2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

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 ($p < 0.05$)

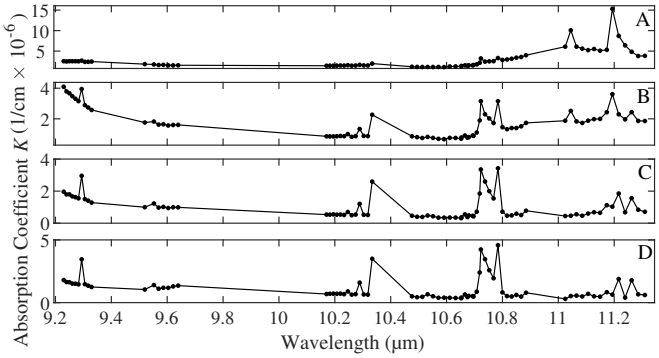

Fig. 1. Example spectra for one subject, desorbed at temperatures A) 75°C, B) 150°C, C) 225°C, D) 300°C.

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.