Convolutional neural networks for automated targeted analysis of raw gas chromatography–mass spectrometry data

To appear in: International Joint Conference on Neural Networks (IJCNN 2018), Rio de Janeiro, Brazil.

Angelika Skarysz*, Yaser Alkhalifah*, Kareen Darnley§, Michael Eddleston¶, Yang Hu*,
Duncan B McLaren||, William H Nailon||, Dahlia Salman‡, Martin Sykora†, C L Paul Thomas‡ and Andrea Soltoggio*

*Computer Science Department
School of Science
Loughborough University
Loughborough, UK

†Centre for Information Management
School of Business and Economics
Loughborough University
Loughborough, UK

‡Centre for Analytical Science
School of Science
Loughborough University
Loughborough, UK

§Clinical Research Facility
Western General Hospital
NHS Lothian
Edinburgh, UK

¶Pharmacology, Toxicology & Therapeutics Unit
University of Edinburgh
Edinburgh, UK

||Edinburgh Cancer Centre
NHS Lothian
Edinburgh, UK

Emails: {A.Skarysz, Y.Alkhalifah, Y.Hu, D.Salman, M.D.Sykora, C.L.P.Thomas, A.Soltoggio}@lboro.ac.uk
Kareen.Darnley@nhslothian.scot.nhs.uk, M.Eddleston@ed.ac.uk, Duncan.McLaren@nhslothian.scot.nhs.uk, W.Nailon@ed.ac.uk

***Abstract*—Through their breath, humans exhale hundreds of volatile organic compounds (VOCs) that can reveal pathologies, including many types of cancer at early stages. Gas chromatography–mass spectrometry (GC-MS) is an analytical method used to separate and detect compounds in the mixture contained in breath samples. The identification of VOCs is based on the recognition of their specific ion patterns in GC-MS data, which requires labour-intensive and time-consuming preprocessing and analysis by domain experts. This paper explores the original idea of applying supervised machine learning, and in particular convolutional neural networks (CNNs), to learn ion patterns directly from raw GC-MS data. The method adapts to machine specific characteristics, and once trained, can quickly analyse breath samples bypassing the time-consuming preprocessing phase. The CNN classification performance is compared to those of shallow neural networks and support vector machines. All considered machine learning tools achieved high accuracy in experiments with clinical data from participants. In particular, the CNN-based approach detected the lowest number of false positives. The results indicate that the proposed method is a promising tool to improve accuracy, specificity, and in particular speed in the detection of VOCs of interest in large-scale data analysis.**

I. Introduction

The typical human breath is estimated to carry over a thousand distinct volatile organic compounds (VOCs) [1]. These are the products of the metabolic processes that occur not only in the lung but, due to the blood-gas exchange, in the whole body. A breath sample contains information that describes physiological and pathological conditions, and thereby the health status of the patient [2]. Recently, the relationship among the changes in VOC patterns and different types of diseases, including breast, colorectal and lung cancers, have been presented in several research studies [3]–[5]. Breath analysis is thought to have a potential to provide a new non-invasive, fast and accurate diagnostic platform.

One of the leading analytical methods to detect VOCs in breath is gas chromatography–mass spectrometry (GC-MS). GC-MS data can be analysed in either a targeted or non-targeted approach. Non-targeted analysis is the study of all detected VOCs and their variability to discover potential biomarkers associated with specific disease. In targeted analysis, a defined panel of VOCs is sought to detect compounds of interest, e.g. known biomarkers [6].

GC-MS produces high dimensional, noisy data: one single sample can contain over 9 million high-resolution variables. For this reason, established data processing approaches employ preprocessing steps such as noise filtering, baseline correction, spectral deconvolution, and peak detection, which are necessary for the identification of VOCs for further multivariate statistical analysis [1], [7]. The result of preprocessing is a list of VOCs with their abundances. The data processing workflow requires analytical expertise and decisions on algorithmic parameter settings. The complexity of the GC-MS data, combined with variability in data processing, often leads to variations in the results. Additionally, GC-MS data process-

This study was partially funded by the EU H2020 TOXI-Triage Project #653409

ing requires about 90 minutes of an experienced analyser to process each breath sample.

The limitations outlined above call for better processing algorithms, now possible by exploiting recent advances in machine learning. Reported attempts to use machine learning on raw GC-MS data are limited. In one study, Shimizu et al. [8] used stacked autoencoders to classify GC-MS urine data from patients with lung cancer and controls, but provided limited evidence on the capability of their system to detect individual VOCs of interest, i.e. potential biomarkers, which are eventually essential for diagnosis. Other applications of machine learning in the area of metabolomics, such as [4], [9], [10], make use of preprocessed data, i.e. a list of VOCs with their abundances, that is time-consuming to obtain and might contain processing errors.

In this study, we propose the use of convolutional neural networks (CNNs) to learn to detect VOCs autonomously and directly from raw data, thereby bypassing the labour-intensive and time-consuming data preprocessing workflow. At first, we exploit expert knowledge to create a database of VOCs and their corresponding patterns, or fingerprints, as found in raw GC-MS data. Such a dataset of patterns is then used to train particular types of CNNs with one-dimensional convolutional filters specifically designed to learn from GC-MS data. Once the system is trained to recognise specific VOCs of interest, it can quickly scan breath data samples to automatically detect the target VOCs. Such a method has the potential to be significantly faster and more scalable than the current state-of-the-art manual procedures. In this study, CNNs are also compared with shallow neural networks and support vector machines. To the best of the authors' knowledge, this is the first study that proposes the use of CNNs to learn VOC-revealing ion patterns directly from raw GC-MS breath data.

The rest of this paper is organised as follows. Section II briefly explains the GC-MS process and data acquisition. Section III gives an overview of the machine learning tools used in this study. Section IV introduces the novel approach to the analysis of raw GC-MS data. Section V illustrates the specific clinical dataset used to train and test the systems. Section VI reports the experimental results. Section VII and VIII discuss and conclude the paper.

II. GC-MS Breath Data

GC-MS is a well-known analytical technology that is the gold standard for biomarkers discovery. The gas chromatography (GC) part of the instrumentation contains a capillary column (narrowed tube) lined with a specific material called stationary phase, where the separation of the compounds takes place. Various compounds interact with the stationary phase in different ways: this affects their time of elution from the column (*retention time (RT)*), and thus results in their separation [9]. The stationary phase degrades with use, and so the RT of VOC changes over the course of an extended GC-MS based campaign. Nevertheless, the elution of a particular compound can be expected in a certain range of time that is related to its chemical properties. Note though that the relationship of a VOC to the other VOCs eluting in the mixture remains constant.

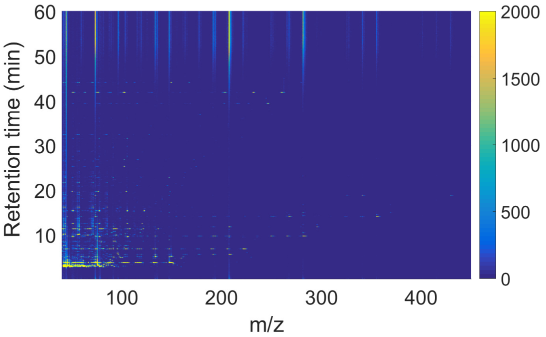

Fig. 1: GC-MS abundance matrix presented as a heat map. On the x-axis is the mass-to-charge ratio (*m/z*). The retention time (RT) is on the y-axis.

The separated compounds are then ionized and broken into fragments that are measured within the mass spectrometer (MS). Ion fragments are characterized by their *mass-to-charge* ratio, also indicated as *m/z*, where *m* is the atomic mass and *z* is the charge number of the ion. Different VOCs produce different ion fragmentation patterns; it is the pattern and the intensities of the fragments that enable to identify the VOC. For more details on the GC-MS process, we refer to [11], [12].

A. Data format

GC-MS produces a two-dimensional data matrix, also known as abundance matrix [13], as the one shown in Fig. 1. A more compact visualization, called chromatogram, is often used as shown in Fig. 2. The intensity on the y-axis, called total ion chromatogram (TIC) is the sum of the intensities across all *m/z* measured at the same time, i.e. at a specific retention time point (x-axis). Each peak generally represents one specific VOC, although superposition of peaks occasionally occurs [1].

B. Data processing

A quadrupole mass spectrometer, commonly used for GC-MS, produces unit resolution *m/z* ratios and has a dimension of 411 along the *m/z* axis. The second dimension is the RT, measured in the number of MS scans, which is approximately 22 500, corresponding to approximately one hour, i.e. the time for all compounds to elute from the GC column. The instrumental scanning rate is approximately 6 Hz. The values in the abundance matrix are affected by instrument and environment-related noise [14]. However, even the smallest value may carry significant information on the individual metabolomics status [15].

Several studies have discussed GC-MS data preprocessing strategies and relevant methods. Smolinska et al. [1] highlighted that data preprocessing is critical to obtain reliable high-level data, i.e. VOCs and their quantities. Trygg et al. [16] summarized several methods to isolate noise and signal. Various automatic baseline correction techniques, mostly based

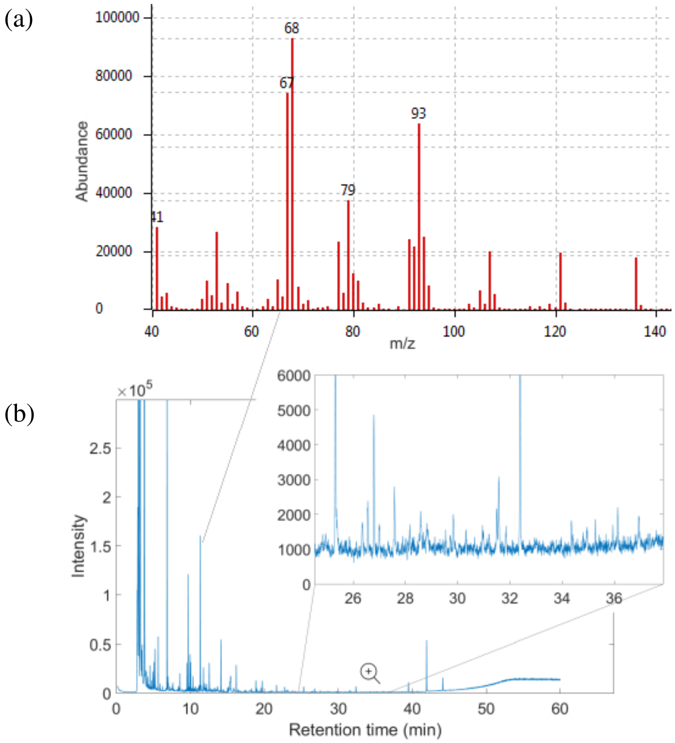

Fig. 2: An example of a breath GC-MS chromatogram. (a) A full mass spectrum corresponds to each point in retention time. (b) The total ion current is plotted along the retention time dimension.

on polynomial fitting, were also proposed in the literature [17]–[19]. Jellema [20] discussed the most common alignment methods used to deal with the fluctuations in retention time across measurements. A common next step of GC-MS data preprocessing is peaks detection by finding all local extrema in TIC chromatogram. Only peaks with signal-to-noise ratio larger than a threshold are taken into account in further analysis [1]. Each peak in the chromatogram is the sum of intensities of all *m/z*, as in Fig. 2(a). The pattern of such *m/z* values can be seen as the fingerprint of a particular compound: the analysis aims to find the closest match to known compounds from a library. Given the complexity of the process that generates these patterns and various sources of noise, the matching is based on a number of ad-hoc rules as detailed in [13]. Several software tools to assist this process have been developed. Details on selected freely available tools and their comparison can be found in Du and Zeisel [21].

The list of VOCs and their abundances is an input required for a further multivariate statistical analysis, in which the objective is to identify VOCs that discriminate between different metabolic derangements of the patients.

The number of steps and the complexity of the process described above may lead to some challenges. Liland et al. [22] noted that typically the choice of the best algorithm to baseline identification and the choice of parameters are based on the visual inspection of the selected chromatogram, which is a highly subjective and laborious task. Furthermore, Coombes et al. [12] highlighted that the preprocessing has the potential to introduce errors and variability.

Nevertheless, ion patterns derived from specific compounds, although noisy, present unique features that distinguish them. The original idea in this study is that such features can be learned using advanced machine learning techniques, such as CNNs, directly from raw GC-MS data and therefore bypassing highly complex preprocessing step. The rest of the paper focuses on this idea.

III. Machine learning tools

This section overviews the machine learning methods that are later applied to GC-MS raw data. The hypothesis is that such methods can be employed to learn ion patterns along the *m/z* dimension to classify compounds directly from raw data. In this study, we propose the use of convolutional neural networks (CNNs), shallow neural networks (NNs) and support vector machines (SVMs) to learn patterns directly from the abundance matrix.

Deep learning techniques, especially convolutional neural networks (CNNs) [23] have demonstrated excellent performance in image recognition and classification tasks [24]–[29]. CNNs can autonomously learn useful features directly from low-level data, e.g. pixels [30], and construct high-level features without human intervention. CNNs can also exploit geometrical properties of the data and are less affected by noise with respect to other techniques [31]. Additionally, an increase of GC-MS as a diagnostic technology will see also an increase of available datasets: a large amount of data is known to benefit the training of deep neural networks [30]. The use of GPU computing and dedicated hardware, which has seen a rapid development in recent years, can help process the large amount of data collected through GC-MS.

Traditional shallow neural networks (NNs) are widely used to learn a mapping from input to output both for classification and regression tasks. A shallow neural network consists of an input layer of neurons, a hidden layer of neurons, and a final layer of output neurons. In this respect, a NN is a much simpler structure than a CNN, but often effective on simple problems. NNs are known for their ability to learn patterns, and for this reason were chosen here as a method for comparison with CNNs. Given the popularity of NNs in machine learning, we omit further general notions and refer to the literature for an overview of neural networks for classification [32].

Support vector machines (SVMs) are widely and successfully used in classification tasks [33], [34]. SVMs try to find, in the optimization process, the hyperplane separating the instances of different classes with maximal margin (maximum margin classification). Furthermore, to handle complex data which may be non-linearly separable, a nonlinear transformation $\Phi : \mathbb{R}^d \to \mathcal{F}$ can be applied along with a kernel function. One limitation of SVMs is the difficulty in determining the best kernel of SVM, whose performance depends on the type of data. Nevertheless, given their success, SVMs were also selected to compare their performance with those of CNNs and NNs on GC-MS data.

Despite the popularity of the methods described above and their achievements in various fields, their applications in the area of metabolomics are limited and almost exclusively concern preprocessed data.

IV. Learning ion patterns and detecting compounds from raw GC-MS data

The novel idea in this paper is to exploit the pattern recognition ability of NNs, SVMs, and in particular CNNs, to learn to recognise ion patterns directly from raw GC-MS data. A recognised ion pattern is effectively a recognised VOC, which in turn could be a biomarker of a given physical condition. Once a learning algorithm can recognise ion patterns, entire breath sample datasets can be scanned quickly and automatically to search for compounds of interest. While the machine learning methods that we used are well established, their precise configurations and experimental designs for the application to raw GC-MS breath data are investigated for the first time in this study.

To realize the above, breath samples are firstly analysed with current methods to identify compounds with the help of chemist expertise [1], [7]. Subsequently, using identified compounds and their positions on the data matrix, segments of raw data are extracted and labelled with the compounds they contain. Then, the labelled dataset is used to train the machine learning systems. Finally, the trained models are used to scan entire breath data samples to identify target compounds.

A. CG-MS data preparation for machine learning

Due to the specific structure of the data, the pattern of each VOC is contained only in a small portion of the abundance matrix, corresponding to a specific range of retention time. Applying current methods [13], and with the supervision of experts, the exact retention time for each target compound, and its classification, are determined. This process generates a dataset of labelled VOCs and their corresponding positions in the matrix of raw data. To link processed with raw data, a dataset structure is created with the following fields: $BreathSample$, $compoundClass$, $startRT$, $peakRT$, $endRT$. $BreathSample$ is the name of the file that contains one sample; $compoundClass$ is the name of the compound found in that sample, e.g. Octane; $startRT$, $peakRT$ and $endRT$ are the indexes along the retention time where the compound was measured to start, peak and end the release from the GC column.

1) Input format and transformation: Each submatrix containing the pattern of the target compound is extracted from the abundance matrix and stored as a *segment* S of dimension $(\delta, 411)$; δ is the width of the segment of the abundance matrix, computed so that $\delta \geq endRT - startRT$, i.e. the segment's width is sufficiently large to include a VOC's entire peak. A segment S so devised is an appropriate 2D input for the CNN. For the NN and SVM, the dimension along the retention time is integrated to one value using a Gaussian filter centred in the middle of the segment, resulting in a 1×411 vector $v = wS$, where w is a vector of Gaussian filter coefficients $(1 \times \delta)$. The coefficients of a Gaussian filter are computed as $w(n) = \exp\left(-n^2/2\sigma^2\right)$ where $n \in \left\{\frac{-\delta+1}{2}, \frac{-\delta+3}{2}, \ldots, \frac{\delta-1}{2}\right\}$, $\sigma = \frac{\delta-1}{2\alpha}$ is a standard deviation of a Gaussian random variable and $\alpha = 2.5$. This procedure results in the approximation of a 2D segment to a 1D mass spectrum (Fig. 2a) without peak detection, i.e. directly from raw data. Thus, NN and SVM receive segments S as 1D inputs that capture the values of *m/z* intensities at specific points along the retention time.

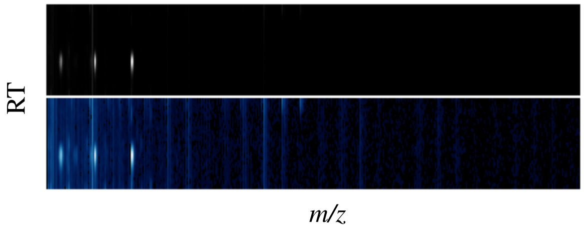

Fig. 3: Image of a segment from an abundance matrix rendered as a one-channel input (top, gray scale) and three-channel input, visualised as RGB values (bottom). More information appear visible in the three-channel format.

2) Three-channel input mapping: The abundance matrix has typically large differences in its values: in the dataset used for this study, the range was [0, 1.6e6] with a mean value of only 24.21. This is because few compounds may be present in large quantities, but other compounds in small quantities are nevertheless relevant in the analysis, see e.g. Fig. 2(b). Thus, after normalisation of the input data, most of the values might be too small to be captured. To overcome this potential problem, we compare two input types: a single-channel input and a three-channel input. For the single-channel input, all input values are normalised in the range [0, 1]. For the three-channel input, additionally, after normalisation, each value x is mapped to three separate values by the function $\mathbf{y} = f(x) = \left[x, x^{\alpha}, x^{\beta}\right]$ with $1 > \alpha > \beta > 0$. Visual inspection of the GC-MS raw matrix data (see Fig. 3) suggested that $\alpha = 0.4, \beta = 0.2$ resulted in visible peaks of high, medium and low intensity; this study only assesses whether a multi-channel input has advantages in classification: further studies can focus on the optimisation of the parameters α and β.

B. Models

This section explains the precise implementations of CNN, SVM and NN to process raw GC-MS data.

1) Implementation for the Convolutional Neural Network: Filters, which are the local receptive fields in the convolutional and pooling layers of the network, exploit the geometrical spatial correlations in the data. With GC-MS data, such a geometrical correlation occurs only in the retention time dimension. Along this dimension, the abundance of different *m/z* increases and decreases thereby creating peaks as the compounds exit the column. On the other hand, the abundance values across different *m/z* channels correlate as the particular ions make up the compounds. Thus, the patterns of abundance along the *m/z* dimension are not local, and cannot be captured by small local filters. Given such property of the data, one hypothesis in this study is that local receptive fields need not be two dimensional as in usual computer vision applications. To test the hypothesis, two types of filters are used: two-dimensional filters, and specific one-dimensional filters along the RT axis only. In the case of two-dimensional filters, sizes were set to (3,3) and (2,2) for convolution and pooling layers. In the case of one-dimensional filters, sizes were set to (3,1) and (2,1).

A VGG-like network architecture [35] was chosen to stack multiple convolutional layers with ReLU activation before pooling layers. Three variants of the architecture were tested in preliminary experiments; each was based on the same four-layered block consisted of two convolutional layers, pooling layer and dropout layer with rate 0.25. The tested architectures were built of respectively one, two and three such blocks, followed by a fully connected layer with ReLU activation, dropout layer with rate 0.5 and the fully connected layer with softmax activation. The batch size and the number of epochs were set up as 128 and 10 respectively. The results of these preliminary tests gave similar performance among these three architectures. The network with the best performance resulted to be the smallest with two convolutional layers. This architecture was selected for the experiments in the rest of the paper. We recognise that further, more thorough investigations on the types of architectures and their parameters are interesting future research directions.

2) *Implementation for the Shallow Neural Network:* For the implementation in this study, we use the standard MATLAB NN toolbox that allowed us to set up a standard three-layer feed-forward network, with sigmoid activation function in the hidden layer and softmax activation function for output layer. The size of hidden layer was set up as 10. The network was trained with scaled conjugate gradient backpropagation.

3) *Implementation for the Support Vector Machine:* To investigate the best performance of SVMs, we tested three different kernels: linear $K(u,v) = u \cdot v$, polynomial $K(u,v) = (\gamma u \cdot v + \delta)^s$, Gaussian (radial basis function, RBF) $K(u,v) = \exp\left(-\|u-v\|^2/2\sigma^2\right)$. We applied grid search with cross-validation for parameters setting. We used a popular library for support vector machines LIBSVM with MATLAB interface [36].

V. Materials

The data used in this study was obtained from 11 participants with different types of cancer receiving radiotherapy. Four breath samples were collected from each participant: prior to radiation, 1, 3, and 6 hours after radiation. The process was not completed for three of the 44 planned samples, and the size of the final dataset is 41. The target compounds in this study are aldehydes because they have been reported in the literature as cancer-related compounds [3]. The set contains 8 aldehydes: Benzaldehyde, Benzeneacataldehyde, Decanal, Furfural, Heptanal, Hexanal, Nonanal and Octanal. The RT location of those compounds was derived from the labelled data processed by experts as described in section II.

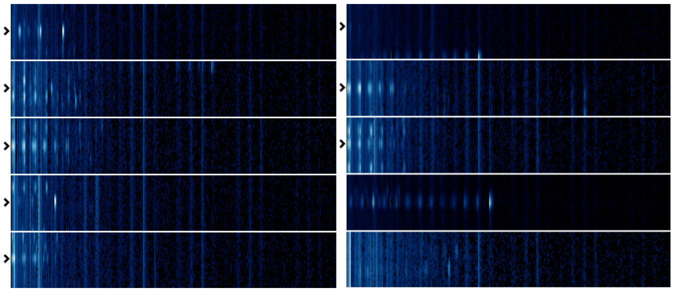

Fig. 4: Examples of segments of aldehydes patterns extracted from the raw GC-MS abundance matrix. Left, from top: Benzaldehyde, Benzeneacataldehyde, Decanal, Furfural, Heptanal; right, from top: Hexanal, Nonanal, Octanal and two examples from the negative group. The black arrows (at the left) indicate the precise points along the RT axis where the target compounds appear in the segments.

A. Extraction of segments of compounds patterns from the abundance matrix

To ensure that segments contain the entire shape of all peaks, the value max($startRT$ - $endRT$) for eight target compounds was measured as 42, and was taken as the segment's width δ for NN and SVM. Such segment's width corresponds to approximately 7 seconds along the RT dimension. A larger segment with $\delta = 70$, ~ 12 seconds, was selected for CNNs to facilitate data augmentation (described below) that involved shifts along the retention time. Besides the labelled compounds, a negative class was created by randomly selecting segments along the RT dimension that did not correspond to any target compound. Fig. 4 shows examples of segments for each of the eight target compounds, and two negative examples.

B. Data augmentation

To increase the robustness of the training and compensate for the limited number of data points, data augmentation was applied. The abundance values were increased by random value in the range 0.01 to 9.99%. Such an augmentation changes the absolute values of *m/z* intensities without changing their proportion, i.e. the pattern. This augmentation step was repeated to obtain four additional data points for each segment. Additionally, for CNN, the 2D segment was augmented 20 times by shifting it along the RT dimension, from -9 to +10 pixels. Therefore, the dataset for CNN was augmented 100 times, while the dataset for SVM and NN was augmented 5 times.

C. Train and test sets

The breath samples dataset was randomly divided into train and test set in the proportion 29/12 according to participants: all breath samples derived from the same participant are in the same set. The train set contained breath samples collected from

8 participants (29 breath samples), while the test set contained samples from 3 participants (12 samples). The segments of target compounds and the negative examples were extracted from the abundance matrices from train and test sets, giving respectively the train and test datasets of aldehydes segments for machine learning models. After augmentation, the train set for SVMs and NNs consisted of 1680 segments and the test set of 720 segments. For CNNs, the train set consisted of 33600 segments and the test set of 14400 segments. The exact sizes of the groups in the train and test sets are listed in the Table I. The inequality of the groups (unbalanced dataset) results from the fact that each of the considered aldehydes does not necessarily occur in each of the considered breath samples.

TABLE I: Number of segments (or datapoints) in the dataset of aldehydes

		NNs and SVMs		CNNs	
Label	Compound	Train size	Test size	Train size	Test size
0	Negative group	840	360	16800	7200
1	Decanal	100	55	2000	1100
2	Nonanal	85	30	1700	600
3	Benzeneacataldehyde	85	40	1700	800
4	Octanal	80	25	1600	500
5	Benzaldehyde	145	60	2900	1200
6	Heptanal	90	40	1800	800
7	Furfural	140	55	2800	1100
8	Hexanal	115	55	2300	1100
		1680	**720**	**33600**	**14400**

VI. Experimental results

This section presents the results of training and testing of the CNN, NN and SVM models on the clinical dataset. All systems were run on a server running Linux Ubuntu with 20 cores, 128GB RAM and NVIDIA Tesla K80 GPU cards.

A. *Evaluation of the models on the segment test set*

Table II presents the class-wise accuracy achieved on the segment test set by the CNNs with single-channel and three-channel inputs with both one-dimensional and two-dimensional convolutional and pooling filters. Confirming our hypotheses, the best performance was achieved by the CNN with one-dimensional filters with a three-channel input.

TABLE II: Class-wise accuracy on the test set for the CNN.

	Class label								
CNN	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1ch-1[a]	99.35	99.55	100	100	99	100	96.88	100	97.09
1ch-2[b]	99.26	99.55	100	100	97	100	91.5	100	95.73
3ch-1[c]	**99.69**	**100**	**100**	**100**	**100**	**100**	**97.25**	**100**	**99.09**
3ch-2[d]	99.39	100	100	100	96	100	95.63	100	97.73

[a] 1ch-1: one-channel input, 1D filters. [b] 1ch-2: one-channel input, 2D filters. [c] 3ch-1: three-channel input, 1D filters. [d] 3ch-2: three-channel input, 2D filters.

The class-wise accuracy achieved on the test set by the shallow neural network with one-channel and three-channel inputs is presented in the Table III.

TABLE III: Class-wise accuracy on the test set for NN

	Class label								
NN	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1ch[a]	96.94	100	100	75	100	100	75	100	100
3ch[b]	**97.78**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

[a] 1ch: one-channel input. [b] 3ch: three-channel input

The performance of the SVM with linear, polynomial and RBF kernels was evaluated on the test set with both one-channel and three-channel input. Table IV shows the comparison of class-wise accuracy achieved by these methods. The highest accuracy was achieved with SVM with the polynomial kernel on the three-channel input. As with CNNs and NNs, the three-channel input seems to lead to an advantage with respect to the one-channel input, as suggested by the performance of all three models.

TABLE IV: Class-wise accuracy on the test set for SVM

	Class label								
SVM	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0-1ch[a]	96.67	100	100	100	100	100	100	100	100
0-3ch[b]	99.17	100	100	100	100	100	100	100	100
1-1ch[c]	97.5	100	100	100	100	100	100	100	90.91
1-3ch[d]	**99.72**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
2-1ch[e]	100	90.91	100	62.5	60	100	50	100	54.55
2-3ch[f]	100	100	100	87.5	100	100	75	100	81.82

[a] 0-1ch: linear kernel, one-channel input. [b] 0-3ch: linear kernel, three-channel input. [c] 1-1ch: polynomial kernel, one-channel input. [d] 1-3ch: polynomial kernel, three-channel input. [e] 2-1ch: RBF kernel, one-channel input. [f] 2-3ch: RBF kernel, three-channel input.

B. *Detecting compounds on entire breath samples*

In practical applications, the entire breath sample requires being analysed to detect whether it contains the target compounds. Thus, the best performing configurations of the CNN, NN and SVM were selected to analyse the entire samples. The procedure involves using as input a segment of the sample at a time, and repeat the operation to scan the entire breath sample along the retention time dimension. The scan of one sample involves over 22500 evaluations for each model (the dimension of the retention time axis). Therefore, the models were extensively validated also on the entire breath samples from the test set.

The results of the scans on the 12 breath samples from the test set are presented in Table V. While all methods achieve 100% sensitivity, CNN reported the lowest number of false positives. A more precise analysis of CNN scans revealed that certain compounds such as Hexanal and Benzaldehyde had no false positives. Other compounds such as Decanal and Furfural were falsely detected more than others.

Interestingly, it was observed that a number of false positives were detected at particular retention times: those RTs correspond to the positions where such VOCs (true positives) can be expected. A more detailed analysis revealed that some of such false positives were indeed true positives. The re-assessment of the ground truth, i.e. whether an ion pattern is or is not a particular compound, requires intense expert-driven analysis with no guarantee of certainty. Therefore, rather than re-evaluate the ground truth, it was decided to present results as true positives and false positives divided as *certain* and *uncertain*. Uncertain false positives occur within the time

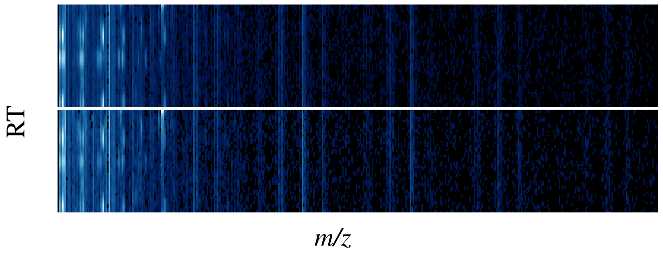

Fig. 5: Certain, i.e. true positive (top) and uncertain, i.e. false positive (bottom) Octane patterns. The similarity in the pattern, and the identical location of such a pattern along the RT axis suggests a high probability of an error in the ground truth.

window where such a target compound was observed in other samples: because the overall false positive rate is low across the entire scan, we can infer that there is a high probability that an uncertain false positive (i.e. occurring within a precise and restricted time window along the RT) is actually a true positive. An example of a comparison between an uncertain false positive and a true positive for the compound Octane is shown in Fig. 5.

TABLE V: Target compounds detection across entire samples. The true positive rate (TPR, the number of target compounds identified in the scan over the number of target compounds in the ground truth) is shown. For false positives, certain and uncertain classifications are shown.

Breath sample	**True positives rate (TPR)**			**False positives certain(uncertain)**		
	CNN	**SVM**	**NN**	**CNN**	**SVM**	**NN**
1	7/7	7/7	7/7	1(1)	6(2)	18(2)
2	5/5	5/5	5/5	1(3)	5(4)	9(4)
3	8/8	8/8	8/8	1(0)	9(0)	19(0)
4	7/7	7/7	7/7	2(1)	11(2)	15(2)
5	7/7	7/7	7/7	4(1)	16(2)	26(1)
6	5/5	5/5	5/5	3(2)	13(2)	30(1)
7	8/8	8/8	8/8	3(0)	13(0)	27(0)
8	6/6	6/6	6/6	3(3)	16(3)	21(2)
9	5/5	5/5	5/5	3(2)	32(3)	28(3)
10	5/5	5/5	5/5	0(3)	10(4)	20(4)
11	5/5	5/5	5/5	0(3)	7(4)	16(5)
12	4/4	4/4	4/4	0(5)	10(5)	34(7)
	1	**1**	**1**	**21(23)**	**146(27)**	**263(31)**

VII. Discussion

The results presented in the previous section reveal that all three machine learning models achieve 100% sensitivity in detecting target compounds in breath samples, demonstrating that the ion patterns can be effectively learnt, from both 2D inputs directly from the raw GC-MS data, and from 1D inputs extracted from the raw data with a Gaussian filter.

These results are encouraging, but become useful only when a high specificity, i.e. low false positives, is also observed. While NN and SVM detected a high number of false positives, CNN performed considerably better in the scans of the entire samples. We expect that NN and SVM will perform better if more preprocessing is applied to the data; this study, however, was focused on raw GC-MS data analysis. The implication is that CNNs are the most promising tool for further studies. In particular, the CNN architecture, including the depth, the alternation of different layers, and the size of the filters can be further investigated to improve the ability to learn and detect target compounds. The current study does not investigate in detail the reason for the higher performance of the CNN with respect to NN and SVM. We speculate that the increased performance might derive from the following reasons. The reading of the 2D matrix and the use of convolutional filters may allow the CNN to learn the shape of the peaks, and not only the relative abundance along the *m/z* axis. In fact, different compounds might exit the GC column at a different speed, resulting in differently shaped peaks. A second reason may be that such filters allow for more robustness in front of low signal-to-noise ratio, in particular ignoring constant levels of noise (column leaks) visible as continuous vertical lines in Fig. 4. Finally, the depth of the CNN may allow for higher level representations of the low-level features that characterise each individual pattern.

The higher performance of the models with the three-channel input confirms the intuition that patterns are revealed by both high and low-intensity signals. A three-channel input may increase the "visual acuity" of the models, resulting in better performance. Although the visual inspection of patterns via a three-channel RGB image appears to confirm the hypothesis, this study shows the critical role played by high and low signals in GC-MS data for machine learning systems.

An interesting result is that the system seems to discover errors in the ground truth. Although their rate was not quantified in this study, their possibility is not surprising because the ground truth is derived from complex human-dependent preprocessing steps. Thus, even as a first documented attempt to learn ion patterns from raw GC-MS data, the proposed systems already demonstrate state-of-the-art performance capable of improving human-guided established methodologies.

One limitation is that the current study considers only eight target compounds. As more compounds are added, chances of misclassification may increase, particularly for ion patterns with similarities. Future studies may assess the approach on a larger set of target compounds.

Finally, the current approach is limited to detecting the presence of compounds, but not their abundance. However, in real life scenarios, further analysis by experts to determine abundance may become necessary only if the system reveals the presence of specific compounds, e.g. cancer biomarkers. Further studies can extend the proposed approach to also measure intensities. Further work on the presented approach may consider a larger breath sample dataset with a higher number of the targeted compounds. A comparison of different neural network architectures to improve performance is also a promising future research direction.

VIII. Conclusions

Machine learning, and in particular convolutional neural networks, were applied to learn and detect volatile organic compounds directly from raw GC-MS data. Due to the high variability, noise, and high dimensionality of GC-MS data, the application of machine learning presents considerable challenges: to the best of the authors' knowledge, this is the first successful machine learning attempt at learning ion patterns and detecting compounds from raw GC-MS data. The complex and noisy patterns present in GC-MS data, derived from breath samples and collected in clinical trials, were used to train convolutional neural networks, shallow neural networks, and support vector machines. The convolutional neural network achieved the best performance when implemented with two particular features: one-dimensional filters to adapt to the particular structure of GC-MS data, and a three-channel input to read high, medium, and low-intensity signals from the highly variable GC-MS spectrum. The novel approach was shown to discover labelling errors from human experts, suggesting better-than-human average performance. Additionally, the proposed methodology can be used to speed up diagnostic processes, remove result variability arising from current methods, and learn from increasingly large amount of samples. The proposed approach has the potential to significantly contribute to the development of a diagnostic platform to detect various diseases quickly, efficiently, and reliably.

References

[1] A. Smolinska, A. C. Hauschild, R. R. Fijten, J. W. Dallinga, J. Baumbach, and F. J. van Schooten, "Current breathomics-a review on data pre-processing techniques and machine learning in metabolomics breath analysis," *Journal of Breath Research*, vol. 8, no. 2, p. 27105, 2014.

[2] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: Acquiring and understanding global metabolite data," 2004.

[3] P. Fuchs, C. Loeseken, J. K. Schubert, and W. Miekisch, "Breath gas aldehydes as biomarkers of lung cancer," *International Journal of Cancer*, vol. 126, no. 11, pp. 2663–2670, 2010.

[4] D. F. Altomare, M. Di Lena, F. Porcelli, L. Trizio, E. Travaglio, M. Tutino, S. Dragonieri, V. Memeo, and G. de Gennaro, "Exhaled volatile organic compounds identify patients with colorectal cancer," *British Journal of Surgery*, vol. 100, pp. 144–150, 2012.

[5] M. Phillips, R. N. Cataneo, C. Saunders, P. Hope, P. Schmitt, and J. Wai, "Volatile biomarkers in the breath of women with breast cancer.," *Journal of breath research*, vol. 4, no. 2, p. 026003, 2010.

[6] K. Hiller, J. Hangebrauk, C. J??ger, J. Spura, K. Schreiber, and D. Schomburg, "Metabolite detector: Comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis," *Analytical Chemistry*, vol. 81, no. 9, pp. 3429–3439, 2009.

[7] S. Ren, A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu, "Computational and statistical analysis of metabolomics data," 2015.

[8] R. Shimizu, S. Yanagawa, Y. Monde, H. Yamagishi, M. Hamada, T. Shimizu, and T. Kuroda, "Deep learning application trial to lung cancer diagnosis for medical sensor systems," *2016 International SoC Design Conference (ISOCC)*, pp. 191–192, 2016.

[9] J. J. B. N. Van Berkel, J. W. Dallinga, G. M. Möller, R. W. L. Godschalk, E. J. Moonen, E. F. M. Wouters, and F. J. Van Schooten, "A profile of volatile organic compounds in breath discriminates COPD patients from controls," *Respiratory Medicine*, vol. 104, no. 4, pp. 557–563, 2010.

[10] A. Baranska, E. Tigchelaar, A. Smolinska, J. W. Dallinga, E. J. Moonen, J. A. Dekens, C. Wijmenga, A. Zhernakova, and F. J. Van Schooten, "Profile of volatile organic compounds in exhaled breath changes as a result of gluten-free diet," *Journal of Breath Research*, vol. 7, no. 3, 2013.

[11] Grob L Robert and Barry F Eugene, *Modern Practice of Gas Chromatography, Fourth Edition.* 2004.

[12] K. R. Coombes, K. A. Baggerly, and J. S. Morris, "Pre-processing mass spectrometry data," in *Fundamentals of Data Mining in Genomics and Proteomics*, pp. 79–102, 2007.

[13] S. E. Stein, "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data," *Journal of the American Society for Mass Spectrometry*, vol. 10, no. 8, pp. 770–781, 1999.

[14] A. B. Fialkov, U. Steiner, S. J. Lehotay, and A. Amirav, "Sensitivity and noise in GC-MS: Achieving low limits of detection for difficult analytes," *International Journal of Mass Spectrometry*, vol. 260, no. 1, pp. 31–48, 2007.

[15] G. A. N. Gowda, S. Zhang, H. Gu, V. Asiago, N. Shanaiah, and D. Raftery, "Metabolomics-based methods for early disease diagnostics.," *Expert review of molecular diagnostics*, vol. 8, no. 5, pp. 617–33, 2008.

[16] J. Trygg, J. Gabrielsson, and T. Lundstedt, "Background Estimation, Denoising, and Preprocessing," in *Comprehensive Chemometrics*, vol. 2, pp. 1–8, 2010.

[17] Y. Xi and D. M. Rocke, "Baseline correction for NMR spectroscopic metabolomics data analysis.," *BMC bioinformatics*, vol. 9, no. 1, p. 324, 2008.

[18] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with B -splines and penalties," *Statistical Science*, vol. 11, no. 2, pp. 89–121, 1996.

[19] P. H. C. Eilers, "A perfect smoother," *Analytical Chemistry*, vol. 75, no. 14, pp. 3631–3636, 2003.

[20] R. H. Jellema, "Variable Shift and Alignment," in *Comprehensive Chemometrics*, vol. 2, pp. 85–108, 2010.

[21] X. Du and S. H. Zeisel, "Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives.," *Computational and structural biotechnology journal*, vol. 4, no. January, pp. 1–10, 2013.

[22] K. H. Liland, T. Almøy, and B. H. Mevik, "Optimal choice of baseline correction for multivariate calibration of spectra," *Applied Spectroscopy*, vol. 64, no. 9, pp. 1007–1016, 2010.

[23] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," 1989.

[24] Y. LeCun, F. J. H. F. J. Huang, and L. Bottou, "Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting," *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–97 – 104, 2004.

[25] D. Cirean, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *International Conference of Pattern Recognition*, no. February, pp. 3642–3649, 2012.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Nips*, pp. 1–9, 2012.

[27] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," *Proceedings of International Conference on Pattern Recognition ICPR12*, no. Icpr, pp. 10–13, 2012.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, p. 3, 2017.

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," *arXiv preprint arXiv*, p. 1312.6229, 2013.

[31] M. Nielsen, "Neural Networks and Deep Learning."

[32] G. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[33] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] V. Vapnik, S. E. Golowich, and A. Smola, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing," in *Advances in Neural Information Processing Systems 9*, pp. 281—-287, 1997.

[35] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, pp. 1–13, 2014.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines,"