A machine learning approach to automated targeted analysis of raw gas chromatography-mass spectrometry data

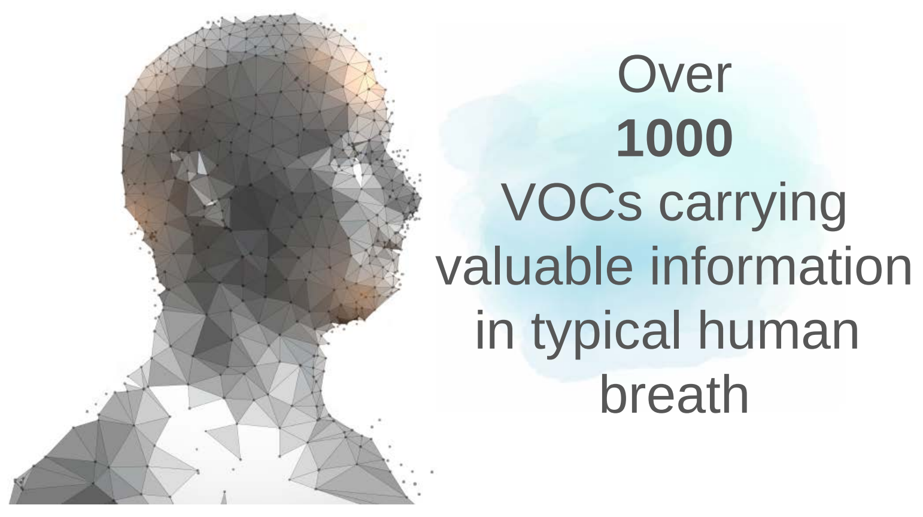
Angelika Skarysz[1], Yaser Alkhalifah[1], Kareen Darnley[2], Michael Eddleston[3], Yang Hu[1], Duncan B McLaren[2], William H Nailon[2], Dahlia Salman[1], Martin Sykora[1], C L Paul Thomas[1] and Andrea Soltoggio[1]

[1] Loughborough University, Loughborough, UK [2]NHS Lothian, Edinburgh, UK [3]University of Edinburgh, Edinburgh, UK

e-mail: a.skarysz@lboro.ac.uk

INTRODUCTION

BREATH AS A DIAGNOSTIC PLATFORM


Over **1000** VOCs carrying valuable information in typical human breath

Volatile organic compounds (VOCs) are the products of the metabolic processes occurring in the body.

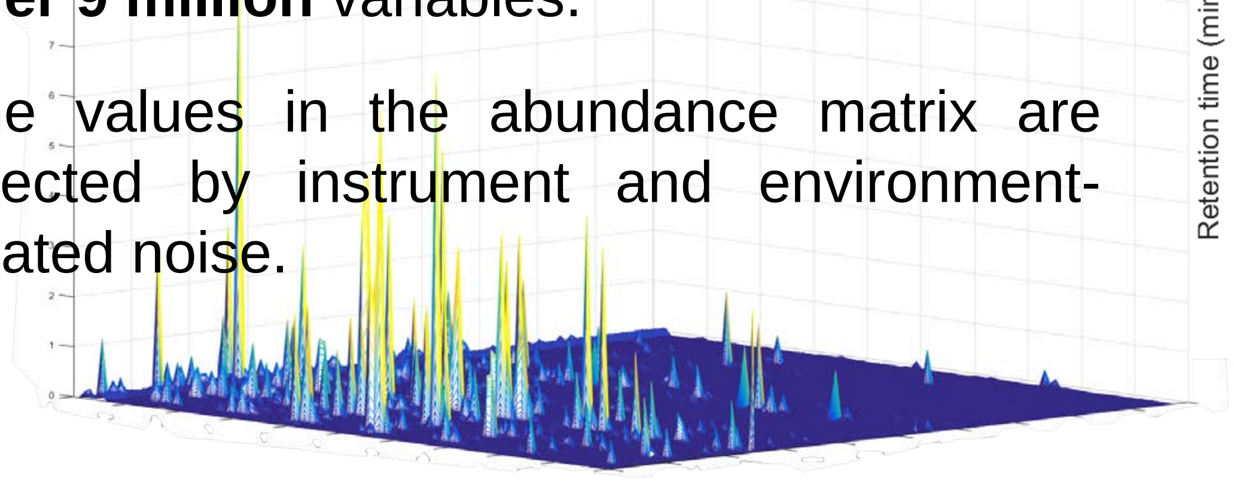
Breath analysis is thought to have the potential to provide a non-invasive, fast and accurate diagnostic platform.

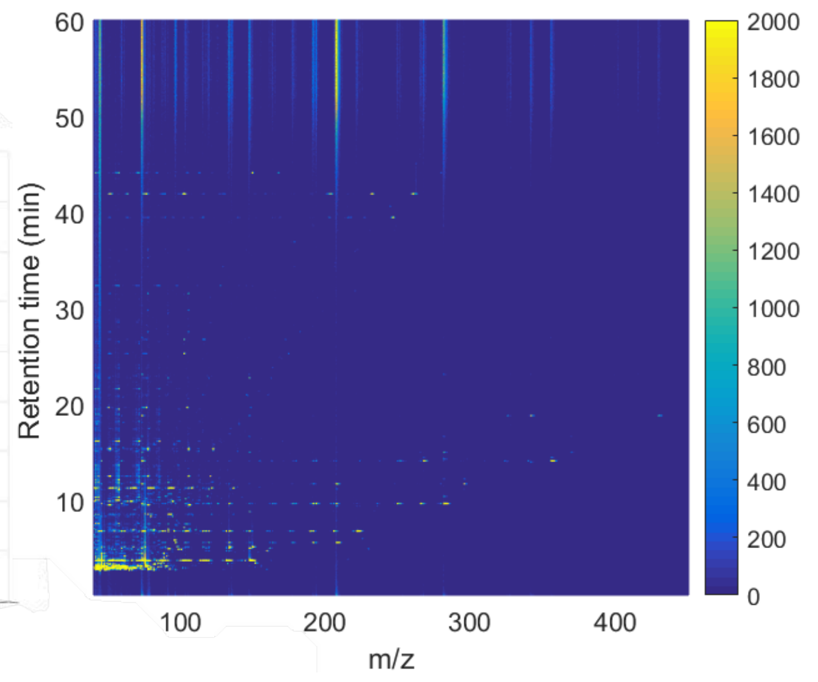
ABUNDANCE MATRIX

One of the leading analytical methods to detect VOCs in breath is gas chromatography-mass spectrometry (GC-MS), which produces a two-dimensional abundance matrix.

One single sample can contain **over 9 million** variables.

The values in the abundance matrix are affected by instrument and environment-related noise.

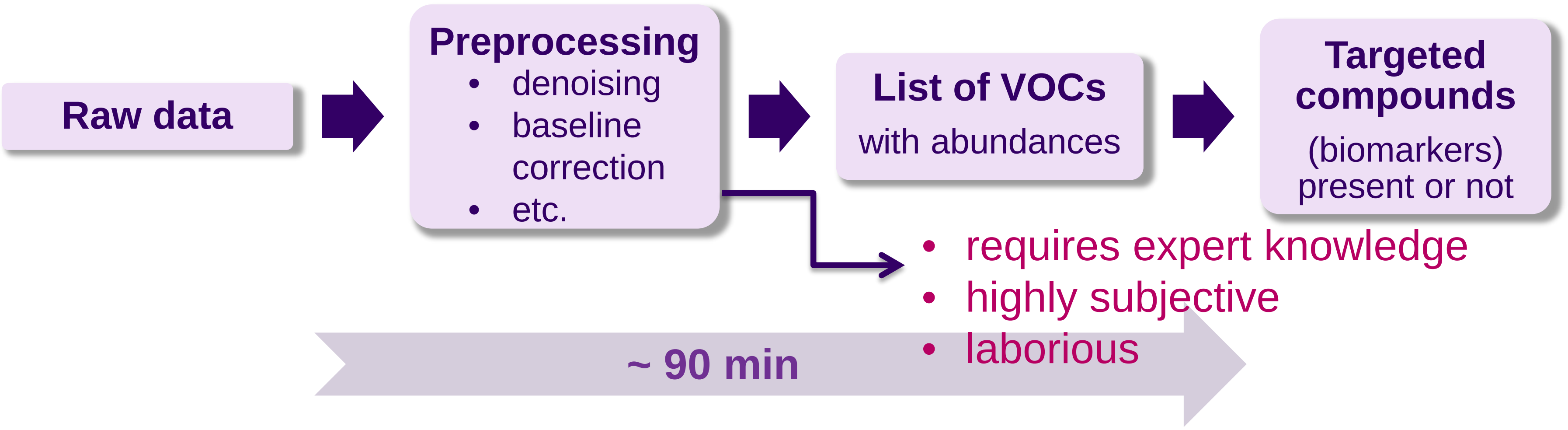




TARGETED ANALYSIS: STANDARD APPROACH

Targeted analysis of GC-MS data seeks a defined panel of VOCs to detect compounds of interest, e.g. known biomarkers.

Due to the data complexity and high level of noise, data pre-processing is a critical step in the standard targeted analysis to obtain reliable results.


Raw data → **Preprocessing**
- denoising
- baseline correction
- etc.

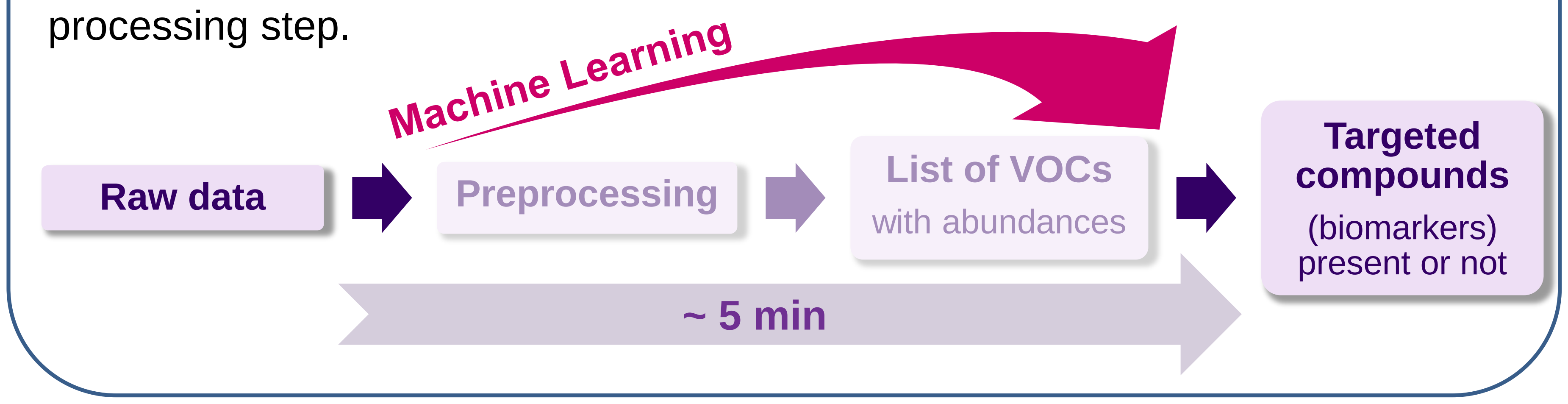
→ **List of VOCs** with abundances → **Targeted compounds** (biomarkers) present or not

- requires expert knowledge
- highly subjective
- laborious

~ 90 min

TARGETED ANALYSIS: PROPOSED APPROACH

The novel idea of our study is to exploit the pattern recognition ability of machine learning to learn to recognise unique patterns of targeted compounds **directly from raw GC-MS data** and therefore bypassing a highly complex pre-processing step.


Machine Learning

Raw data → Preprocessing → List of VOCs with abundances → **Targeted compounds** (biomarkers) present or not

~ 5 min

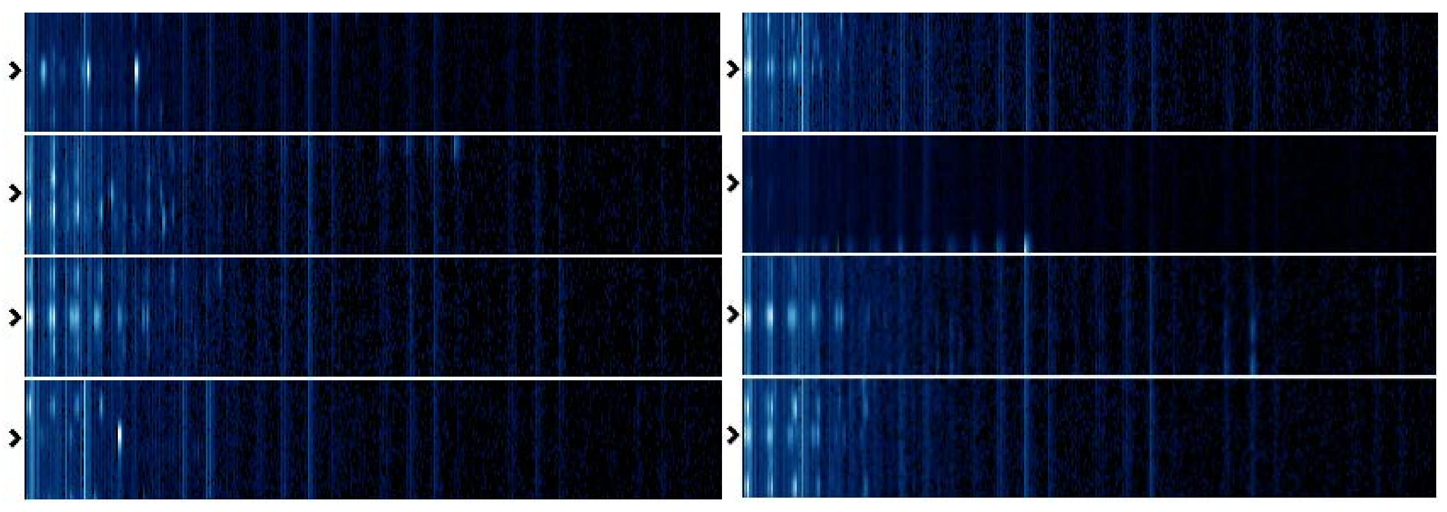
MATERIALS

BREATH SAMPLES DATASET

Materials: 41 breath samples obtained from participants with different types of cancer receiving radiotherapy.

The breath samples dataset was randomly divided into train and test set in the proportion 29/12.

Target VOCs: 8 aldehydes (cancer-related compounds)



METHODS

AUTOMATED TARGETED ANALYSIS

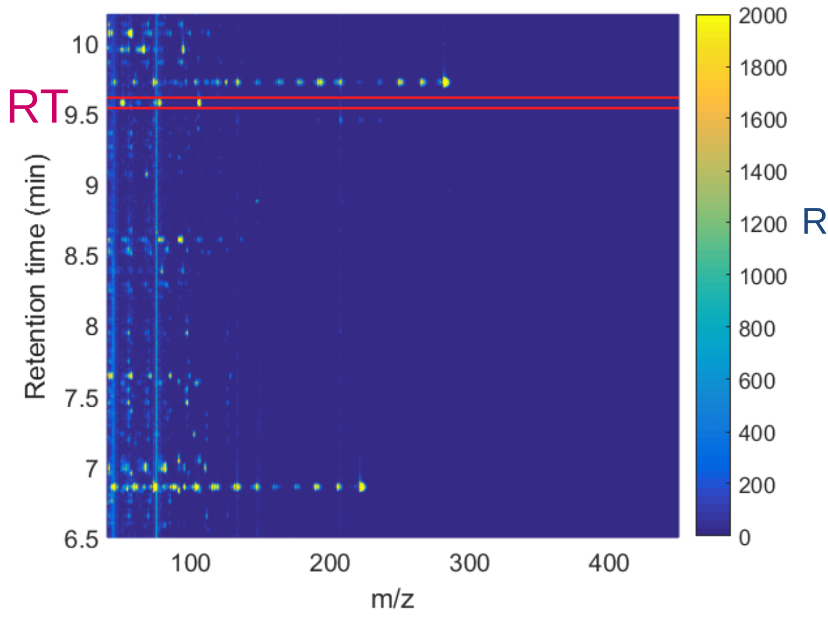
Proposed research idea includes three steps:

1. Preparation of a labelled dataset of targeted VOCs' patterns
2. Supervised learning of ion patterns
3. Detection of targeted VOCs in raw GC-MS data

DATASET OF TARGETED VOCs' PATTERNS

The RT location of target VOCs was derived from the labelled data processed by experts. Based on that, pattern segments of target compounds were extracted from the abundance matrix.

Raw data



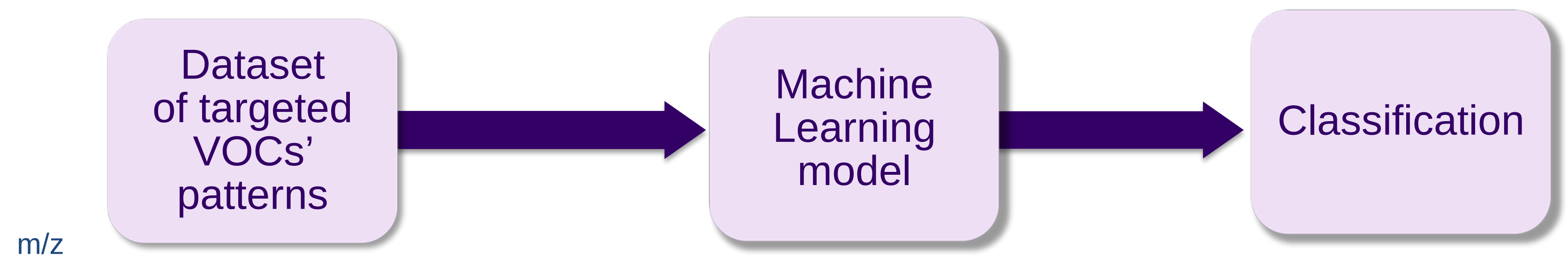
VOC pattern segment

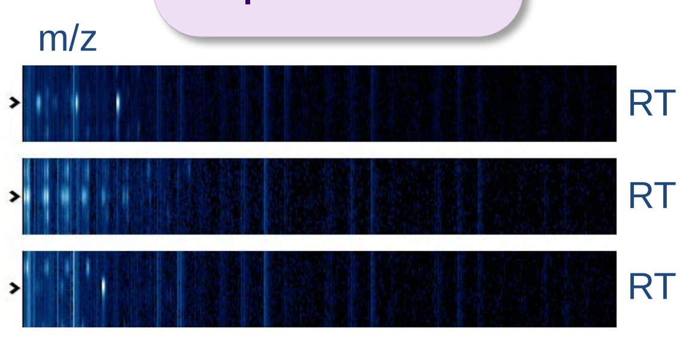
Patern segment size: 70 (12 sec.) x 411

To increase the robustness of the training we applied data augmentation, giving **33 600** pattern segments in train dataset and **14 400** in test dataset.

SUPERVISED LEARNING OF ION PATTERNS

The machine learning models were trained in a supervised manner to classify patterns of target VOCs. Our study investigated Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs) and Shallow Neural Networks (NNs).

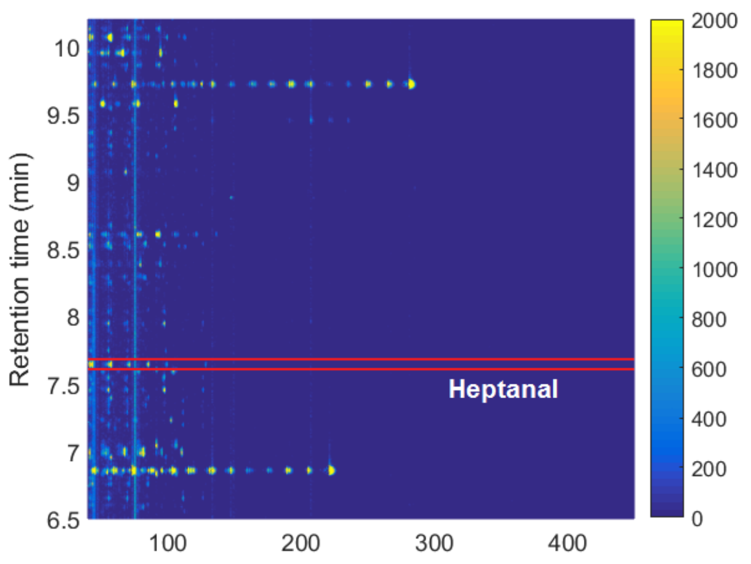

Dataset of targeted VOCs' patterns → Machine Learning model → Classification



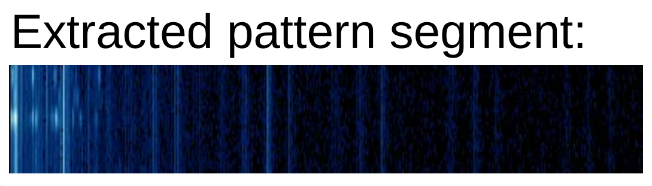
- Convolutional Neural Networks
- Support Vector Machines
- Shallow Neural Networks

- Decanal
- Hexanal
- Octanal
- ...

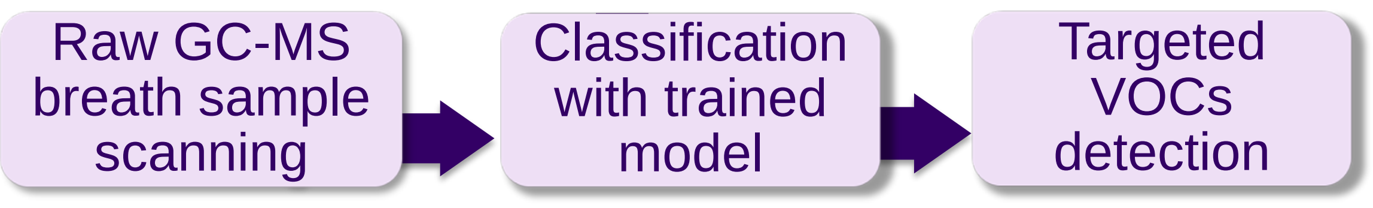
DETECTION OF TARGETED VOCs IN RAW GC-MS DATA



Extracted pattern segment:



Once a learning algorithm can recognise ion patterns, the entire **raw breath sample** can be scanned **quickly** and **automatically** to search for compounds of interest. Consecutive pattern segments are extracted from the abundance matrix and classified by the trained model.


Raw GC-MS breath sample scanning → Classification with trained model → Targeted VOCs detection

Detected compound: **Heptanal**

RESULTS

BREATH SAMPLES SCANNING RESULTS

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)**

The table presents the results of the scans of the 12 breath samples from the test set.

All methods achieved 100% sensitivity. CNN reported the lowest number of false positives.

The scan of one sample involves over 22500 evaluations (the dimension of the RT axis).

UNCERTAIN FALSE POSITIVES

While scanning the raw breath samples, the trained models detected a number of compounds, which were false positives, at positions on the RT axes specific for these compounds. We called such false positives uncertain. Since 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 is actually a true positive.

CONCLUSIONS

Machine learning was applied to detect volatile organic compounds **directly from raw GC-MS data**. Convolutional neural network achieved the best performance.
The proposed methodology can **speed up biomarkers detection** and has the significant potential to contribute to the **development of a breath-based diagnostic platform**.

REFERENCE:

SKARYSZ, A. et al, 2018. Convolutional neural networks for automated targeted analysis of gas chromatography-mass spectrometry data; International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8-13 July 2018.