafluoride
Acetone	Thioglycol
Acetonitrile	Texanol
Acrolein	Tetrahydrofuran
Ammonia	Naphthalene
Aniline	Silane
Benzaldehyde	Propyne
Benzene	Propyleneimine

Table A.1: Continued from previous page.

Compound Name	**Compound Name**
Benzyl alcohol	Propylene sulfide
Butane	Propylene glycol
Butyl acetate	Propargyl chloride
Butyraldehyde	Phosphorous oxychloride
Butyric acid	Perfluoroisobutylene
Carbon dioxide	Perfluorobutane
Carbonyl sulfide	Octafluoropropane
Chlorobenzene	N,N-Diethyl formamide
Chlorodifluoromethane	Nitromethane
Chloroform	Nitric acid
Cineole	Vinyl chloride
Crotonaldehyde	Water
Cumene	alpha-Pinene
Cycloheptane	m-Cresol
Cyclohexane	m-Xylene
Cyclohexanol	n-Butanol
Cyclohexanone	n-Decane
Cyclopentane	n-Hexadecane
Cyclopentene	n-Nonane
D-Limonene	n-Pentadecane
DL-Limonene	n-Tetradecane
Dichloromethane	n-Tridecane
Diethyl ketone	n-Undecane
Diethylether	o-Toluidine
Dimethyl ether	o-Xylene
Dimethyl sulfide	p-Xylene
Dimethylamine	tert-Butyl alcohol
Ethane	tert-Butyl methyl ether

Appendix B

Pre-Processing Techniques

The following paper is reprinted, with permission, from: Larracy, R., Phinyomark, A., and Scheme, E. "Data Pre-Processing of Infrared Spectral Breathprints for Lung Cancer Detection". Accepted to the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2021.

Data Pre-Processing of Infrared Spectral Breathprints for Lung Cancer Detection

Robyn Larracy[1], Angkoon Phinyomark[1], and Erik Scheme[1], *Senior Member, IEEE*

***Abstract*— Though breath analysis shows promise as a non-invasive and cost-effective approach to lung cancer screening, biomarkers in exhaled breath samples can be overwhelmed by irrelevant internal and environmental volatile organic compounds (VOCs). These extraneous VOCs can obscure the disease signature in a spectral breathprint, hindering the performance of pattern recognition models. In this work, pre-processing pipelines consisting of missing value replacement, detrending, and normalization techniques were evaluated to reduce these effects and enhance the features of interest in infrared cavity ring-down spectra. The best performing pipeline consisted of moving average detrending, linear interpolation for missing values, and vector normalization. This model achieved an average accuracy of 73.04% across five types of classifiers, exhibiting an 8.36% improvement compared to a baseline model ($p < 0.05$). A linear support vector machine classifier yielded the best performance (79.75% accuracy, 67.74% sensitivity, 87.50% specificity). This work can serve to guide pre-processing in future lung cancer breath research and, more broadly, in infrared laser absorption spectroscopy in general.**

I. Introduction

Exhaled breath analysis is an increasingly prominent area of research, with potential applications in the detection and monitoring of innumerable diseases. Breath is composed of both exogenously and endogenously produced compounds, the latter of which give insight into the body's metabolic processes and consequently, the state of the individual's health [1]. Since sample collection is non-invasive, painless, and does not require skilled medical staff, breath analysis is an attractive alternative to traditional imaging and biopsy techniques.

One of the most natural applications for breath testing is screening for lung cancer–the world's most common and deadliest cancer [2]–which is most often discovered too late to be effectively treated [3]. Some health organizations have recommended low-dose computed tomography for early lung cancer detection, but the technology's costs and tendency for overdiagnosis have hindered the implementation of widespread screening programs [4]. The need for more effective and accessible screening has motivated considerable effort in recent years to identify breath biomarkers for lung cancer. Studies have used various technologies for breath profile analysis, such as ion-mobility spectrometry, proton transfer reaction-mass spectrometry, and e-nose sensor arrays, but the most common technique by far is gas chromatography-mass spectrometry (GC-MS) [5]. GC-MS is popular because it is able to identify volatile organic compounds (VOCs) in a sample with near-certainty, allowing researchers to define lung cancer by the presence or elevation of particular VOCs in the breath.

Unfortunately, the VOCs concluded to be biomarkers across studies of this type are inconsistent and occasionally contradictory [6]. In a review of fifty mass-spectrometry studies, the most frequently confirmed biomarkers were each found only five times (i.e., 10% of the studies) [7]. Jia et al. [6] attributed this lack of agreement to a number of factors, many of which would be difficult or impossible to control in a screening context. These include environmental conditions at the time of collection like ambient temperature, humidity, and exogeneous VOCs, as well as individual-specific differences like diet, smoking habits, gender, and comorbidities. Given the complex relationships, origins, and metabolic pathways for the VOCs in exhaled breath, which are generally not well understood [1], it may not be possible to define the wide range of potential lung cancer breath profiles in terms of a handful of VOCs and their concentrations.

Indeed, the entire composition of the breath sample may be necessary to fully characterize the lung cancer in an individual. Rather than identifying and quantifying specific VOCs, the 'breathprinting' approach to breath analysis considers the overall sensor or analyzer response, a complex pattern encompassing the blend of all VOCs in the breath. Machine learning techniques can then be used to extract the underlying disease signature from these patterns, enabling the recognition of the disease in future samples. In this way, distinguishing features are captured that might otherwise be missed with a VOC-specific approach.

Though less common than GC-MS in breath analysis research, laser absorption spectroscopy (LAS) is an attractive alternative for capturing breathprints. In recent years, advances in analyzer hardware and laser sources have progressed LAS techniques to a degree comparable to GC-MS in sensitive, effective breath profiling [8]. Furthermore, the costly, time-consuming GC-MS analysis is generally restricted to laboratory research, whereas laser-based technologies have the potential to advance breath testing to real-world, clinical applications. These optical techniques offer comparatively quick analysis times, require little to no maintenance or calibration, and the analyzers can be operated by non-experts [9]. With sufficient spectral range and resolution, LAS is a practical, robust, efficient method

*This work was supported in part by Mitacs Canada, the New Brunswick Innovation Foundation, and the New Brunswick Health Research Foundation.

[1]All authors are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada, rlarracy@unb.ca, aphinyom@unb.ca, escheme@unb.ca

for attaining breath profiles for lung cancer screening.

However, even with low instrumentation noise, spectral breathprints vary greatly for different individuals and environments due to the diversity of VOC profiles. The critical information in the spectrum, the lung cancer's signature, may be very subtle and easily masked by these highly variable, irrelevant signals [10]. Pre-processing techniques are therefore necessary to remove noise, improve uniformity across spectra, and enhance important, discriminating features prior to training a learning algorithm. This integral step allows for the recovery of the disease's true spectral biomarkers, and thus the development of robust classification models that can generalize to the entire population of lung cancer individuals.

Hence, in this study, a comprehensive investigation of preprocessing techniques was performed for an ultra-sensitive form of LAS, cavity ring-down spectroscopy (CRDS). Various normalization, detrending, and missing value replacement techniques were evaluated for CRDS spectra based on their ability to reveal the spectral features that accurately distinguish non-small cell lung cancer patients from control subjects. An analysis of this type has not yet been performed for spectral lung cancer breathprints, nor for CRDS spectra in general. This study aims to recommend techniques that can reduce the effects of irrelevant VOCs in the spectra, which should translate to various LAS breath analysis applications.

II. Methods

A. Data

One hundred biopsy-confirmed lung cancer patients and 98 control subjects were enrolled in the study to provide breath samples. Subjects gave informed consent as per the Horizon Health Network's Research Ethics Board (#100099), and these analyses were conducted as approved by the University of New Brunswick's Research Ethics Board (#2019-068).

Collection was performed at three different hospitals using Picomole's exhaled breath sampler [11], which tracks CO_2 levels to collect alveolar breath into Tenax TA sorbent tubes. Subjects were asked to abstain from smoking for 4 hours and drinking alcohol for 8 hours prior to collection, where they were instructed to breathe deeply and exhale into a single-use filter on the sampler's mouthpiece until 10-litre (10L) samples were amassed. Post-collection, the inclusion criteria for lung cancer subjects were amended to exclude patients with ambiguous or small-cell histologic subtypes and patients that had undergone any form of lung cancer treatment. Also disqualifying subjects that had missing data (for example, if they were unable to provide the full 10L sample), the remaining 62 pre-treatment, non-small cell lung cancer (NSCLC) and 96 control subjects were used for this analysis. A comparison of demographics and clinical factors for the two cohorts is provided in Table I.

Infrared breath profiles were measured for each of the 10L samples with CRDS. CRDS uses highly reflective mirrors to increase the effective path length of light trapped in an optical cavity. For a gas sample within the cavity, the decay rate of the trapped light is measured to establish the sample's absorption spectrum. Measurements are ultra-sensitive due

TABLE I

Subject Demographics and Clinical Factors

Factor	Lung Cancer	Control	*p*-value
Sample size	62	96	-
Sex			
Female	50%	53.1%	.75
Male	50%	46.9%	
Age ($\mu \pm \sigma$, years)			
Female	68.2 ± 9.1	61.0 ± 14.3	.01*
Male	71.3 ± 8.3	65.9 ± 12.1	.03*
Smoking			
Current smokers	19.4%	6.7%	
Former smokers	75.8%	48.9%	$< .0001^{\dagger}$
Never smokers	4.8%	44.4%	
Other lung conditions	44(71.0%)	36(37.5%)	$< .0001^{\dagger}$
Diagnosis			
Adenocarcinoma	58.1%		
Squamous cell carcinoma	37.1%	-	-
Unspecified NSCLC	4.8%		

* t-test indicated a significant difference between groups ($p < 0.05$)
† Fisher's exact test indicated a significant difference between groups ($p < 0.05$)

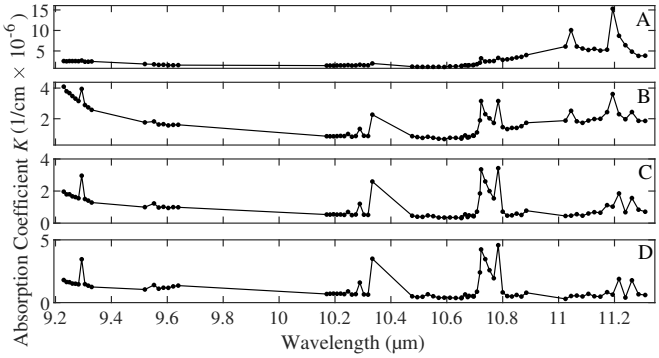

Fig. 1. Example spectra for one subject, desorbed at temperatures A) 75°C, B) 150°C, C) 225°C, D) 300°C.

to the long path length of the laser, which is approximately one kilometer in total, and are unaffected by fluctuations in laser intensity. Two CO_2 lasers with carbon isotopes ^{12}C and ^{13}C were tuned to a combined 73 lines in the mid-infrared region, a favourable spectral range for the detection of small molecules [12]. At each wavelength, the average times from 500 ring-downs were measured for the breath sample (τ) and for a baseline nitrogen sample (τ_0). The absorption coefficients K comprising each spectrum were calculated from the average ring-down times according to $K = \frac{\tau_0 - \tau}{c \cdot \tau_0 \cdot \tau}$, where c is the speed of light. The analysis was performed four times for each sample, at desorption temperatures 75, 150, 225, and 300°C, yielding four different spectra per subject. Fig. 1 shows the four spectra obtained for one subject.

B. Data Pre-Processing Techniques

a) Missing Value Replacement: The imputation techniques considered in this work fall into two categories: 1-Way and n-Way. The 1-Way methods interpolate missing values in a spectrum using only information from within that spectrum, independent of all other subjects. The four employed methods of this type are one (linear interpolation) that uses only the absorption coefficients for the two closest available wavelengths, and three (cubic spline, PCHIP: piece-

TABLE II

LIST OF n-WAY MISSING VALUE REPLACEMENT TECHNIQUES

Technique	Description
Euclidean	$D = \sqrt{\sum_{i=1}^{N}[x_j(i) - x_k(i)]^2}$
Stand. Euclidean[a]	$D = \sqrt{\sum_{i=1}^{N}[x_j(i)/\sigma_1 - x_k(i)/\sigma_2]^2}$
City Block	$D = \sum_{i=1}^{N} \lvert x_j(i) - x_k(i)\rvert$
Chebyshev	$D = \max(\lvert x_j - x_k\rvert)$
Minkowski[b]	$D = \sqrt[p]{\sum_{i=1}^{N}[x_j(i) - x_k(i)]^p}$
Cosine	$D = 1 - \frac{\sum_{i=1}^{N} x_j(i)x_k(i)}{\sqrt{\sum_{i=1}^{N} x_j(i)^2 \cdot \sum_{i=1}^{N} x_k(i)^2}}$
Correlation[c]	$D = 1 - \frac{\sum_{i=1}^{N}[x_j(i)-\mu_j]\cdot[x_k(i)-\mu_k]}{\sqrt{\sum_{i=1}^{N}[x_j(i)-\mu_j]^2 \cdot \sum_{i=1}^{N}[x_k(i)-\mu_k]^2}}$

[a] σ_j and σ_k denote standard deviations of spectra x_j and x_k
[b] p denotes the Minkowski distance order ($p = 3$ in this study)
[c] μ_j and μ_k denote means of spectra x_j and x_k

TABLE III

LIST OF DETRENDING TECHNIQUES

Technique	Description
Constant Offset	$y_j = \min(x_j)$
Linear[a]	$y_j = a_j \cdot x_j + b_j$
Quadratic[a]	$y_j = a_j \cdot x_j^2 + b_j \cdot x_j + c_j$
Moving Average	$y_j(i) = \frac{1}{2l+1}\sum_{k=i-l}^{i+l} x_j(k), i = l, 2, ...N - l$

[a] a_j, b_j and c_j denote the estimated coefficients for the fit to x_j

wise cubic hermite interpolating polynomial, and modified Akima interpolation) that fit splines to the entire spectrum to extrapolate the missing points.

Contrarily, n-Way techniques use information from multiple spectra, leveraging measurements from other subjects to find suitable replacement values. The seven equations provided in Table II describe the N-dimensional distance (D) between two spectra, x_j and x_k, for $N = 73$ coefficients. For a spectrum with missing coefficients, D is used to define the 10 most similar spectra from the training set. A mean of the corresponding values in these neighboring spectra are used to replace the missing points for the spectrum in question. To account for spectra with multiple missing coefficients, all calculated distances were adjusted by a correction factor N/n, where n is the number of wavelengths that are non-missing for both x_j and x_k.

b) Detrending: Four 1-Way detrending techniques were considered in this work, presented in Table III. In each case, the trend y_j was re-estimated for each spectrum x_j and subtracted to obtain the corrected residuals. For the constant offset technique, this trend is simply a 0$^{\text{th}}$ degree polynomial representing the minimum value in the spectrum. The linear and quadratic detrending techniques are based on least-squares polynomial fitting. The moving average method, also referred to as 0$^{\text{th}}$ order Savitzky-Golay detrending [13], captures the trend by smoothing the spectrum with a very large moving window of $2l + 1$ points ($l = 18$ in this study).

c) Normalization: Six different techniques are defined in Table IV; five 1-Way methods and one n-Way method. The presented 1-Way normalization methods are common within spectroscopy fields [10] and elsewhere, employing properties

TABLE IV

LIST OF NORMALIZATION TECHNIQUES

Technique	Description
1-Way	
Min-Max	$\tilde{x}_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}$
1-Norm	$\tilde{x}_j = \frac{x_j - \mu_j}{\sum_{i=1}^{N} \lvert x_j(i)\rvert}$
Vector	$\tilde{x}_j = \frac{x_j - \mu_j}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_j(i)^2}}$
Peak	$\tilde{x}_j = \frac{x_j - \mu_j}{\max(x_j)}$
Standard Normal Variate (SNV)	$\tilde{x}_j = \frac{x_j - \mu_j}{\sigma_j}$
***n*-Way**	
Multiplicative Scatter Correction (MSC)	$\tilde{x}_j = \frac{x_j - a_j}{b_j}$

within the spectrum like its mean μ_j and standard deviation σ_j for correction. The n-Way method, multiplicative scatter correction [13], first requires an ideal reference spectrum, which is estimated by taking the mean across all M subjects in the training set (x_{avg}). To correct a spectrum x_j, it is regressed onto x_{avg} using a least-squares criterion, yielding coefficients a_j and b_j to be used for normalization.

C. Evaluating Techniques

A pre-processing pipeline consists of one technique from each of the three categories (Section II-B), applied sequentially to the spectra in a given order. Cases without normalization and detrending were also considered. Each pipeline was evaluated by its ability to enhance the distinguishing characteristics of the spectra.

Specifically, for a given set of techniques, the four spectra for each subject were first pre-processed and combined into a single feature matrix, consisting of a total 292 features (73 per spectrum). The minimum redundancy maximum relevance (mRMR) feature selection algorithm [14] was then used to sequentially rank these features based on a difference of Pearson correlation coefficients. Linear discriminant analysis (LDA) classifiers with feature sets ranging from 1 to 79 mRMR-ranked features (half the sample size) were trained and validated using leave-one-out cross-validation. This procedure was repeated for every permutation of the techniques under consideration. In the end, the model that provided the best classification accuracy was recorded, thereby tuning both the order of techniques and the number of selected features. In case of a tie, models with fewer selected features were preferred.

The 20 best-performing pipelines were further tested with four other types of learning algorithms: a support vector machine (SVM) with a linear kernel, quadratic discriminant analysis (QDA) classification, k-nearest neighbor (KNN) classification ($k = 5$), and a random forest (RF) with 100 decision trees. The average classification accuracies across all learning algorithms were used for the final pipeline ranking to ensure that a pipeline's performance was not dependent on a single type of classifier.

Additionally, to replicate real-world settings in which the classifier's test subjects would not be seen during the pre-processing stage, some modifications were made for

TABLE V

FIVE TOP PERFORMING PRE-PROCESSING PIPELINES

Rank	Missing Value Replacement	Detrending	Normalization	Accuracy $\mu(\sigma)\%$
1	Linear[2]	Moving Average[1]	Vector[3]	73.04 (6.33)
2	PCHIP[3]	Moving Average[1]	Peak[2]	72.91 (4.27)
3	Akima[1]	Moving Average[2]	SNV[3]	72.66 (4.39)
4	Stand. Euclidean[2]	Quadratic[1]	SNV[3]	72.41 (5.31)
5	PCHIP[1]	Moving Average[2]	Vector[3]	72.28 (5.45)

[1,2,3] Pipeline order (1st, 2nd, 3rd step, respectively)

TABLE VI

TECHNIQUES' AVERAGE LDA PERFORMANCE ACROSS ALL PIPELINES

Category	Rank	Technique	Accuracy $\mu(\sigma)\%$
Missing Value Replacement	1	Stand. Euclidean	72.35 (2.35)
	2	PCHIP	72.12 (3.79)
	3	Linear	71.36 (3.36)
	4	M-Akima	70.96 (3.65)
	5	Euclidean	70.76 (3.24)
	6	Cityblock	70.51 (3.40)
	7	Cosine	70.51 (2.81)
	8	Minkowski	70.49 (3.34)
	9	Chebyshev	70.13 (3.07)
	10	Cubic Spline	70.11 (3.59)
	11	Correlation	69.75 (2.77)
Detrending	1	Moving Average	71.91 (2.85)
	2	Quadratic	71.27 (3.38)
	3	Linear	70.59 (2.72)
	4	Offset	70.28 (4.21)
	5	None	70.05 (4.21)
Normalization	1	SNV	73.03 (2.23)
	2	Vector	72.93 (1.83)
	3	1-Norm	72.31 (1.36)
	4	Peak	71.89 (1.93)
	5	MSC	71.46 (2.57)
	6	Min-Max	68.75 (2.52)
	7	None	65.37 (2.18)

pipelines containing n-Way techniques compared to those containing only 1-Way techniques. The n-Way techniques were applied using information from training subjects only, requiring both the pre-processing and feature selection steps to be repeated for each cross-validation fold. For pipelines consisting of only 1-Way techniques, a single pre-processing step was sufficient since the calculations for one subject were independent from all other subjects.

III. RESULTS AND DISCUSSION

Of a possible 385 sets of techniques (equivalent to 1815 pipelines total, considering all permutations), the twenty best performing pipelines with the LDA classifier were further reduced to the top five in Table III based on performance in SVM, QDA, KNN and RF classifiers. The top pipeline across these five classifiers consisted of (1) moving average detrending, (2) linear interpolation for missing value replacement, and (3) vector normalization. With this pipeline, the SVM classifier achieved the highest accuracy, 79.75% (67.74% sensitivity and 87.50% specificity). Compared to a baseline model that applied only standardized Euclidean imputation without detrending or normalization, the top pipeline produced a significant improvement in classification performance (on average across the five learning algorithms, $73.04\% > 64.68\%$; $p < 0.05$). This finding demonstrates the importance of a data pre-processing step, which is strongly recommended for the development of future models using spectral breathprints.

TABLE VII

FREQUENCY OF SELECTED PIPELINE ORDERS WITH LDA

Instances	First	Second	Third
94 (35.6%)	Detrending	Normalization	Missing Value Replacement
91 (34.5%)	Missing Value Replacement	Detrending	Normalization
66 (25.0%)	Detrending	Missing Value Replacement	Normalization
5 (1.9%)	Missing Value Replacement	Normalization	Detrending
5 (1.9%)	Normalization	Missing Value Replacement	Detrending
3 (1.1%)	Normalization	Detrending	Missing Value Replacement

Notably, there were no statistically significant differences between the accuracies for the top pipeline and the next four highest ranked ones in Table III ($p > 0.05$), indicating that these five pipelines are essentially interchangeable for this application. In fact, many of the top pipelines are quite similar, most often employing a form of shape-preserving interpolation combined with moving average detrending and 1-Way normalization.

To observe the performance of individual pre-processing techniques more generally, the LDA model accuracies were averaged across all pipelines employing the same technique (Table VI). The top techniques from each category, though by a small margin in each case, are standardized Euclidean distance imputation, moving average detrending, and standard normal variate normalization. These results reflect the selected methods in the top pipelines of Table III. For other types of laser absorption spectroscopy and/or diseases in future studies, the pre-processing optimization process can be narrowed down to include only the top few techniques in each category, rather than requiring a full search.

While there are a handful of machine learning techniques that permit missing data, missing values typically need to be replaced to perform effective classification with so few subjects and variables. In Table VI, standardized Euclidean distance may work well as an n-Way technique for remedying missing CRDS coefficients because it corrects the spectra by their standard deviation prior to comparison, perhaps improving robustness to differences in large peaks compared to other distance measures. Further, lower order Minkowski variants, such as Euclidean ($p = 2$) and Cityblock distance ($p = 1$), have been shown to outperform higher order variants in high dimensions [15], evidenced in Table VI by the poorer performance obtained by third-order Minkowski and Chebyshev ($p \to \infty$) distances. Interestingly, the PCHIP, linear and M-Akima 1-Way techniques also performed well despite the unfounded relationships they necessarily assume between neighboring wavelengths. It is possible that more subjects are needed for the n-Way techniques, since there may be a lack of similar spectra in the nearest neighbor search.

Given that CRDS absorption coefficients are calculated

with respect to baseline nitrogen measurements, the baseline or background correction necessary with many other forms of spectroscopy (e.g. Raman and FTIR: Fourier transform infrared spectroscopy) do not apply. In this context, the purpose of detrending is instead to aid in the removal of trends caused by highly concentrated, extraneous VOCs. Considering the six distinct spectral branches for the two CO_2 lasers (9R, 9P, 10R, and 10P of the $^{12}CO_2$ laser, 10R and 10P of the $^{13}CO_2$ laser), moving average detrending outperformed other techniques in this respect by capturing the baselines presented in each branch, acting almost as piecewise detrending. The spectra in Fig. 1 may exemplify why quadratic detrending was also successful, since the baseline tended to be highest in the outer branches of the measured region.

In combination with detrending, normalization techniques remedy the high variability across spectra and improve the classifier's ability to recognize the lung cancer signature in the group. The best performing normalization technique over all LDA models, SNV normalization, is similar to standardized Euclidean missing value replacement in that it corrects spectra by their standard deviation. Vector, 1-norm and peak normalization performed similarly well, even appearing in the top 5 pipelines of Table III. As with the n-Way missing value replacement techniques, it is possible that more subjects are needed to achieve satisfactory results with MSC. The results were significantly worse when no normalization was applied, though, establishing it as an essential step.

The order of pre-processing steps is also an important consideration, as evidenced in Table VII. This table presents the frequencies of the selected pipeline orders with LDA, considering all pipelines that incorporated three techniques (excluding no-detrending and no-normalization cases). Notably, detrending was preferred before normalization 95.1% of the time. This order is typically adopted in FTIR and Raman spectroscopy [10], in fact, and ensures that 1) normalization is not affected by noisy trends and 2) normalized scales are maintained. Future studies should therefore adopt this standard while optimizing the position of missing value replacement in the pipeline if necessary.

It should be noted that, first, feature extraction was omitted in this study to avoid tying the results to specific extraction methods. However, in a previous study with this dataset, spectral derivatives were extracted along with one-dimensional local binary patterns (1D-LBP), achieving a classification accuracy of 86.10% (89.60% sensitivity and 80.70% specificity) [16], indicating that the addition of feature extraction is advantageous for spectral breathprints. Second, although this study was limited to non-small cell lung cancer and cavity ring-down spectroscopy, the general guidelines regarding the best individual techniques and orders should extend to other applications, and provide a reasonable starting point for further optimization. Finally, the utilized sample size was relatively small for an optimization of this type, and a larger dataset may be necessary to substantiate the findings. Importantly, a larger sample size would also permit the implementation of deep learning techniques, which may improve performance and bypass the need for certain pre-processing steps like missing value replacement.

In conclusion, this study demonstrates the value of pre-processing techniques for extraneous VOC management and imparts a productive starting point for pattern recognition with various diseases and forms of LAS.

ACKNOWLEDGMENT

The authors would like thank Picomole Inc. for providing the raw CRDS data, their expertise and their support, especially Dr. Steve Graham, Dr. Gisia Beydaghyan, and Chris Purves, P.Eng. Additionally, we would like to acknowledge principal investigators Dr. Tony Reiman, Dr. Luisa Galvis-Gomez, and Dr. Mahmoud Abdelsalam as well as the clinicians that contributed to sample collection at the Saint John Hospital, Dr. Everett Chalmers Hospital, and the Moncton Hospital, Canada.

REFERENCES

[1] A. Dent, T. Sutedja, and P. Zimmerman, "Exhaled breath analysis for lung cancer," *J Thorac Dis*, vol. 5, no. S5, pp. S540–S550, 2013.

[2] F. Bray *et al.*, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, pp. 394–424, 2018.

[3] N. Howlader *et al.*, "SEER cancer statistics review, 1975-2016," Bethesda, MD, 2018.

[4] E. Patz Jr *et al.*, "Overdiagnosis in low-dose computed tomography screening for lung cancer," *JAMA Intern Med*, vol. 174, no. 2, pp. 269–274, 2014.

[5] G. Pennazza and M. Santonico, *Breath Analysis*. Elsevier Science, 2018.

[6] Z. Jia, A. Patra, V. K. Kutty, and T. Venkatesan, "Critical review of volatile organic compound analysis in breath and in vitro cell culture for detection of lung cancer," *Metabolites*, vol. 9, no. 52, 2019.

[7] Y. Saalberg and M. Wolff, "VOC breath biomarkers in lung cancer," *Clin Chim Acta*, vol. 459, pp. 5–9, 2016.

[8] C. Wang and P. Sahay, "Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits," *Sensors*, vol. 9, no. 10, pp. 8230–8262, 2009.

[9] B. Henderson *et al.*, "Laser spectroscopy for breath analysis: Towards clinical implementation," *Appl Phys B: Lasers Opt*, vol. 124, no. 8, pp. 1–21, 2018.

[10] R. Gautam, S. Vanga, F. Ariese, and S. Umapathy, "Review of multidimensional data processing approaches for raman and infrared spectroscopy," *EPJ T*, vol. 2, no. 1, 2015.

[11] Picomole Inc., "Breath sampler," https://www.picomole.com/breath-sampler, 2020.

[12] T. Stacewicz, Z. Bielecki, J. Wojtas, P. Magryta, J. Mikolajczyk, and D. Szabra, "Detection of disease markers in human breath with laser absorption spectroscopy," *Opto-Electronics Review*, vol. 24, no. 2, pp. 82–94, 2016.

[13] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemom Intell Lab Syst*, vol. 117, pp. 100–114, 2012.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[15] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory - ICDT 2001*, J. Van den Bussche and V. Vianu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 420–434.

[16] T. Reiman *et al.*, "Analysis of exhaled breath of lung cancer patients using infrared spectroscopy," in *2020 ASCO Virtual Scientific Program*, no. 38, 2020.

Appendix C

Evaluation Frameworks

Machine Learning Model Validation for Early Stage Studies with Small Sample Sizes

Robyn Larracy[1], Angkoon Phinyomark[1], and Erik Scheme[1], *Senior Member, IEEE*

***Abstract*— In early stage biomedical studies, small datasets are common due to the high cost and difficulty of sample collection with human subjects. This complicates the validation of machine learning models, which are best suited for large datasets. In this work, we examined feature selection techniques, validation frameworks, and learning curve fitting for small simulated datasets with known underlying discriminability, with the aim of identifying a protocol for estimating and interpreting early stage model performance and for planning future studies. Of a variety of examined validation configurations, a nested cross-validation framework provided the most accurate reflection of the selected features' discriminability, but the relevant features were often not properly identified during the feature selection stage for datasets with small sample sizes. Ultimately, we recommend that: (1) filter-based feature selection methods should be used to minimize overfitting to noise-based features, (2) statistical exploration should be conducted on datasets as a whole to estimate the level of discriminability and the feasibility of the classification problems, and (3) learning curves should be employed using nested cross-validation performance estimates for forecasting accuracy at larger sample sizes and estimating the required number of samples to converge towards best performance. This work should serve as a guideline for researchers incorporating machine learning in small-scale pilot studies.**

I. Introduction

Machine learning has enabled groundbreaking advances in the analysis of biomedical signals, images, and omics data in recent years. Due to the complexity of the physiological processes that produce these samples, data-driven pattern recognition techniques are often able to outperform conventional statistical tools [1]. Among the countless applications for machine learning in biomedical research are detecting disease biomarkers, monitoring injury rehabilitation, developing prostheses, and predicting predisposition for injury or illness.

However, while very large datasets are preferred for machine learning applications, these are often unattainable in biomedical studies. Due to the expense associated with data collection involving human participants or labeling of data by domain experts, pilot studies with small sample sizes are commonly used for determining feasibility, securing further funding, and/or for subsequent sample size planning. The future of the given projects may therefore hinge on these preliminary results meaning that, despite the small number of samples, performance estimates must reflect the true error rate of the problem.

Unfortunately, there are very real challenges facing performance estimation with small datasets. The most valuable measure of a classifier's performance is its generalization error, which reflects its performance for samples not seen at any point during the algorithm's creation. Simply splitting the available samples into random groups for training and testing, however, is generally unsuitable for low sample sizes [2]. With too few samples used for training, the quality of the model and its decision rules are weakened. Similarly, with too few test samples, performance estimates are unreliable. Hence, more efficient approaches are required to appropriately utilize the limited samples.

Endeavouring to exploit the available data as much as possible, on the other hand, can result in overly optimistic performance estimates for a problem. This occurs when the same or highly related samples that were used for feature selection, hyperparameter selection, and/or classifier training are reused for estimating a model's performance [3], [4], [5]. This information leakage makes it difficult to detect and prevent overfitting, where the model learns the noise in the dataset in addition to, or rather than, truly valuable patterns [6]. Overfitting can be especially severe with small datasets since strong spurious patterns may occur by chance, and perfect or near-perfect accuracy can be achieved by continually tuning model parameters and hyperparameters to the available samples. While many researchers are careful to avoid information leakage in the classifier training step, information leakage in the other stages of model development, especially feature selection [7], can result in similarly inflated performance estimates.

In this study, simulated datasets were created with varying degrees of discriminability to investigate performance estimation in small-sample binary classification problems. First, the impact of different feature selection techniques was investigated, based on their ability to identify relevant features and provide performance estimates that reflect the true underlying predictive power of the features. Second, these same metrics were then used to compare six configurations for developing and validating models, built on commonly used holdout, cross-validation, and bootstrapping techniques. Lastly, these validation configurations were further assessed using fitted learning curves to evaluate each configuration's ability to forecast model performance at larger sample sizes. Ultimately, the goal of this work is to provide an effective, data efficient framework for validating classification models with small sample sizes.

*This work was supported in part by Mitacs Canada.

[1]All authors are with the Institute of Biomedical Engineering, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada, rlarracy@unb.ca, aphinyom@unb.ca, escheme@unb.ca

II. Methods

A. Data

The data used in this work were simulated, and can be categorized into four types based on their level of class discriminability: (1) low, (2) moderate, (3) high, and (4) varying. Each dataset consisted of two equally sized subsets, constituting a balanced binary classification problem. Fifty features per class were randomly drawn from unit variance Gaussian distributions. Forty of these features were drawn from the same zero-mean distribution for both classes, representing irrelevant features (or noise), while 10 features were made to be differentiable between the two classes by increasing the distribution's mean for only the positive group, i.e., (1) low, with a 0.1 difference in mean for the positive and negative class distributions for each feature; (2) moderate, with a 0.5 difference; (3) high, with a 0.9 difference; and (4) varying, with differences spanning from 0.1 to 1.0 in steps of 0.1 for the 10 features.

The size of the datasets ranged across 20 sample sizes: beginning with 10, samples were added to each dataset in increments of 10 until 100 samples were reached, and then in increments of 50 until 600 samples were reached. For each of the four discriminability cases, an additional dataset of 100,000 samples was simulated to be used for estimating the true classification ability afforded by the distributions of the 10 discriminative features, with 50% randomly selected for training and 50% for testing a linear support vector machine (SVM) classifier.

B. Feature Selection

For the first experimental aim, three feature selection methods were examined for small sample sizes. First was feature selection using simple variable ranking, a filter method that orders features by the value of a scoring function (here, the 10 highest ranked features according to independent t-tests). This method is based solely on independent feature relevance to the class labels so it may not produce optimal predictors, but it is computationally efficient. A second filter method, minimum redundancy maximum relevance feature selection (mRMR), uses a scoring function (a difference of Pearson correlation coefficients in this work) to sequentially select features that exhibit high relevance to the binary class labels but also low redundancy compared to higher ranked features [8]. Third, sequential forward selection (SFS), a wrapper method, uses classification accuracy to establish a well-performing subset of features from the dataset. Successively, features are added to the subset that exhibit the best improvement in accuracy of all remaining features. In this work, linear SVM classifiers were used to obtain resubstitution performance estimates for the SFS procedure.

These three methods were assessed using a holdout validation configuration for each of the twenty sample sizes, repeated for 100 iterations with the varying discriminability data. Linear SVM classifiers were used to estimate both training accuracy, for the 80% subset of samples that were used to train the algorithm, and test accuracy, for the 20% test set unseen during development or training. The feature selection methods were further evaluated based on their ability to identify the truly relevant features, those that were simulated with real underlying differences rather than those that exhibited coincidental noise-based differences, based on the percentage of correctly selected features.

C. Model Validation

For the second experimental aim, six model validation techniques were considered for assigning samples to training and test groups. The first two 'non-nested' techniques (A and B) utilized all available samples for model development, comprised of only a feature selection step in this work, while four 'nested' techniques (C, D, E and F) performed model development using only training subsets of the data.

A) Leave-one-out cross-validation (LOOCV): model development is performed using all samples, then the model's performance is evaluated using LOOCV where one sample is used as the model's test set and the remaining samples as the training set. This is repeated for each of the n samples, shifting the test exemplar each time, and the final performance estimate is averaged from all samples. The same selected features and hyperparameters are retained for each surrogate model.

B) Bootstrapping: model development is performed using all samples, then the model's performance is estimated using 50 bootstrap subsamples of n samples. Specifically, bootstrapping performs random subsampling with replacement for establishing the model's training set and all remaining samples (those 'out-of-bag') are used as test samples. This procedure is repeated for several iterations and the out-of-bag performance estimates are averaged across all surrogate models to determine the final estimate. As with the LOOCV configuration, the same selected features and hyperparameters are retained for each surrogate model.

C) Holdout: 20% of samples are randomly selected for testing, the other 80% are used for both model development and training.

D) Nested k-fold cross-validation (Nested 10-CV): samples are randomly partitioned into k approximately equal sized groups (in this work, $k = 10$). In turn, each group is used as the test set while all remaining groups are used for model development and training, creating k surrogate models. The final performance estimate is the average test accuracy across all surrogate models. Note that model development is repeated in each iteration using only training samples.

E) Nested LOOCV: n partitions are used for the nested cross-validation procedure. Again, model development is performed independently for each surrogate model.

F) Nested Bootstrapping: n samples are subsampled with replacement for 50 iterations. Model development is repeated in each iteration using only training samples.

For each sample size and discriminability level, the six validation configurations were applied to one hundred iterations of simulated data. The t-test variable ranking method

was used for feature selection and linear SVM classifiers were used to estimate classification performance.

D. Learning Curve

For the third experimental aim, learning curves were used to evaluate the forecasting capability of each of the validation configurations. The commonly-used inverse power law model was adopted for this purpose, as in (1) where Y is the fitted curve, n is sample size, a is the best achievable error rate, b is the learning rate, and c is the decay rate [9].

$$Y = (1 - a) - b \cdot n^c \tag{1}$$

Briefly, the learning curve fitting procedure is as follows:

1) Randomly select a stratified subset of n_0 samples from the dataset.
2) Apply model development and validation to obtain a classification performance estimate for the subset.
3) Add m stratified samples to the existing subset and re-apply model development and validation to estimate the new performance. Repeat this step until all available samples have been added, yielding a sequence of classification performance estimates for sample sizes n_0 to the total number of samples, n_{max}.
4) Use least-squares regression to fit an inverse power law to the sequence of performance estimates.
5) Infer the model's performance at larger sample sizes from the fitted curve.

For each of the one hundred iterations performed using varying discriminability data in Section II-C, learning curves were fit to the first ten measured performance estimates, ranging in sample sizes from 10 (n_0) to 100 (n_{max}) in increments of 10 (m) samples. An n_{max} of 100 was selected to exemplify a typical small-scale biomedical dataset. The remaining estimates, ranging in sample sizes from 150 to 600, were used to determine the fit's root-mean-square error (RMSE) at larger sample sizes, and thus, the quality of the forecast. While the value of c was unconstrained during the fitting procedure, a was bounded between 0 and 1 and b was made to be positive to enforce the desired convergence to $1 - a$ (the stabilized accuracy).

III. Results and Discussion

A. Feature Selection

Fig. 1(a) illustrates the average performance of the three feature selection methods using the holdout validation method for 100 simulated datasets of varying discriminability. The training and test curves are revealing: first, models built using small sample sizes clearly suffer from training set overfitting, considering the extremely high training accuracies but poor test accuracies. In these cases, generalization ability can be improved, at least to a certain degree, with the addition of more training data. This exemplifies the well-established advantages of larger sample sizes in machine learning model development. Second, SFS, the wrapper method, suffered from overfitting more than both the variable ranking and mRMR filter methods, especially for small

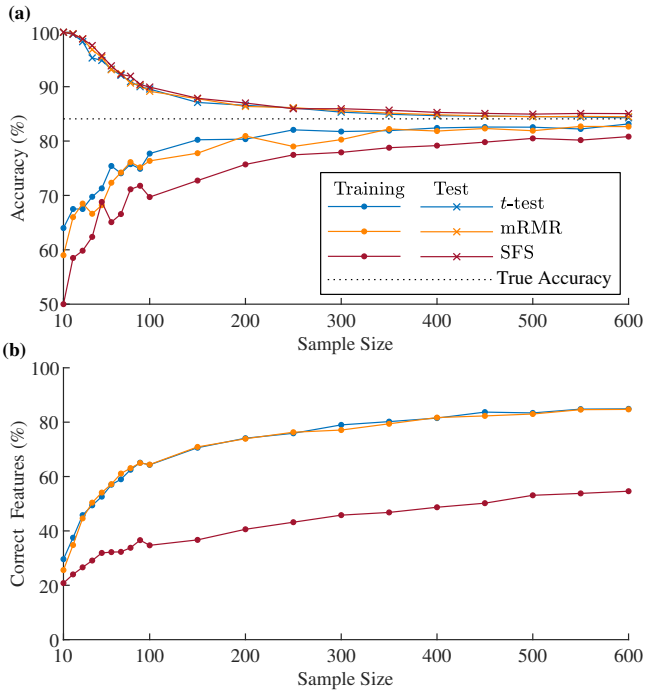

Fig. 1. (a) Average training and test classification accuracies and (b) percentage of correctly selected features using the holdout validation technique with the feature ranking based on t-test, mRMR, and SFS methods for sample sizes of 10-600.

sample sizes where a 10%-15% disparity in testing accuracy was evident. Correspondingly, it can also be observed from the percentage of correct features selected (Fig. 1(b)) that SFS was least effective in identifying the features with true underlying differences, selecting on average over 20% more random noise features than the filter methods. These results suggest that filter-based feature selection methods may be most appropriate in early stage studies with small sample sizes, and more advanced wrapper-based methods should be reserved for larger sample sizes. Lastly, the simple variable ranking technique based on t-tests had the best performance in this study. This may be due in part to the fact that the simulated datasets consisted of normally distributed and uncorrelated features which are ideal conditions for this method. For complex real-world data where correlated, non-normal features can be expected, more advanced filter-based feature selection methods like mRMR, measuring both relevance and redundancy, may be a better fit.

B. Model Validation

The average classification performance estimates for each of the six validation techniques over 100 trials are presented in Fig. 2, for all four levels of dataset discriminability. As in the comparison of feature selection methods, the proportions of correctly selected features are also presented.

For the low discriminability case (Fig. 2(a)), the underlying predictive power was barely better than random data (50% for binary classification), exhibiting a true accuracy of only 56%. With such a small effect, the truly relevant features can be indistinguishable from noise-based features.

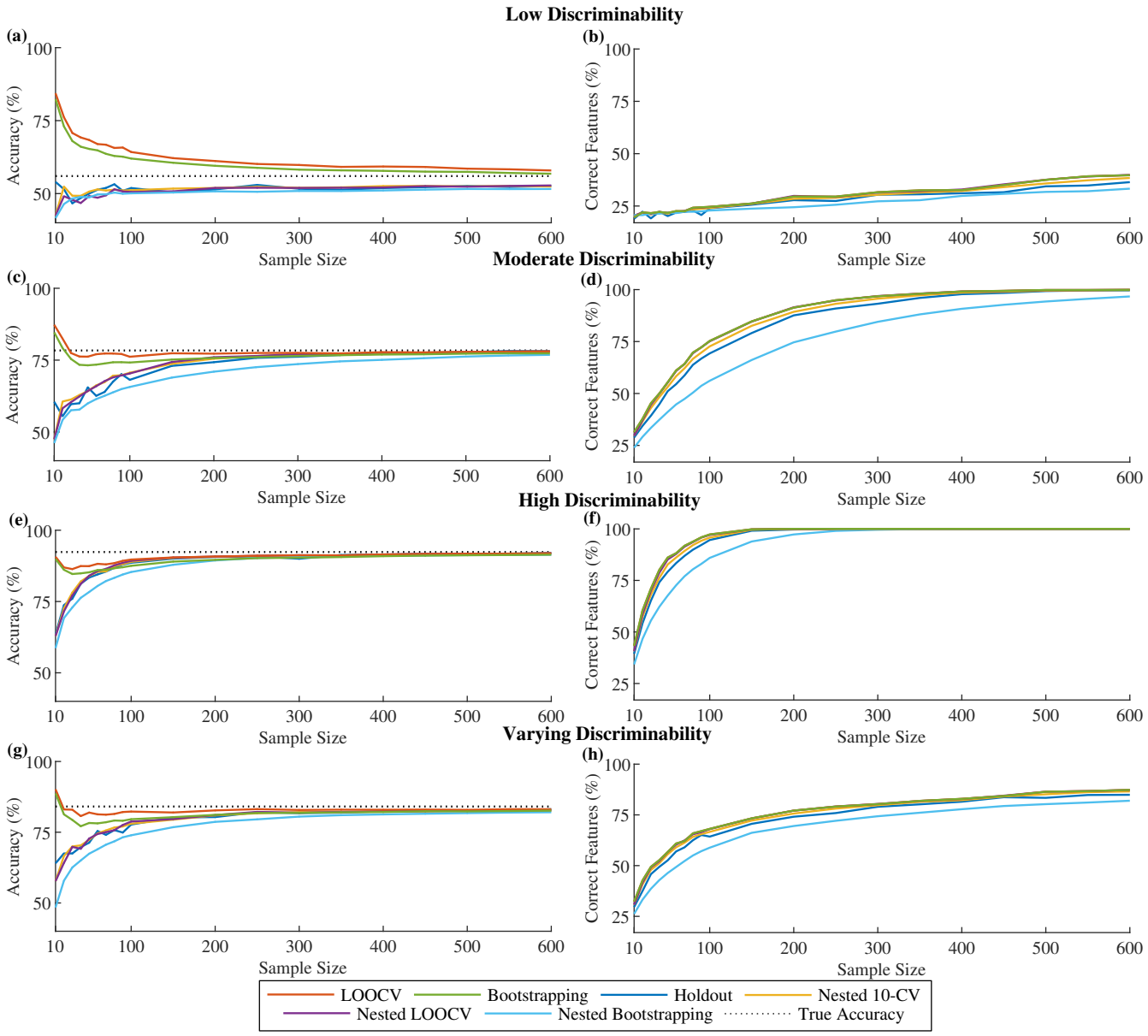

Fig. 2. (a, c, e, g) Average test accuracies and (b, d, f, h) percentage of correctly selected features for each validation configuration and level of data discriminability: (a, b) low, (c, d) moderate, (e, f) high, and (g, h) varying. The feature ranking based on t-testing was applied for feature selection and the linear SVM classifier was used for determining test accuracies.

This was evident during the simulations given the low proportions of correctly selected features with this data, which remained below 50% across all examined sample sizes and validation configurations. Due to information leakage during model development, the non-nested validation techniques, LOOCV and bootstrapping, yielded highly over-optimistic performance estimates for this mostly random data. The problem was most apparent with sample sizes below 100, but the optimism remained even up to 600 samples. The nested validation techniques provided much more accurate performance estimates, better reflecting the large amount of noise-based features and the low effect size of the truly differing distributions. For a real-world problem, designers should therefore perform some initial exploration of the data to assess the quality of the features for classification. Data that does not demonstrate a reasonable level of discriminability in an early stage study (considering both the anticipated learning curve trajectory and the desired end stage performance) should likely not be pursued further.

The moderate discriminability features provide a more viable classification problem (Fig. 2(c)). In this case, however, overfitting was once again evident with the non-nested configurations. Despite their performance estimates more closely exemplifying the true accuracy (78.4%) than the nested configurations for sample sizes less than 200, the proportions of correctly selected features indicate that these estimates were based partially on noise-based features. The four nested configurations, on the other hand, had lower estimates that converged after only a few hundred samples to the true accuracy. This is the gradual improvement in

accuracy expected as the sample size, and therefore the model's ability to learn the appropriate features and class associations, is increased. Since the true underlying accuracy of a real problem is unknown, and cannot be estimated without a sufficient number of samples, it may be beneficial to implement both nested and non-nested configurations to provide a rough idea of the accuracy range when sample size is small. A non-nested configuration will almost certainly experience overfitting and may closely match or exceed the true accuracy, while a nested configuration will provide a conservative estimate that is generally lower than or equal to the true accuracy.

Notably, with the effect sizes of the features increased to a high discriminability case, all six of the validation configurations required fewer samples to converge to the true accuracy (92.4%, Fig. 2(e)). The improved discriminability enabled the models to better identify the correct features and exploit useful patterns, with each configuration reaching an average accuracy estimate within 2% of the true accuracy by 300 samples. With fewer than 300 samples, incidentally, all of the validation methods had pessimistic estimates. Even the non-nested configurations, which still showed evidence of overfitting at very low sample sizes (< 50 samples), did not provide estimates that exceeded the true accuracy as they had in the low and moderate discriminability cases. A real world dataset, however, may likely consist of features with diverse effect sizes. This condition was represented by the varying discriminability case in Fig. 2(g), which exhibited similar results to the moderate and high discriminability cases.

Regardless of the level of discriminability, certain trends were evident in the proportions of correctly selected features (Fig. 2(b), 2(d), 2(f), and 2(h)). The non-nested configurations, LOOCV and bootstrapping, generally yielded the highest percentage of correct features among all configurations, reflecting the increased statistical power afforded during the feature ranking by utilizing all available samples. However, this also allowed the model to find the most convenient noise-based features for the test samples, promoting overfitting and inflating performance estimates. Of the nested configurations, the nested LOOCV configuration, which uses all samples but one for feature selection, was the best at identifying the appropriate features. This was followed closely by the nested 10-CV and then holdout configurations, which used 90% and 80% of samples for feature selection, respectively. The nested bootstrapping configuration had the worst performance in this regard, presumably since it uses the fewest unique samples for model development and training (on average, 63.2% [2]) and repetitions in the bootstrap subsamples can emphasize noise-based patterns. This aspect of the bootstrap procedure explains the more conservative model performance estimates with the bootstrapping and nested bootstrapping configurations compared to their cross-validation counterparts.

While the level of discriminability of the features was varied, this work considered only a fixed number of features that were all drawn from unit standard deviation Gaussian distributions. Assorted feature distributions, correlated features, and highly disproportionate feature-to-sample size ratios are expected for many real-world biomedical problems, so a tailored simulation experiment using dataset-specific characteristics may be a valuable step for future implementations. The number of irrelevant features, in particular, has been shown to greatly disrupt the feature selection procedure and widen the gap between estimates from the nested and non-nested frameworks [10]. Further, there are several additional aspects of a classification model and its development that can affect performance and performance estimates. Though this work examined a handful of feature selection methods, the suitability of the selected algorithms and techniques for pre-processing, feature extraction, and classification can have a large impact. Future studies should incorporate a variety of methods in each of the classification stages to gain a greater understanding of the effect of these methods on model validation.

C. Learning Curve

Fig. 3 depicts the fitted learning curves for the averaged varying discriminability performance estimates from the results shown in Fig. 2. The non-nested configurations clearly did not fit well with the inverse power law model and showed the worst forecasting ability by greatly underestimating performance at higher sample sizes. The nested configurations performed much better in this regard. Nested bootstrapping, in particular, provided the best forecasting ability overall with a mean RMSE of 5.18%. This is, however, the most computationally intensive method, and its RMSE was not significantly improved ($p > 0.05$) compared to the less costly nested 10-CV and nested LOOCV. Though the holdout configuration was the least computationally intensive method, it exhibited the highest RMSE for projected performance of all the nested methods, reaffirming the necessity for more efficient sample use with small datasets. Hence, the nested CV configurations provided the best trade-off between forecasting ability and computation time.

While previous works examining learning curves for sample size planning have assumed fixed feature sets [9], [11], the protocol employed in this work incorporated feature selection to accommodate changes in the optimal feature set for the model as the sample size was varied. To further improve the flexibility of the learning curve procedure, future works should allow the number of selected features to vary at each sample size as well, since in practice, the required number will be unknown. Additionally, though standard techniques were adopted in this work, there are several other models and fitting procedures that could be adopted for learning curve analysis [12]. Certain methods may outperform others given the specific dataset, validation framework and set of utilized algorithms, so these factors should be further explored. Likewise, it would be beneficial to assess the presented techniques with real datasets.

IV. Conclusions

Ultimately, this work has shown that when sample size is sufficiently large, the selected model validation techniques

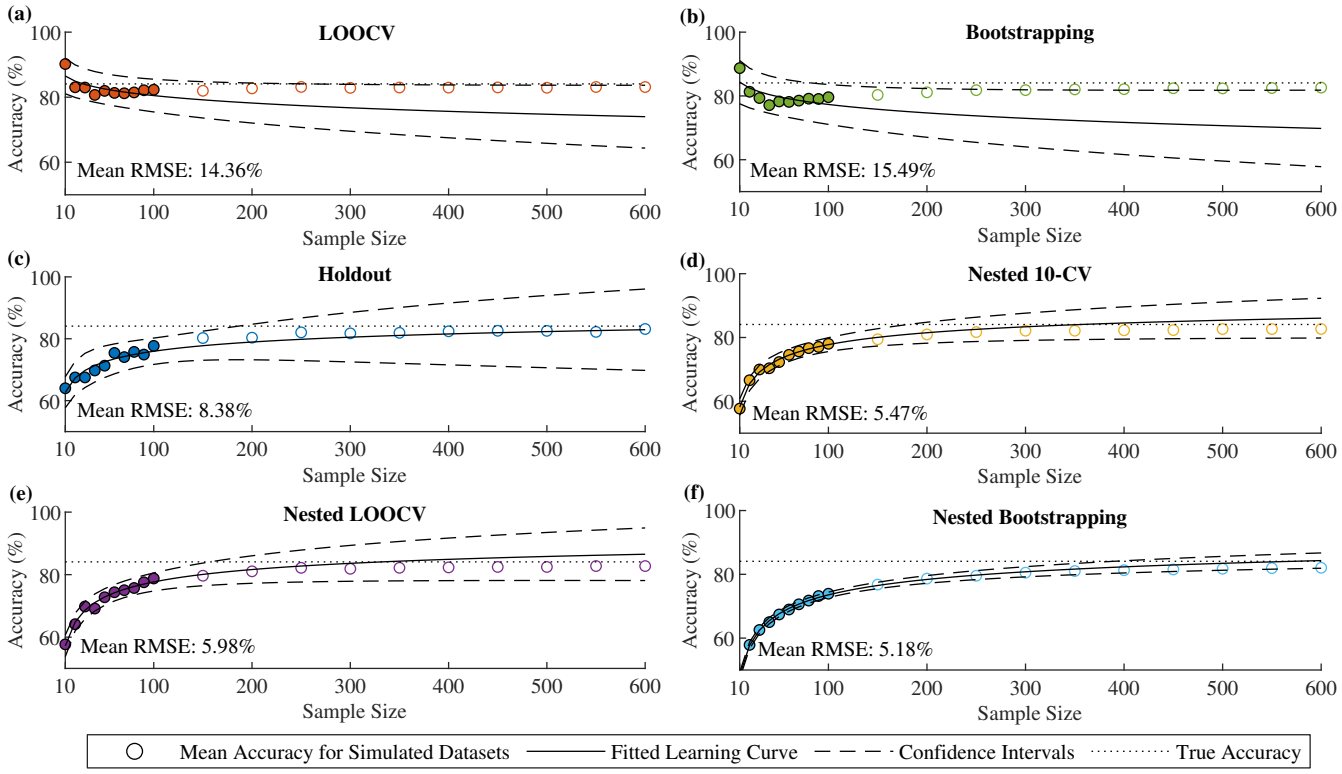

Fig. 3. Fitted learning curves for the averaged varying discriminability performance estimates using (a) LOOCV, (b) bootstrapping, (c) holdout, (d) nested 10-CV, (e) nested LOOCV, and (f) nested bootstrapping. For each of the 100 datasets, the first 10 points (filled) were used for curve fitting and the final 10 points (unfilled) were used for estimating RMSE. The presented RMSE values were averaged across all 100 datasets.

are largely inconsequential, with all techniques eventually converging to the same or similar results. When sample size is small, however, these methods have a much larger impact. Of all the compared validation configurations, the nested CV frameworks had the most success in reflecting the true accuracy of the selected features. The performance estimates were limited by the quality of the selected feature sets, however, which included a large proportion of noise-based features when sample sizes were low. Hence, simply applying nested CV to obtain a single performance estimate may not be adequate, and additional considerations are necessary. First, rather than wrapper-based techniques, filter-based feature selection techniques should be used for small sample sizes to avoid overfitting to irrelevant features. We also recommend that researchers perform an initial exploration of their early stage dataset to assess the level of discriminability of the features, which will aid in the interpretation of performance estimates and gauge the project's potential. Lastly, to estimate the maximum achievable accuracy of the problem and the sample size required to reach this accuracy, we suggest a learning curve fitting procedure using nested CV performance estimates. These recommendations should serve as a practical starting point for researchers performing small-scale feasibility studies and sample size planning.

References

[1] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larranaga, and J. Lozano, "Machine learning: An indispensable tool in bioinformatics," *Methods Mol Biol*, vol. 593, pp. 25–48, 2010.

[2] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer, and M. G. Sowa, "Variance reduction in estimating classification error using sparse datasets," *Chemom Intell Lab Syst*, vol. 79, pp. 91–100, 2005.

[3] R. G. Brereton, "Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data," *Trends Anal Chem*, vol. 25, no. 11, pp. 1103–1111, 2006.

[4] M. Hosseini, M. Powell, J. Collins, C. Callahan-flintoft, W. Jones, H. Bowman, and B. Wyble, "I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data," *Neurosci Biobehav Rev*, vol. 119, pp. 456–467, 2020.

[5] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J Mach Learn Res*, vol. 11, pp. 2079–2107, 2010.

[6] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, S. Huang, M. Brooks, M. J. Lee, and H. Asadi, "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods," *Am J Roentgenol*, vol. 212, pp. 38–43, 2019.

[7] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, pp. 1–20, 2019.

[8] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, 2005.

[9] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Med Inform Decis Mak*, vol. 12, 2012.

[10] G. Aldehim and W. Wang, "Determining appropriate approaches for using data in feature selection," *Int J Mach Learn Cybern*, vol. 8, no. 3, pp. 915–928, 2017.

[11] K. R. Hess and C. Wei, "Learning curves in classification with microarray data," *Semin Oncol*, vol. 37, no. 1, pp. 65–68, 2015.

[12] T. Viering and M. Loog, "The shape of learning curves: a review," 2021, arXiv preprint arXiv:2103.10948, March 2021.

Vita

Candidate's full name: Robyn Larracy

University attended: BScEE, University of New Brunswick, 2015-2019

Peer-Reviewed Publications:

1. Phinyomark, A., Larracy, R., and Scheme, E. "Fractal Analysis of Human Gait Variability via Stride Interval Time Series," *Frontiers in Physiology - Fractal and Network Physiology*, Vol. 11, No. 333, 2020.

2. Larracy, R., Phinyomark, A., and Scheme, E. "Data Pre-Processing of Infrared Spectral Breathprints for Lung Cancer Detection," *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Accepted, 2021.

3. Larracy, R., Phinyomark, A., and Scheme, E. "Machine Learning Model Validation for Early Stage Studies with Small Sample Sizes," *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Accepted, 2021.

4. Larracy, R., Phinyomark, A., and Scheme, E. "Laser Absorption Spectroscopy for the Detection of Lung Cancer Biomarkers in Exhaled Breath Samples," *Journal of Breath Research*, Under Review.

Conference Presentations & Industry Reports:

5. Larracy, R., Phinyomark, A., Reiman, T., Galvis, L., Abdesalam, M., Graham, S., Purves, C., Beydaghyan, G., and Scheme, E. "Identification of Lung Cancer Breath Biomarkers Using Infrared Cavity Ring-Down Spectroscopy," *Breath Biopsy Conference 2020*, Online, Nov 2020.

6. Phinyomark, A., Larracy, R., Scheme, E. "Machine Learning Methods for the Analysis of Lung Cancer from Breath Samples," *NBIF Innovation Voucher Report for Picomole Inc.*, Moncton, NB, Canada, July 2020