Cláudia Martins Alves

Bachelor of Science in Biomedical Engineering

Signal processing and automatic classification tools in the development of a new opto-electronic nose

Dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Biomedical Engineering

Adviser: Susana Isabel Conde Jesus Palma, Post-Doctoral Research Fellow, FCT-NOVA

Co-adviser: Ana Cecília Afonso Roque, Associate Professor, FCT-NOVA

Examination Committee

Chairperson: Prof. Dr. Carla Maria Quintão Pereira
Raporteur: Prof. Dr. Pedro Manuel Cardoso Vieira
Member: Dr. Susana Isabel Conde Jesus Palma

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

September, 2018

Signal processing and automatic classification tools in the development of a new opto-electronic nose

This document was created using the (pdf)LaTeX processor, based in the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Acknowledgements

I would first like to thank my thesis advisors, Doctor Susana Palma and Professor Cecíclia Roque. Professor Cecíclia is a tireless leader and Doctor Susana is a great example of what a Biomedical Engineer is capable of doing. Their guidance and motivational words made this journey more easy.

A special thanks to Professor Hugo Gamboa, whom without his input and expertise in Machine Learning it would not be possible to successfully conduct this work. I would also like to thanks his PhD students who also had an important participation.

I would like to thank all people from the Biomolecular Engineering Laboratory for their feedback and cooperation. The miscomprehension that sometimes arised caused by my lack of knowledge of the Chemistry world, was clarified with patience and kindness. I was very lucky to make part of this interdisciplinary team that is willing to grow together.

Last, but not least, I have to thank my parents for their unconditional support. I would not be here without the sacrifices they made to provide me and my sister with a good education.

Abstract

Bacterial infections are a worldwide concern due to the increasing microbial resistance to antibiotics. Therefore, a need to create fast diagnose methods has risen.

Electronic noses are devices that try to mimic the olfactory system. These systems became popular due to their fast response time and portability, and for that reason, they are seen as a possible diagnose method.

In the Biomolecular Engineering laboratory, a project involving an electronic nose is being developed, in which the final goal is the diagnosis of bacterial infections.

The objective of the present dissertation was to develop an analysis tools to complement the system that is being developed.

First, some preprocessing methods were chosen and applied to the acquired data, then a classification tool was developed. Machine learning algorithms were used: a recursive feature selection method was applied to select the best features to characterize the signals and a Support Vector Machine classifier trained to distinguish eleven volatile classes.

Five experiments were analysed and three different sensor formulations tested. Since the device is yet not fully developed, samples which were used were not from bacteria. Instead, simple volatile organic compounds were used.

The results showed that it was possible to efficiently distinguish all compounds with the proposed methods. Moreover, important conclusions related with the current state of the project where drawn. Namely, sensor stability is possible during long, continuous periods of time, but limitations in the reproducibility of the production method may influence the performance of the classifier.

Keywords: Electronic nose, volatile organic compounds, machine learning, recursive feature elimination, Support Vector Machine

Resumo

As infeções bacterianas são um problema mundial e tem afetado cada vez mais pessoas devido ao aumento da resistência dos micro-organismos a antibióticos. Existe, portanto, a necessidade de métodos de diagnóstico rápidos.

Os narizes eletrónicos são dispositivos que tentam imitar o sentido olfativo do ser humano. Estes aparelhos tornaram-se populares devido à sua capacidade de resposta rápida e ao facto de serem portáteis. Devido a isso, tornaram-se interessantes como um possível método de diagnóstico.

No laboratório de Engenharia Biomolecular, está a ser desenvolvido um nariz eletrónico que tem como objetivo final diagnosticar infeções bacterianas.

Esta dissertação teve como objetivo desenvolver ferramentas de análise para complementar o dispositivo que está a ser desenvolvido.

Primeiro, foram usados métodos de pre-processamento para tratar os dados adquiridos. Depois, foram usados algoritmos de aprendizagem automática (do inglês, *machine learning*). Para a escolha das melhores características que descrevem o sinal foi usado um método de seleção recursiva (do inglês, *recursive feature selection*), e para a classificação de onze voláteis foi escolhida uma Máquina de Vetores de Suporte (do inglês, *Support Vector Machine*) como o modelo de aprendizagem.

Foram analisadas cinco experiências e três tipos de sensores diferentes. Uma vez que o sistema ainda não está suficientemente desenvolvido, as amostras utilizadas não foram provenientes de bactérias, em vez disso, usaram-se compostos orgânicos voláteis simples.

Os resultados obtidos mostram que é possível distinguir eficientemente todos os compostos com os métodos propostos. Além disso, foram obtidas conclusões importantes relativamente ao estado atual do projeto. Nomeadamente, os sensores são estáveis durante longos e contínuos períodos de tempo, mas as limitações existentes na reprodutibilidade do método de produção dos géis pode influenciar o bom desempenho do classificador.

Palavras-chave: Nariz eletrónico, compostos orgânicos voláteis, aprendizagem automática, seleção recursiva de características, Máquina de Vetores de Suporte

Contents

List of Figures xiii

List of Tables xvii

Acronyms xix

1 Introduction 1
- 1.1 Objectives . . . 2
- 1.2 Thesis overview . . . 2

2 Theoretical Concepts 3
- 2.1 The device . . . 3
  - 2.1.1 The hybrid gas-sensing gel . . . 3
  - 2.1.2 Generation of the opto-electronic nose's signal . . . 5
  - 2.1.3 The opto-electronic nose . . . 5
- 2.2 Machine Learning . . . 7
  - 2.2.1 Preprocessing . . . 8
  - 2.2.2 Learning . . . 9
  - 2.2.3 Validation . . . 11
  - 2.2.4 Prediction . . . 11

3 State of Art 13

4 Development of signal processing tools 15
- 4.1 Introduction . . . 15
- 4.2 Methods . . . 15
  - 4.2.1 Experimental methods . . . 15
  - 4.2.2 Computational methods . . . 16
- 4.3 Results and discussion . . . 17
  - 4.3.1 Sensors stability . . . 17
  - 4.3.2 Normalization . . . 17
- 4.4 Conclusion . . . 19

5 Development of classification tools 25

5.1 Introduction . . . 25
5.2 Methods . . . 27
5.2.1 Experimental methods . . . 27
5.2.2 Computational methods . . . 27
5.3 Results and discussion . . . 29
5.3.1 Feature extraction . . . 29
5.3.2 Feature selection . . . 32
5.3.3 Tuning the hyperparameters of the classifier . . . 34
5.4 Conclusion . . . 36

6 Case studies **39**
6.1 Introduction . . . 39
6.2 Methods . . . 40
6.2.1 Experimental methods . . . 40
6.2.2 Computational methods . . . 40
6.3 Results . . . 41
6.3.1 Data from sensors made with the standard formulation . . . 41
6.3.2 Data from sensors made with the Ionic Liquid [BMIM][Cl] . . . 44
6.3.3 Data from sensors made with starch instead of gelatine . . . 46
6.4 Discussion of results . . . 46
6.5 Conclusions . . . 49

7 Concluding Remarks and Future Work **51**
7.1 Future Work . . . 52

Bibliography **53**

I Experiment 1 **59**

II Experimental protocol **65**

III Experiment 2 **67**

IV Experiment 3 **73**

V Experiment 4 **75**

VI Experiment 5 **81**

List of Figures

2.1 Section of the sensor observed by POM. 4
2.2 Graphical representation of the changes in morphology of a LI-LC droplet. . 5
2.3 Biomolecular Engineering Laboratory's opto-electronic nose. 6
2.4 Representation of a sensor with a POM image of the sensing area. 6
2.5 Representation of a cycle with its exposure and recovery times. 7
2.6 Steps for the development of a machine learning algorithm. 7

4.1 Response of six of the eighteen sensors exposed for 100 minutes. 18
4.2 Overlapped cycles of the acquired signals. 19
4.3 DTW similarity measures for the sensors exposed to 6 different VOCs 20
4.4 Overlapped cycles of all sensors per VOC. 21
4.5 Mean DTW similarity measure and respective standard deviation for each VOC. 22
4.6 Comparison between the waveform of non-normalized and normalized cycles for the sensors exposed to ethanol and isopropanol. 23

5.1 Schematic representation of Support Vector Machines. 26
5.2 Example of a signal generated with the proposed function for curve fitting. . 28
5.3 Example of a cycle and its first and second derivatives. 30
5.4 Estimated curves for the two parts of a cycle 31
5.5 Results from the RFECV algorithm. 34
5.6 Normalized confusion matrix obtained for the validation of the RFECV algorithm. 35
5.7 Visualization of the classifier's decision regions in two dimensions with use of the two principal components of a PCA. 35
5.8 Molecular structure of the 11 VOCs used. 37
5.9 Molecular structure of 5CB. 37

6.1 Comparison of two batches of standard sensors before they were exposed to VOC's 42
6.2 Comparison of two standard sensors before being exposed to any volatile and after being exposed to a sequence of 11 VOCs, twice. 43
6.3 Saturated response of sensor *D_67* when exposed to acetone in Experiment 4.a. 43
6.4 Normalized confusion matrix for the test with standard sensors. 44

6.5 Saturated response of sensor *C_45* when exposed to acetone in Experiment 4.b. 45
6.6 Comparison of the two batches of sensors that differ the IL from the standard formulation before they were exposed to VOC's 45
6.7 Comparison of two sensors that differ the IL from the standard formulation before being exposed to any volatile and after being exposed to a sequence of 11 VOCs, twice. 46
6.8 Normalized confusion matrix made for the sensors that differ from the standard formulation in the IL. 47
6.9 Normalized confusion matrix made for the sensors that have starch instead of gelatine. 47

I.1 Experiment 1 - Responses of all sensors exposed to acetone. 59
I.2 Experiment 1 - Responses of all sensors exposed to isopropanol 60
I.3 Experiment 1 - Responses of all sensors exposed to hexane 60
I.4 Experiment 1 - Responses of all sensors exposed to ethyl acetate 61
I.5 Experiment 1 - Responses of all sensors exposed to diethyl ether 61
I.6 Experiment 1 - Responses of all sensors exposed to ethanol 62
I.7 Experiment A - Overlaped cycles of all sensors exposed to acetone. 62
I.8 Experiment A - Overlaped cycles of all sensors exposed to isopropanol. . . . 62
I.9 Experiment A - Overlaped cycles of all sensors exposed to hexane. 63
I.10 Experiment A - Overlaped cycles of all sensors exposed to ethyl acetate. . . . 63
I.11 Experiment A - Overlaped cycles of all sensors exposed to diethyl ether. . . . 63
I.12 Experiment A - Overlaped cycles of all sensors exposed to ethanol. 64

III.1 Experiment 2 - Responses of all sensors exposed to heptane. 67
III.2 Experiment 2 - Responses of all sensors exposed to hexane 68
III.3 Experiment 2 - Responses of all sensors exposed to toluene 68
III.4 Experiment 2 - Responses of all sensors exposed to chloroform 68
III.5 Experiment 2 - Responses of all sensors exposed to dichloromethane 69
III.6 Experiment 2 - Responses of all sensors exposed to diethyl ether 69
III.7 Experiment 2 - Responses of all sensors exposed to ethyl acetate 69
III.8 Experiment 2 - Responses of all sensors exposed to acetone 70
III.9 Experiment 2 - Responses of all sensors exposed to acetonitrile 70
III.10 Experiment 2 - Responses of all sensors exposed to ethanol 70
III.11 Experiment 2 - Responses of all sensors exposed to methanol 71

IV.1 Experiment 3 - Responses of all sensors exposed to hexane 73
IV.2 Experiment 3 - Responses of all sensors exposed to toluene 74
IV.3 Experiment 3 - Responses of all sensors exposed to acetone 74
IV.4 Experiment 3 - Responses of all sensors exposed to ethanol 74

V.1 Experiment 4 - Responses of all sensors exposed to heptane. 75

V.2 Experiment 4 - Responses of all sensors exposed to hexane 76
V.3 Experiment 4 - Responses of all sensors exposed to toluene 76
V.4 Experiment 4 - Responses of all sensors exposed to chloroform 76
V.5 Experiment 4 - Responses of all sensors exposed to dichloromethane 77
V.6 Experiment 4 - Responses of all sensors exposed to diethyl ether 77
V.7 Experiment 4 - Responses of all sensors exposed to ethyl acetate 77
V.8 Experiment 4 - Responses of all sensors exposed to acetone 78
V.9 Experiment 4 - Responses of all sensors exposed to acetonitrile 78
V.10 Experiment 4 - Responses of all sensors exposed to ethanol 78
V.11 Experiment 4 - Responses of all sensors exposed to methanol 79

VI.1 Experiment 5 - Responses of all sensors exposed to heptane. 81
VI.2 Experiment 5 - Responses of all sensors exposed to hexane 82
VI.3 Experiment 5 - Responses of all sensors exposed to toluene 82
VI.4 Experiment 5 - Responses of all sensors exposed to chloroform 82
VI.5 Experiment 5 - Responses of all sensors exposed to dichloromethane 83
VI.6 Experiment 5 - Responses of all sensors exposed to diethyl ether 83
VI.7 Experiment 5 - Responses of all sensors exposed to ethyl acetate 83
VI.8 Experiment 5 - Responses of all sensors exposed to acetone 84
VI.9 Experiment 5 - Responses of all sensors exposed to acetonitrile 84
VI.10 Experiment 5 - Responses of all sensors exposed to ethanol 84
VI.11 Experiment 5 - Responses of all sensors exposed to methanol 85

List of Tables

2.1 Examples of supervised learning algorithms 10

4.1 Experimental conditions of Experiment 1. 16

5.1 Numeration and description of the first 16 features proposed to describe the cycles. 31
5.2 Numeration and description of the features extracted from the curve fitting model. 32
5.3 Ranking of the selected features by RFECV with LOO and 10-fold cross validation. 33

6.1 Different sensor formulations used. 39
6.2 Set of the 12 selected features to describe the cycles and to be used as input in the classifier. 41
6.3 Scores obtained by comparing to different batches of sensors. 48
6.4 Scores obtained from the same batch of sensors. 48
6.5 Cross valitadion scores of datasets obtained from different experiments. . . . 49

II.1 Relative concentrations of sample in the e-nose chamber. 65

Acronyms

[BMIM][Cl]	1-butyl-3-methylimidazolium chloride.
[BMIM][DCA]	1-butyl-3-methylimidazolium dicyanamide.
5CB	4-cyano-4'-pentylbiphenyl.
DTW	Dynamic Time Warping.
DWT	Discrete Wavelet Transform.
FN	False Negative.
FP	False Positive.
IL	Ionic Liquid.
LC	Liquid Crystal.
LED	Light-emitting Diode.
LOO	Leave-one-out.
PCA	Principal Component Analysis.
POM	Polarized Optical Microscopy.
RFECV	Recursive Feature Elimination with Cross Validation.
SVM	Suport Vector Machine.
TN	True Negative.
TP	True Positive.
VOC	Volatile Organic Compound.
WHO	World Health Organization.

Chapter 1

Introduction

Bacterial infections are a worldwide problem and have been causing uprising concern due to the increasing microbial resistance to antibiotics. Nowadays, it is estimated that, in some regions, around 50% of infections are caused by resistant bacteria. For this reason, the World Health Organization (WHO) has launched a campaign to reduce the prescription of antibiotics [1].

Therefore, there is a need, not only to reduce the use of antibiotics, but also to develop fast infection diagnostic devices [2].

Electronic noses are devices which mimic the biological sense of smell. In the biological olfactory system, there are hundreds of olfactory receptors. Each odour excites a specific set of them, and the combination of responses is then processed by the brain, which corresponds it to a smell. A similar process occurs in electronic noses: there is an array of gas sensors, with different compositions, each with a distinct selectivity towards distinct odorant molecules. The combined information received by a transduction system is then processed and classified for the specific purpose wanted [3, 4].

The popularity of these devices has risen, as they are seen as fast and portable tools [5]. Nowadays, electronic noses are being used for food and air quality control [6–9] and there is also a great interest in the clinical field for diagnose purposes [10–14].

Hybrid gel materials sensitive to volatile organic compounds (VOC) have been developed and being used as sensors in an in-house built opto-electronic nose [15]. In ISO 16000-6, a VOC is defined as: "any organic compound with a boiling point in the range from (50 °C to 100 °C) to (240 °C to 260 °C), corresponding to having saturation vapour pressures at 25 °C greater than 100 kPa" [16].

In the current stage of the project, functional characterization of the sensors is ongoing. Namely, regarding optimization and reproducibility of the production method and stability and durability of the sensors.

One of the final goals of the Biomolecular Engineering Laboratory's ongoing investigation is to use these types of devices to identify bacterial infections. There has been research showing that bacteria produce patterns of VOCs that contribute to the distinction between different bacterial species [17–19], and that some VOC's produced by bacteria are not normally produced by the human body. Furthermore, some VOC's are produced mainly by one type of bacteria, thus making them possible indicators of bacterial presence [20].

1.1 Objectives

The next step to be closer to the desired application is to create a signal analysis tool to complete the opto-electronic nose. Therefore, the main goal of this dissertation is the development of algorithms for sensor characterization during the current development and validation stages. Firstly, to ascertain if the reproducibility of responses caused by the gas samples can be achieved, and also if the sensors developed so far have the ability to distinguish a wide range of compounds.

Firstly, a set of functions will be developed for signal pre-processing and response characterization. Then, machine learning will be used for the classification of the VOCs, mainly because of its ability to extract important statistical information from big sets of data and find pattern that humans suspect exists but are not able to quantify.

1.2 Thesis overview

The present chapter contextualizes the reader on the need of the proposed thesis, as well as the main objectives.

Chapter 2 summarizes the concepts needed to understand both the current state of the Biomolecular Engineering Laboratory research and the basic theory behind machine learning. After the notation that will be used throughout the document is clarified, Chapter 3 presents the state of art on electronic noses and classification tools that have been used.

Then, in Chapter 4, the signal processing tools used to treat the acquired data are described and applied to study sensor stability. After establishing the tools for preprocessing, the following steps for constructing the machine learning algorithm were made: feature extraction, feature selection and tuning the classifier to achieve best performance. This can be found in Chapter 5. Finally, Chapter 6 contains the application of the preprocessing and classification tools on different case studies.

Chapter 7 concludes the dissertation with some final remarks and suggestions for future works.

There are also 5 annexes (I ,II,III,IV,V and VI), which contain details about the laboratorial experiments that were used to generate data for the analysis.

Chapter 2

Theoretical Concepts

This chapter is divided into two parts. In the first part, an explanation of the specifications of the Biomolecular Engineering Laboratory's device will be made for better understanding of what will be analysed. In the second part, there is an overview on the general aspects of machine learning.

2.1 The device

Together with sensor development, a signal acquisition device is also being assembled to record the responses of the hybrid gas sensing gels [21]. This in-house developed system is referenced along this work as electronic nose (e-nose). A general description of sensor composition, generation of the signal and finally, of the device itself will be made.

2.1.1 The hybrid gas-sensing gel

The biggest difference between other Electronic Noses and the device being developed in the Biomolecular Engineering Laboratory is the sensor's composition. A new hybrid gel material with specific optical properties that respond to the interaction of VOCs was developed. A hybrid gas-sensing gel is composed by molecules of liquid crystal (LC) and Ionic Liquid (IL) that self-assemble into droplets in a matrix of biopolymer, forming a hybrid gas-sensing gel [15]. The standard formulation is the biopolymer gelatine, the IL 1-butyl-3-methylimidazolium dicyanamide ([BMIM][DCA]) and the LC 4-cyano-4´-pentylbiphenyl (5CB).

Liquid crystals have mechanical properties between a solid and a liquid. Crystal components form regular structures, while liquids are disorganized. LC linger between these two states in a way that forms defined phases in which they can exist but with a degree of anisotropy [22]. The orientational phases are: nematics, smectics and cholesteric. The

phase is related to the manner in which the molecules are ordered. In nematic phase, molecules orientation remains the same even if they move around; in smetics phase, not only the orientation but also the movement is limited, and particles are arranged in layers; lastly, in cholesteric phase there are also layers but the molecular alignment rotates among them [23, 24].

The LC phase can be identified by using optical methods – Polarized Optical Microscopy (POM). This is possible due to the birefringent properties of LC molecules. When light passes through a LC sample that is placed between two crossed polarizers (90 degrees phase between them), the order in which the molecules are arranged is evidenced, since LC appear bright under crossed polarizers [25].

When it comes to LC droplets, when seen through POM with crossed polarizers, there are three known configurations: bipolar, radial and preradial [25]. Dark spots are called defects and the radial configuration is characterized for having only one point defect [25].

Ionic Liquid are salts with melting points lower than 100°C. They have useful properties such as conductivity and are known as a group of green solvents [26]. In the hybrid gas-sensing gels, the IL promotes gelatine dissolution and the alignment of the LC molecules inside LC-IL droplets [15].

In the hybrid gas-sensing gels, the IL molecules self-assemble with the LC and therefore the majority of the droplets present a radial configuration [15]. This LC configuration rotates the plane of the polarized light and is what creates the typical morphology that is observed in the POM image presented in Figure 2.1.

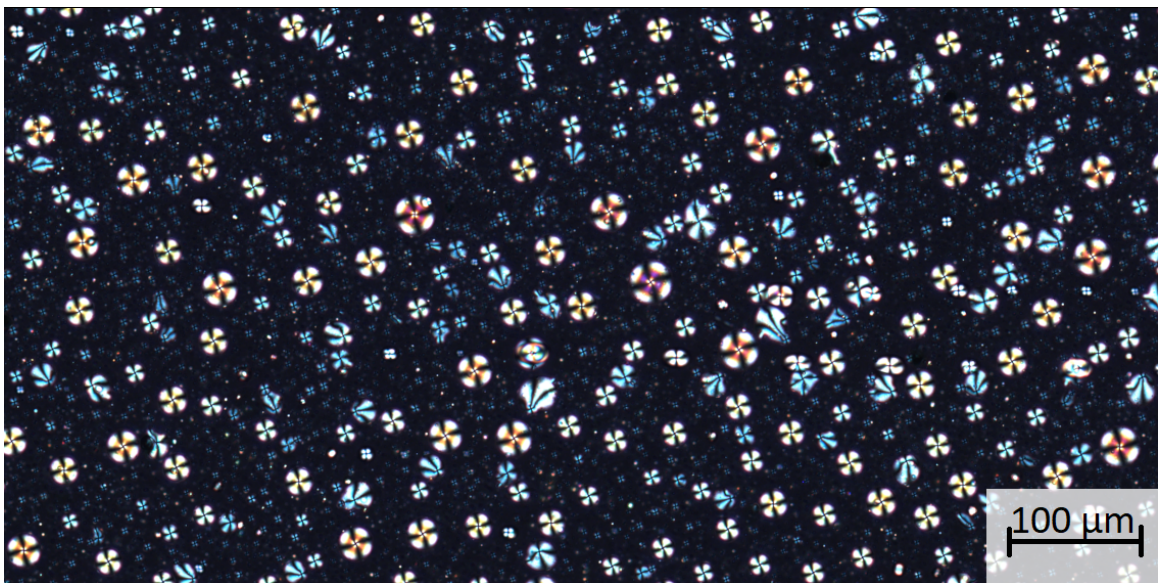


Figure 2.1: POM image of a 1 x 0.5 mm section of the gel observed with a 100x magnification.

2.1.2 Generation of the opto-electronic nose's signal

The configuration of the LC-IL droplets changes when exposed to a volatile organic compound (VOC).

When the droplets are in a radial configuration, they rotate the plane of polarized light and it is able to pass between the crossed polarizers (Figure 2.2a). When the sample enters the chamber, organisational order is lost and rotation of the plane of polarized light is no longer possible, therefore, the intensity of the light drops after the second polarizer (Figure 2.2b). When air is flushed over the sensor to clean the VOC away, the LC recovers the radial configuration and lets light pass again. The variation of intensity of light that passes through the mask during this optical phenomenon is then recorded by the e-nose.

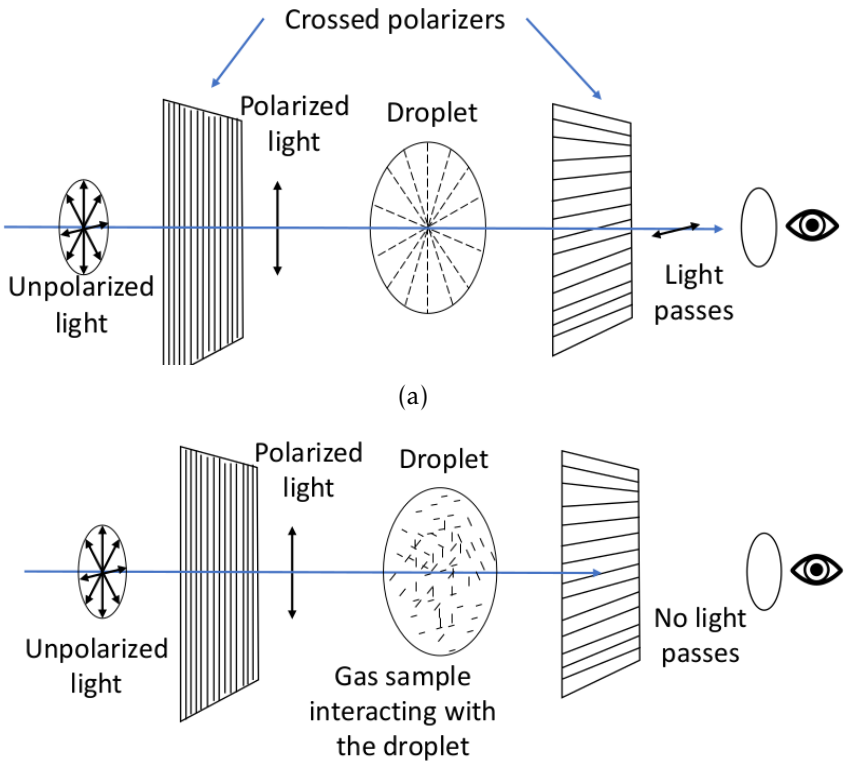


Figure 2.2: Graphical representation of the changes in morphology of a LI-LC droplet when (a) it is not exposed to any sample and (b) when interacts with the VOC sample (b). Adapted from [15].

2.1.3 The opto-electronic nose

The device has a chamber that is isolated from ambient light and that accommodates six sensors. Each one has a thin film of sensing gel on one side of a glass slide and a black mask on the other (Figure 2.4). The circle cut in the mask is aligned with a light-emitting

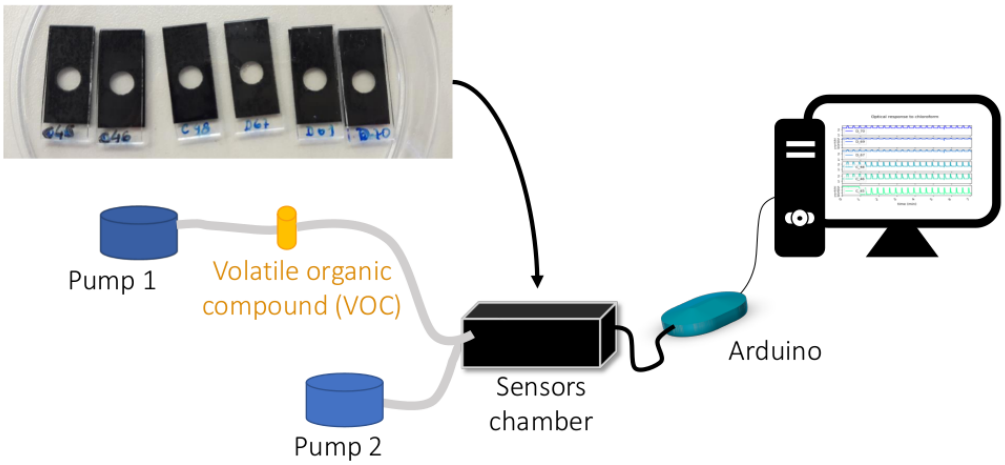


Figure 2.3: Biomolecular Engineering Laboratory's opto-electronic nose.

diode (LED) and a photodiode, thus ensuring that the gel sensing area seen is always the same.

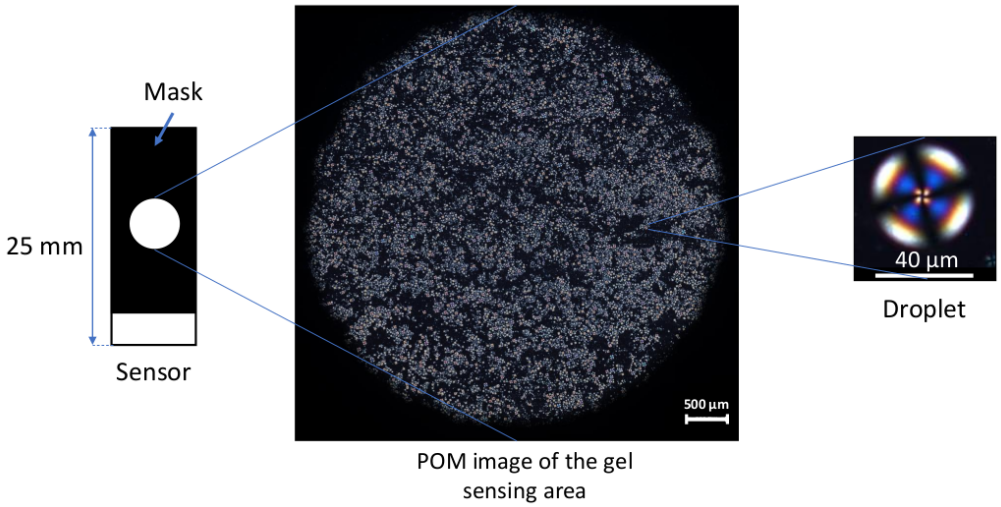


Figure 2.4: Representation of a sensor with the mask and a zoomed picture of the sensing area, obtained by POM.

To obtain a signal, two pumps work alternately. The sample in liquid form is placed in a vial, and since it is volatile it evaporates and is pushed into the chamber by the exposure pump, where it will interact with the sensors. VOCs begin to interact with the sensing gels and a change in the intensity of light detected by the photodiode is evidenced. Then the recovery pump is activated an it injects air in order to clean the chamber. This enables the sensors to recover their morphology and the light intensity captured by the photodiode returns to its baseline.

The instructions to control the device are given by a computer. An Arduino receives them, and it activates the pumps, powers the LEDs and acquires the signal from the photodiode. A scheme of the e-nose is found in Figure 2.3.

In terms of software, the latest version of the open access language Python (Python 3.6) is used. Python has a variety of libraries that suit the aim of the work, such as *Pymata*, that allows the communication between the computer and the Arduino, *pandas*, used for intuitive and easy data manipulation, and *sklearn* for machine learning.

In this work, an exposure to gas sample followed by a recovery period, in which air restores the radial configuration, will be referred to as a cycle (Figure 2.5). Due to the way the transduction circuit is assembled, when the sample interacts with the material and the luminosity decreases, the signal increases. After the acquisition of a few cycles, the signal is saved and later processed.

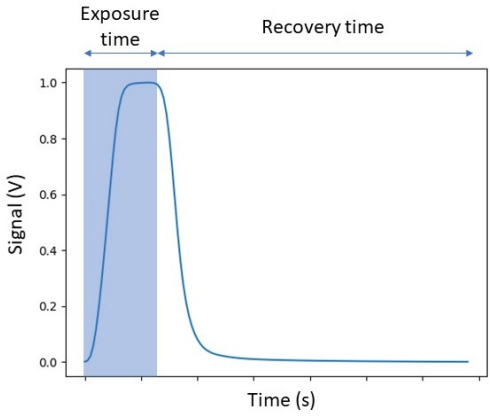


Figure 2.5: Representation of a cycle with its exposure and recovery times.

2.2 Machine Learning

Author Ethem Alpaydin defines machine learning as "programming computers to optimize a performance criterion using example data or past experience" [27].

In the early days, machines took decisions based on a handmade rule system of "if" and "else". Nowadays, with the advances in computer science and the junction with statistics, computers are able to learn from a big set of data and use that knowledge to provide a prediction. The flow chart illustrated in Figure 2.6, which is recommended by Sebastian Raschka in his book [28], will be used to explain the necessary steps for the development of a machine learning algorithm.

Figure 2.6: Steps for the development of a machine learning algorithm.

2.2.1 Preprocessing

Raw signals need to be preprocessed before features can be extracted. The first step is filtering, to make sure the signal's features will be described and not the noise. The most common contamination comes from the electrical circuit, which can be present through all the signal or appear as peaks. In the first case a band pass filter can be used to attenuate the unwanted frequency and in the other case a median filter can be used. After filtering the signal, it is possible to perform normalization to attenuate the effects of the inherent variability of the data.

2.2.1.1 Feature Extraction

First, for a machine to be able to learn, the data it will be given must be in a format that it can understand. For example, if we want to differentiate flowers from each other it is not enough to give the learning algorithm pictures of said objects. Since a machine is not able to decide how to characterize and object, one must define what information should be passed to the learning phase, such as the petal length, colour and others [29].

The extracted information is usually presented in the form of a table, where each row corresponds to an example and each column to a feature. Taking the flower example, each flower is an example and petal length and the other characteristics are the features [15].

There are three main methods to extract features [30]: morphological feature extraction (characteristics such as the difference between the maximum value and the baseline, time to reach the maximum, area under the curve, among others); methods such as Fast Fourier Transform and Discrete Wavelet Transform, in which coefficients are used as features; and curve fitting where the parameters of the fitted model serve as features.

Among the examples, there are usually outliers which are, sometimes, the result of errors in measurements; these can be removed so that the algorithm does not learn from samples that are not representative of the object in study. This normally requires an experts' opinion. Normalization can also be performed to enable comparison between features [29].

Lastly, one must have into account the possibility that the dataset could be unbalanced. In other words, if the dataset presents more samples of flower A and almost none of flower B, the algorithm will learn how to recognize flower A faster that flower B. In these cases, if possible, the best way to oppose this effect is to collect more data. If even then the problem remains, there are techniques which help balance the outcome [31].

2.2.1.2 Feature Selection

Feature Selection brings many advantages. With fewer features, less data will be processed so the algorithm will run faster. It can also help avoid overfitting [32]. Overfitting happens when the learning algorithm works well with the data given but when exposed to new data the performance gets worse. This can happen for example if the data set

is not representative enough thus making the algorithm not being able to generalize its outcomes [29].

There are several techniques of feature selection, which are classified in three main groups: filter methods, embedded methods and wrapper methods [33].

Filter methods are the simplest, but they only evaluate intrinsic properties of the data and can eliminate features that alone do not bring more information, but when combined with others into the algorithm can provide better results. This happens because the filter methods do not have into consideration the interaction with the classifier. On the other hand, wrapper and embedded methods do [33].

Embedded methods are the ones, that as the name implies, are integrated in the learning algorithm [33].

Wrapper methods are iterative algorithms than in each iteration take into account the performance of the learning model. They perform better than filter methods and work better with fast modelling algorithms such as Naïve Bayes classifiers and Support Vector Machines [33].

2.2.1.3 Dividing the data set

Once it is ensured that the information is not redundant, the data set is divided into three sets: training set, test set and validation set. The training and the validation sets are used to train the algorithm, while the test set remains untouched for final performance evaluation. Before doing so, is recommended to shuffle the examples of the dataset in order to remove trends that could influence the learning process [28, 31].

2.2.2 Learning

There are three types of machine learning: supervised learning, reinforcement learning and unsupervised learning [28, 29]. Each one serves a different purpose depending on the type of data and goal one has. Bellow follows an explanation of each kind.

- **Supervised Learning**: This type of machine learning is used when one has a finite set of known possible outcomes characterized by determined features. The algorithm learns which characteristics correspond to which outcome and when presented with a new input, it tries to fit the output that matches it more accurately. This means that if, for example, the algorithm learned the differences between four flowers, when presented with a fifth kind of flower it would try to fit it to the most likely from the known four types of flowers. When the output is a continuous value, one subtype of supervised learning can be used: regression. Here the features are the parameters of a model, such as a slope [28]. Table 2.1 lists some examples of supervised learning algorithms.

- **Reinforcement learning**: This type of learning tries to mimic the way living beings learn in a way that the algorithm's learning process is enhanced with feedback

through interaction with the environment. The reinforcement can be in the form of reward or punishment [34]. The typical example is a chess game. Each action affects the probability of winning. A move is made considering the state of the board and winning is seen as the reward [28].

- **Unsupervised learning**: Unlike the previous two types, here there is not an expected result. Unsupervised learning is mainly used to analyse the structure of the data [28]. Taking the flowers example, the algorithm does not know how many different kinds of flowers there are or even if there are flowers. There could be roses and daisies, or a more complex set of objects.

Table 2.1: Examples of supervised learning algorithms

Algorithm	Description
k-Nearest Neighbours	The prediction is made by finding the "nearest neighbour(s)" of the sample input. The nearest can be as simple as the one having the smaller Euclidian distance. It is considered the simplest algorithm of supervised learning [29].
Naïve Bayes Classifiers	Algorithm that applies Bayes' theorem. Admits features to be independent from each other. There are three kinds: Gaussian, Multinomial and Bernoulli [29].
Decision Trees	Is a series of if/else questions of the type "is a greater than b?". Each question is created through statistical analysis. The dimensions of the three can be controlled by limiting the number of nodes (questions) it can have [29].
Random forests	A collection of decision trees, usually used to avoid overfitting [29].
Support Vector Machine (SVM)	SMV maps the inputs into a n-dimensional space and divides it with a hyperplane. The divisions set the margin between classes [29].

For an initial approach, some authors suggest beginning with the simplest algorithm, and if that does not fit the problem then explore more complex options. This considers computational cost.

Overall, it is good practise to test different algorithms and compare their performance. One can analyse the confusion matrix, in which each column represents a predicted class and each row the actual class. The confusion matrix allows to calculate metrics to measure performance, for instance: accuracy, precision, sensitivity, specificity, F1 score and Matthews correlation coefficient (useful for unbalanced data sets). Another useful performance score is the area under the curve of two common curves: Precision-Recall curve and Receiver Operating Characteristic [31, 35].

The last thing to have into account in a machine learning methodology is that the algorithm's default hyperparameters may not be the most suitable and may lead to underfitting, where the model is to simple to explain the variance in the data, or overfitting, where the model is too optimized and is not able to generalize for the proposed problem. Therefore, hyperparameter optimization may be required. A hyperparameter is, for example, the number of neighbours to consider in the k-nearest neighbours [28, 31].

2.2.3 Validation

After selecting the best features and choosing the best algorithm for the training set, the test set will be used. A test in unseen data is needed to confirm if no overfitting is contaminating the performance results. Since the work developed will use mostly supervised learning, the focus of validation will be in three most common cross-validation methods [28, 29].

- **k-fold cross-validation**: The data set is split randomly into k smaller sets (folds) of equal size. One of them will serve as the test set and the others as training sets. This will happen for each fold. In the end there will be k accuracy values, calculating the mean one obtains the average accuracy [29].
- **Stratified k-fold cross-validation**: Same procedure as the previous one but here the folds have the same proportions between classes as the complete dataset [29].
- **Leave-one-out cross-validation**: Here each fold is a single sample. The algorithm is trained will all the samples except one and, then, it is exposed to a single new input to see how it reacts. For small data sets it can work very well, but if a large number of samples is present it can take too long to compute [29].

2.2.4 Prediction

When all the previous steps are concluded, the model can finally be used for prediction. If all steps where completed successfully, then the outcomes are reliable and thus the resultant algorithm can be used for the application it was designed for. It's important to highlight that an algorithm developed for a specific problem cannot be used broadly. This means that if another problem arises, all the steps must be redone to find a new optimum solution.

Chapter 3

State of Art

There are many applications for e-noses [4, 36] including food quality control [6, 7], environmental monitoring [8, 9] and disease diagnosis [10–14], among others. This state of art will focus on the latter, especially those aimed at bacterial infection diagnosis, and will be more focused on the analysis techniques employed.

Many researchers are using commercial e-noses for bacterial infection diagnosis such as: AirSense Portable Electronic Nose system from Airsense Analytical, Schwerin, Germany [37]; The electronic noses from the eNose company [38]; Cyranose Electronic Nose from Sensigent, Intelligent Sensing Solutions [39, 40]; Bloodhound BH-114 [41]; Osmetech Microbial Analyser [42]; and ChemPro [43].

Other e-noses are being developed, and the trend is always to make them reliable and portable [44–46].

As mentioned above, researches have been able to diagnose a disease by means of this type of device. For example, Nakhleh *et al* [10] studied the possibility of using an artificially intelligent nanoarray to differentiate between 17 diseases. Moreover, the study took patients from five different countries which verifies the broad applications of the technology.

Saidi *et al* [12] used an in-house build e-nose to test if it was possible to diagnose chronic kidney disease and diabetes mellitus. Morphological characteristics were extracted from the signal and the algorithms used for classification were Hierarchical cluster analysis (a method of cluster analysis – samples are divided into groups following a criteria of similarity, usually a distance), Principal Component Analysis (PCA) and Support Vector Machine.

Not all diseases are caused by bacteria, but other researches have proven that e-noses are useful for their detection/ identification [18, 37, 39, 41, 43, 47, 48]. Most of the research has been made to diagnose bacterial infections in the upper respiratory tract [38],

urinary tract [41, 43], ear nose and throat [49], and wounds [50].

Qinghua He *et al* [51] did an experiment with infected wounds. The signal was extracted with an in-house built e-nose, and feature extraction was made through DWT. Performance of radial basis function network and SVM algorithms were compared, the last being the one with better performance.

Geffen *et al* [38] managed to distinguish patients with chronic obstructive pulmonary disease without infection, and patients suffering from viral infection or bacterial infection. They used the Aeonose to acquire samples, an artificial neural network for the learning algorithm and leave-10%-out method for validation.

Trincavelli *et al* [18] used NST 3220 Emission Analyzer to differentiate ten types of bacteria. They used the static response (value of the signal at the end of the acquisition minus the baseline value) and the average of the derivate as features, Linear Discriminant Analysis for feature selection (Discriminant Analysis was used as a supervised method used to maximize feature separability), SVM with grid search for hyperparameter optimization and k-fold cross for validation.

Liang *et al* [48] used SVM, Principal Component Analysis and Independent Component Analysis to differentiate between simple samples of *Escherichia coli*, *Staphylococcus aureus* and *Pseudomonas aeruginosa* as well as samples with mixtures of these bacteria. For feature selection Kennard Stone sequential method was used.

Moens *et al* [37] used AirSense Portable Electronic Nose to differentiate ten micro-organisms. Morphological characteristics of the signal were extracted; a wrapper approach was used on feature selection to obtain better results. The performance of an artificial neural network and the k-nearest neighbour algorithm were compared with leave-one-out validation.

Aksebzeci *et al* [47] used Cyranose 320 to differentiate 7 strains. Morphological characteristics were obtained from the signal, and pre-processing of the baseline was performed with three different methods. For feature selections two methods were used: the first was to choose the three sensors that showed higher differentiation; the second was using PCA. Three Discriminant Analysis methods were compared and the best was used for classification. The validation method used was k-fold cross validation.

Roine *et al* [43] used ChemPro 100i to differentiate micro-organism through Linear Discriminant Analysis and Logistic Regression. To validate the results, leave-one-out and k-fold cross methods were used.

Many applications can result from the use of an e-nose, but one thing in common is the need for methods to analyse the collected data.

Chapter 4

Development of signal processing tools

This chapter has two main goals. First, to ascertain sensor stability in order to infer if a sensor can be used for long periods of time. Second, to establish the preprocessing tools to be used when analysing data from the e-nose.

4.1 Introduction

In previous work, it was proven that the sensors were able to distinguish VOCs [15]. Besides the discriminatory capacity of the sensors it was necessary to ascertain in the responses remained the same after some periods of time.

First, it was necessary to select the best filtering method and to obtain the individual cycles of the signals acquired.

Then, the change in cycle waveform over time was quantified for each sensor using a measure of similarity to ascertain sensor durability.

Finally, signals from sensors exposed to the same sample were compared and a normalization method was chosen to attenuate the effects of variations in amplitude.

4.2 Methods

4.2.1 Experimental methods

The data from Experiment 1, involving six volatile samples (acetone, ethanol, isopropanol, hexane, diethyl ether and ethyl acetate), was used to study sensor stability. The experimental conditions are annotated in Table 4.1. Each sample interacted with three sensors (all made with the same formulation) for 100 minutes in cycles of 5 seconds of exposure and 25 seconds of recovery. The signal responses can be found in annex I.

Table 4.1: Experimental conditions of Experiment 1.

Volatile	quantity (mL)	temperature (ºC)	sampling rate (Hz)
Acetone	20	24	5
Diethyl ether	15	24	10
Ethanol	20	23	5
Ethyl acetate	20	24	5
Hexane	18	24	10
Isopropanol	20	24	5

4.2.2 Computational methods

4.2.2.1 Filtering

To remove electronic noise, the signal was first filtered with the median filter from the *SciPy* library, kernel size equal to 11, to remove inductive noise caused mainly by the activation of the pumps. Then, the signal was filtered with *smooth* function from the *novainstrumentation* library, default parameters were used.

4.2.2.2 Cutting the signals

The e-nose records information that indicates when the recovery or the exposure pumps are working. This information was used to split the signals and obtain the individual cycles. Every time there is an indication that the exposure pump is on, a cycle begins and a cut in the signal is made.

4.2.2.3 Similarity measure

Euclidean distance and Dynamic Time Warping (DTW) are two common methods to measure similarity between time series [52]. Euclidean distance is the simplest to implement thus making it the most frequently used measure [53]. DTW is used mainly because of its ability to compare sequences that are not aligned in time or that have different lengths. It works by aligning the time series by choosing the path that minimizes a given cost function [54].

DTW was used because the acquisition rate of the system was not always constant, thus making the cycles have different lengths. Also, it has shown the best results in previous experiments [52, 55]. The function *dtw* from *tslearn* library that calculates DTW similarity measure is given by Equation 4.1 and corresponds to the Euclidean distance between the aligned time series, where X and Y are the time series and P is the alignment path.

$$DTW_{measure}(X,Y) = \sqrt{\sum_{(i,j)\in P} (X_i - Y_j)^2} \tag{4.1}$$

4.2.2.4 Normalization

The two proposed methods of normalizing are: method A - scaling the values into range [0,1] (Equation 4.2); and method B - subtracting the cycles' mean and dividing by the cycle' standard deviation (Equation 4.3).

$$signal_{normalized} = \sum_{n} \frac{value_n - \bar{x}}{S} \tag{4.2}$$

$$signal_{normalized} = \sum_{n} \frac{value_n - min(signal)}{max(signal) - min(signal)} \tag{4.3}$$

4.3 Results and discussion

4.3.1 Sensors stability

Figure 4.1 shows the response for 6 of the 18 sensors of Experiment 1, one for each volatile. Individual cycles are not distinguishable due to the time window chosen. The entirety of the experimental time was represented to show that the responses do not always have the same amplitude or baseline. In some cases these two are maintained (for example sensors *c9* and *contf*), and in other cases (like sensor *c12*), the signal varies. However, variations in these two characteristics may not imply that the shape of the cycles change.

Taking into consideration that the baseline varies, all baselines were set to zero for visualization purposes. This was made subtracting the minimum value of each cycle to every value of that cycle. The results are plotted in Figure 4.2, where the cycles of the signals in Figure 4.1 were overlapped. All cycles for the same sensor have, approximately, the same waveform.

To quantify if the format of the response remained the same along time, the similarity DTW measure was calculated for all cycles taking the first one as reference. The higher the DTW similarity measure, the less similar the cycles are to the initial one.

As expected, Figure 4.3 shows that although the score tends to grow through time, it does not thrive far from zero. The sensors that present a bigger change in the similarity measure are those exposed to diethyl ether (Figure 4.3b), but sensor *contf* only starts to show significant changes in the last part of the experiment. The only sensor that changes abruptly in the first minutes is *cont3* (Figure 4.3d), stabilizing its response towards the end.

4.3.2 Normalization

As Figure 4.4 shows, there is some amplitude variability among different sensors exposed to the same sample. Nonetheless, as seen in the previous section, its shape remains similar.

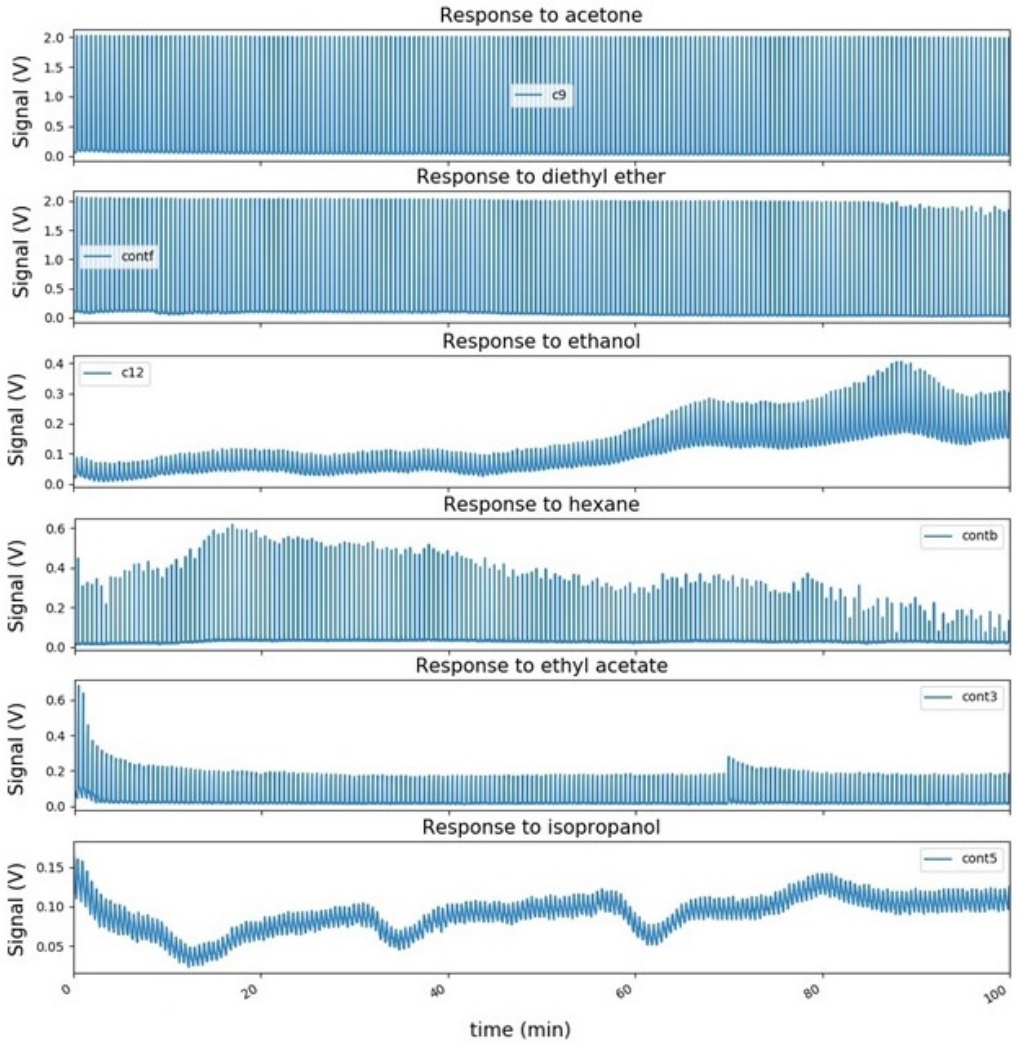


Figure 4.1: Response of six of the eighteen sensors exposed for 100 minutes. Sensors are labeled *cX*, *X* is the unique identifier.

In order to ascertain which of the proposed normalization methods better enhanced the similarity between cycles from different sensors, the DTW score between the average normalized cycle and each normalized cycle was calculated for each volatile. The mean and standard deviation (std) of those results can be found in Figure 4.5.

Overall, method A had better performance, as it is able to reduce the score for most of the samples. Although ethanol and isopropanol signals normalized by this method show higher scores than without normalization, the results are still relatively close to zero. Comparing normalized with non-normalized scores, it is also worth mentioning that the standard deviation is smaller for all VOCs except for isopropanol and ethanol, meaning that there is less variation in the cycle format.

The fact that isopropanol and ethanol did not follow the trend for method A may be

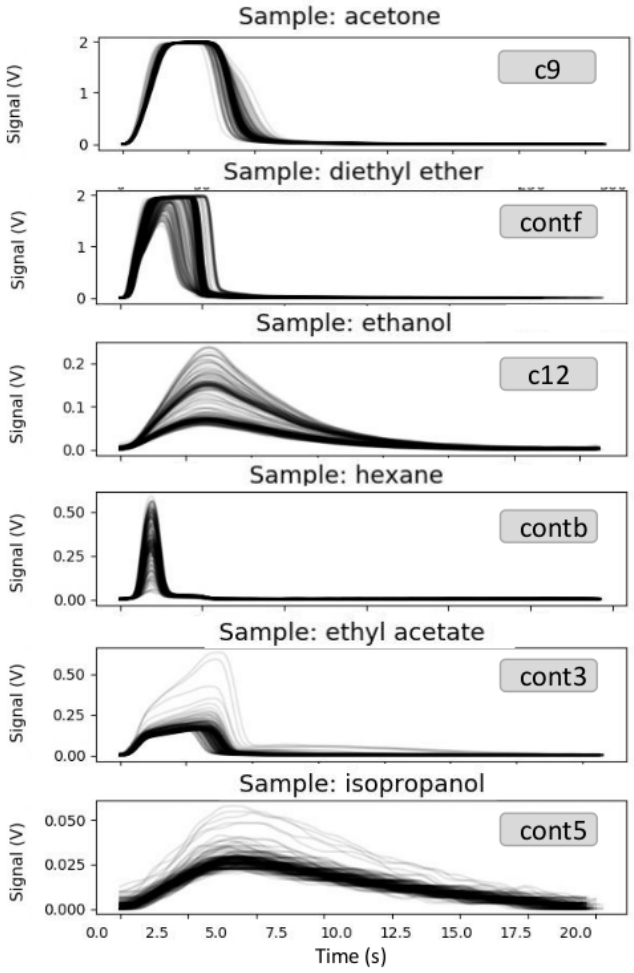


Figure 4.2: Overlapped cycles of the signals in Figure 4.1, each volatile creates a different waveform.

because of the presence of excessive noise. Figure 4.6 shows the overlapped cycles, as well as the mean cycle and respective variation. For both samples, the darker part of the overlapped normalized cycles matches the mean signal; noise is more noticeable in the normalized cycles due to the amplification of the cycles with smaller amplitude, on which noise and signal amplitude are almost the same. For this reason, the results for these were taken as outliers and the method A was approved and used throughout the remaining analyses.

4.4 Conclusion

In conclusion, this initial analysis showed that sensors are able to maintain their response even after long exposure times. This is important because it allowed the characterization of the stability of the sensor's response and, therefore, infer that a sensor can be exposed for long periods of time. Moreover, this information is necessary since the features extracted from the signal are related to its waveform: if the signal does not

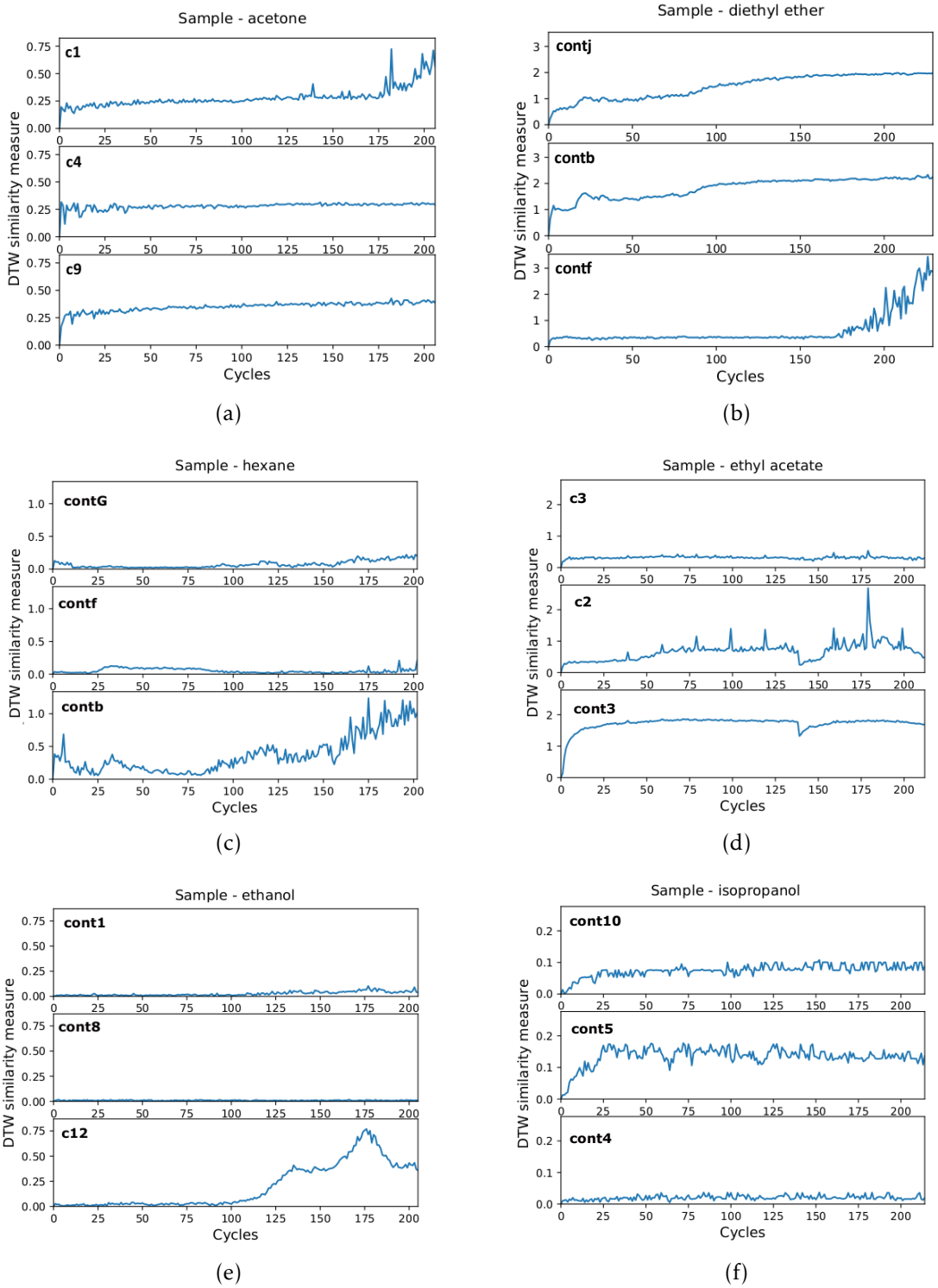


Figure 4.3: DTW similarity measure for: (a) acetone; (b) diethyl ether; (c) hexane; (d) ethyl acetate; (e) ethanol; (f) isopropanol.

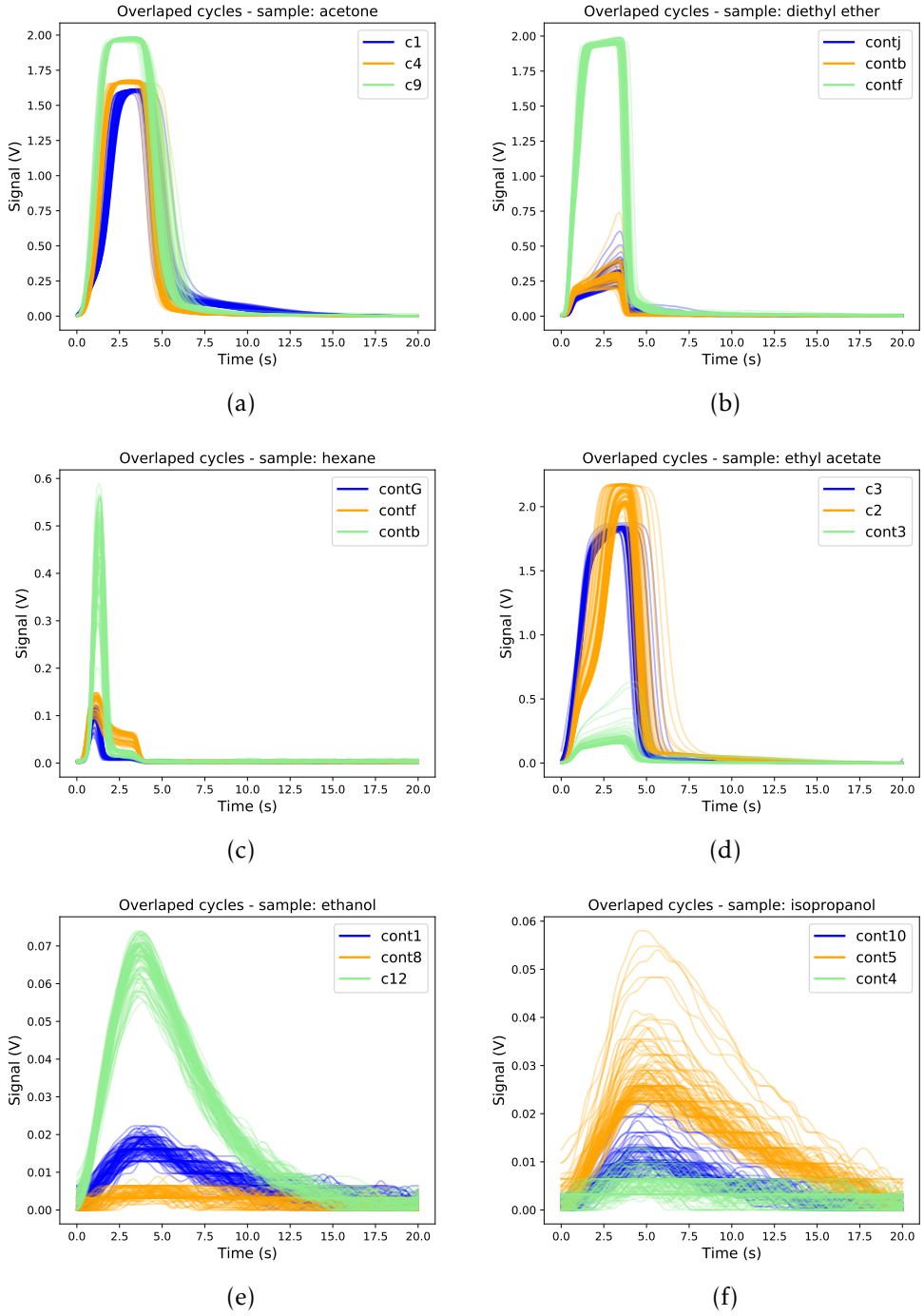


Figure 4.4: Overlapped cycles of all sensors per VOC: (a) acetone; (b) diethyl ether; (c) hexane; (d) ethyl acetate; (e) ethanol; (f) isopropanol.

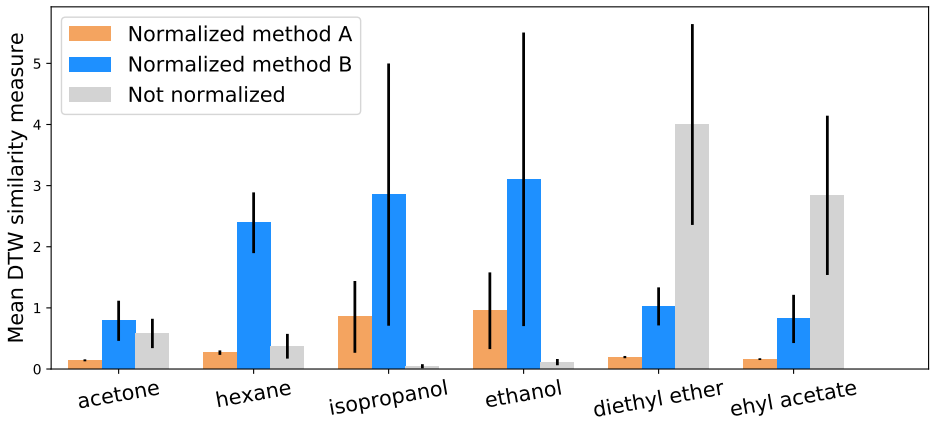


Figure 4.5: Mean DTW similarity measure (bars) and respective standard deviation (black lines) for each VOC.

maintain its format, classification performance becomes limited.

Experiment 1 also evidenced that variations in amplitude and baseline may occur. Therefore, it also helped to establish the necessary pre-processing steps for the rest of the work: filtering, separating the cycles and normalizing them between zero and one (method A).

It is important to note that this dataset was not used in the remaining analyses due to several reasons. Firstly, acquisitions herein presented were made with an older version of the e-nose. The relevant differences between the older version and the one used in the rest of the work, is related to the transduction circuit, in which photodiodes take the role of photoresistors in the acquisition of the optical signal, and also to the sampling rate which was increased. In addition, the final goal of the investigation is to acquire biological samples. The human body is normally at 37 degrees Celsius, and in this experiment the samples were only heated to room temperature (24 degrees Celsius). Finally, it was decided that more VOCs should be used.

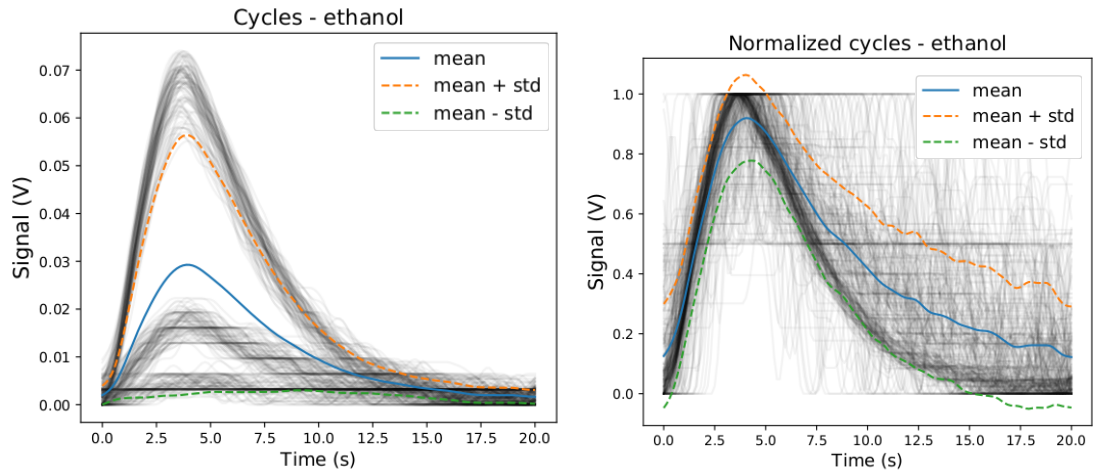


Figure 4.6: Comparison between non-normalized ((a) and (c)) and normalized cycles ((b) and (d)) in terms of waveform for the sensors exposed to ethanol and isopropanol.

Chapter 5

Development of classification tools

After the pre-processing tools were established, a set of features was proposed to describe the cycles. Then, the determination of the most informative features was performed together with the selection and tuning of a classifier. Finally, a validation test was made to ascertain if the selected classification tool was able to give a good performance when exposed to new data.

5.1 Introduction

In the present chapter, a time domain analysis is proposed, combining two types of features: morphological features and parameters of curve fitting models.

The selected approach for feature selection was a wrapper method, specifically Recursive Feature elimination with cross validation (RFECV). This algorithm starts with the complete set of features; a learning model determines feature importance and the worst performing feature is eliminated. This happens in each iteration until there are no more features left. Features are ranked by their performance. The result is a subset with the optimal number of features [29, 56].

The next step was selecting the learning model. A. Jović *et al* stated that linear SVM were feasible to use with wrapper methods [33]. Therefore, it was the used method.

The standard SVM, illustrated in Figure 5.1, is a binary classifier. It finds a hyperplane that bests separates the two classes and creates a decision boundary. The examples of the classes that are closer to the hyperplane are called the support vectors. The hyperplane can be described by the set of points that satisfy $\mathbf{w}^T\mathbf{x} = 0$, $\mathbf{w}$ being the vector perpendicular to the hyperplane. The points that satisfy $\mathbf{w}_o + \mathbf{w}^T\mathbf{x}_{pos} = 1$ form the *positive* hyperplane, and the ones that satisfy $\mathbf{w}_o + \mathbf{w}^T\mathbf{x}_{neg} = -1$ form the *negative* hyperplane. By subtracting

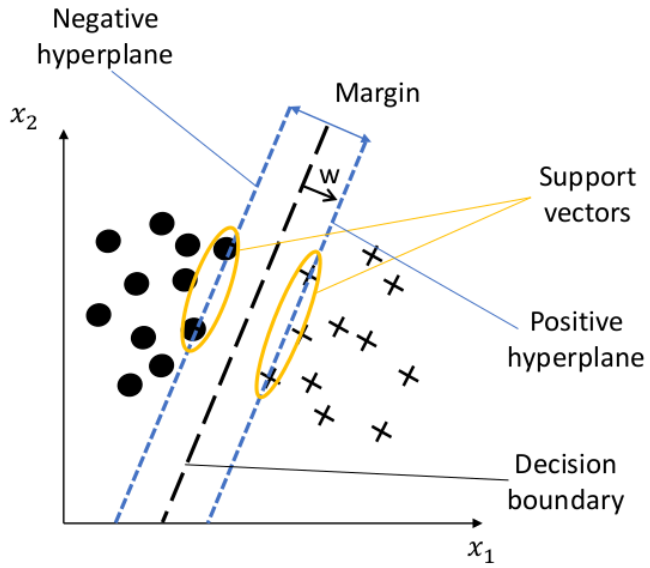


Figure 5.1: Schematic representation of Support Vector Machines. Adapted from [28].

those expressions and normalizing by the length of the vector $\mathbf{w}$, equation 5.1 is obtained.

$$\frac{\mathbf{w}^T(\mathbf{x}_{pos} - \mathbf{x}_{neg})}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \tag{5.1}$$

The left term of equation 5.1 is the distance between hyperplanes, the objective is to maximize it.

There are different methods for generalizing the SVC into a multi-class classifier. The one used in this work was one-vs-rest [57].

To classify an example by one-vs-rest, a classifier per class is created and, in each one, one of the classes is compared with the remaining ones. The algorithm then decides in which hyperplane it is on, and a confidence score is computed. The class with the highest score is the selected one [58].

The Support Vector Classifier performance depends on the kernel, the kernel parameter γ (if kernel is non-linear) and the parameter C.

The kernel is used to transform the original features and project them onto a higher dimensional space via a mapping function [28].

The parameter γ influences the reach of the training examples, a higher γ makes the boundaries closer to the examples. By default, its value is $1/nr_{samples}$. The variable C influences the penalty for misclassification, a larger C means a higher cost [28].

At the end, the SVC with the best combinations of parameters was trained with the selected features as input and validated with unseen data.

5.2 Methods

5.2.1 Experimental methods

To ensure that all data was acquired under the same conditions, an experimental protocol was established by the Biomolecular Engineering Laboratory team (See annex II). This provides comparability to information extracted from different experiments.

The data used in this chapter was obtained with an experiment that followed the proposed protocol - Experiment 2.a (See annex III). Only the responses from the three sensors were made with the standard formulation were analysed in this chapter. These sensors were exposed to a sequence of 11 VOCs, with periods of 5 seconds of exposure and 15 of recovery.

Although sensors are able to respond for long times in continuum, when using e-noses, it is important to have the ability to do quick screenings. Fifteen minutes was defined as the time to use for this experiment. This time segment is shorter than the one used in the previous chapter, and still creates enough samples in order to be used in training algorithms.

5.2.2 Computational methods

5.2.2.1 Preprocessing

Before extracting the features, the preprocessing steps defined in the previous chapter were applied (signal filtering, obtaining the individual cycles and normalization). Since the sampling rate increased, it was necessary to change window length for the *smooth* filter to 50.

5.2.2.2 Feature extraction

Starting with the simpler type, there are some features that can be extracted directly from the cycle signal, as well as its first and second derivatives. The maximum and minimum values of these curves as well as the position in time were they occur, can be directly extracted with *Numpy* functions.

Another characteristic that describes the signal is its area. Although there are many ways of calculating this, the selected method was a *Numpy* function which integrates the signal using the composite trapezoidal rule.

A suggestion made by R. Dutta and colleagues [49] was followed. Skewness and kurtosis were extracted with functions available in *SciPy* subpackage, *stats*. These are measures of the asymmetry of the probability distribution of a random variable around its mean. Although the signal is not a probability density function, extracting these features helped increase the accuracy of their electronic nose classifier. Hence, they were included to see if could also be fitted for the current problem.

For the curve fitting, a *Python* function that uses nonlinear least-squares to estimate the unknown parameters was used. Least squares works by minimizing the sum of the squared deviations between the real data and the model [59].

For the fitting model, the function considered was proposed by Holmberg, M. *et all* (Equation 5.2) where θ_1 corresponds to the amplitude, θ_2 to the rising slope, θ_3 to the position where the signal reaches half the rise, θ_4 to the descent slope and θ_5 to position where the signal reaches half of the descend [60]. This model is based on the logistic function (Equation 5.3).

$$f(x) = \theta_1 . \frac{1}{1 + e^{\theta_2 . (\theta_3 - x)}} . \left(1 - \frac{1}{1 + e^{\theta_4 . (\theta_5 - x)}}\right) \tag{5.2}$$

$$f(x) = \frac{a}{1 + e^{b.(c-x)}} \tag{5.3}$$

Figure 5.2 presents a signal made with Equation 5.2, as well as its first and second derivatives.

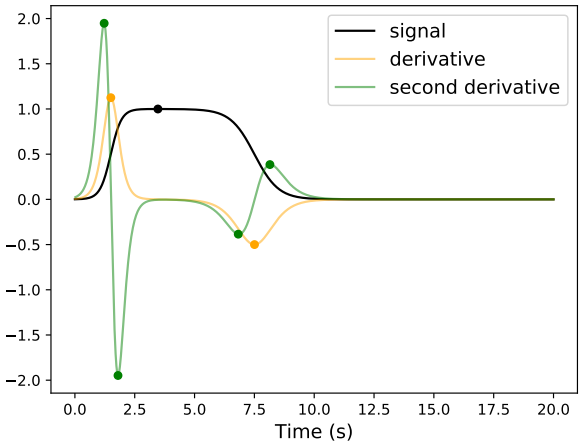


Figure 5.2: Example of a signal generated with the proposed function for curve fitting and its first and second derivatives. The dots mark the peaks. Chosen parameters - θ_1: 1.0 ; θ_2: 4.5; θ_3: 1.5; θ_4: 2.0; and θ_5: 7.5

After the feature extraction, it was necessary to scale the features. To do so, *Standard-Scaler* from *scikit-learn* library was used.

Then, the dataset obtained from the extracted features was divided in two halves (training and validation sets). Rows were randomly chosen, but the proportion of examples for each volatile was the same in both halves.

5.2.2.3 Feature Selection

The function *RFECV* (recursive feature elimination with cross validation), from *Yellowbrick* library [56] was used for feature selection. This function takes as input a classifier, a cross validation method and a scoring metric.

Support Vector Classification (SVC) from *scikit-learn* library was the implemented classifier.

To select the validation method, it was taken into consideration that wrapper methods tend to be time consuming. Therefore, a greedy cross validation method such as leave-one-out could not be the most appropriate if time is a constriction. For that reason, stratified k-folds can be a good alternative since is less time consuming. Both methods were used and compared.

Accuracy, given by Equation 5.4, was chosen as the scoring method, since it provides a measure of global performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

It gives the number of correct classified examples, True Positives (TP) and True Negatives (TN), divided by the number of examples, TP, TN, False Positives (FP) and False Negatives (FN).

5.2.2.4 Tuning the SVC hyperparameters

For the RFECV algorithm to work, the classifier must have a feature importance coefficient to rank their relevance, for that reason the linear kernel was chosen to use with this algorithm since it was the only one that has that functionality.

In terms of the final classifier, there are other kernels available to test, namely polynomial kernel, radial basis function kernel, and sigmoid kernel [57].

Parameters C and γ should be tuned together. For that, is possible to use *GridSearchCV* from *scikit-learn* library that performs a grid search with cross validation to choose the best parameters. This function was used with: values 0.1, 1, 10 and 100 for C; values 0.01, 0.1, 1 and 10 for gamma; and the kernels available, already mentioned.

5.3 Results and discussion

5.3.1 Feature extraction

The typical signal of a cycle rises until a maximum is reached, while the sensor is being exposed, and it decreases while the volatile leaves the chamber. This leads to a typical shape in the derivatives (Figure 5.3). The first derivative has two prominent peaks, with their position corresponding to the inflexion points in the rise (positive peak) and the decrease (negative peak) of the signal. The second derivative has four prominent peaks; two positives when the signal is concave up and two negatives where it is concave down.

To efficiently extract the four peaks of the second derivative, the signal was divided into two parts (from the beginning of the cycle until its maximum and from the maximum until its end), then the maximum and minimum of the respective second derivatives was calculated. The relative position of these peaks in relation to the beginning of the cycle,

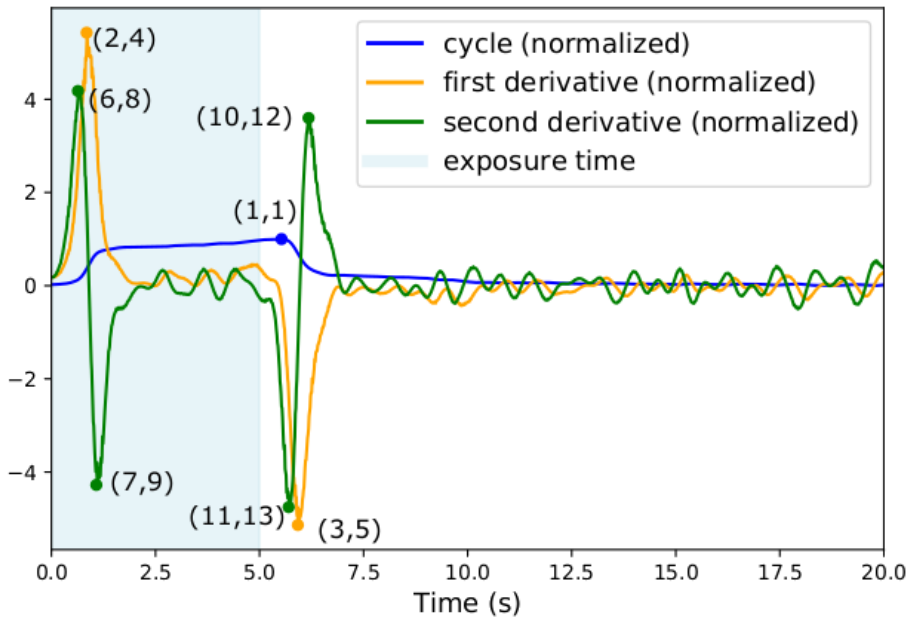


Figure 5.3: Example of a cycle and its first and second derivatives. Dots mark the peaks, darkened part of the plot correspond to the exposure time. The coordinates of the peaks are represented by the numbers of the features. For example, coordinate (2,4) corresponds to the time to reach the maximum of the first derivative of the signal (2) and the value of that maximum (4)

as well at its amplitudes, correspond to the first thirteen extracted features (the cycles' maximum amplitude is always 1 due to the normalization, therefore it is not accounted for). Table 5.1 enumerates these features, as well as the area under the signal, kurtosis and skewness. The numbers 1-13 are represented in Figure 5.3 as the coordinates of the marked points.

For the curve fitting, the approach of the article (Equation 5.2) was promising, and the resemblance of Figure 5.2 with Figure 5.3 suggested that this function could in fact be appropriate to model the signal. However, the algorithm was not able to converge into a solution for the parameters. In order to fit a model to the signal, it was necessary to simplify the approach. The solution was to divide the signal in two parts (from cycle beginning until its maximum and from the maximum until its end) and apply the logistic function (Equation 5.3) to each part of the signal.

There were three parameters to find: a, that corresponds to the amplitude; b that corresponds to the slope, positive if the signal rises and negative if it descends; and c that corresponds to where the signal bisects with the value of half of the rise/descend.

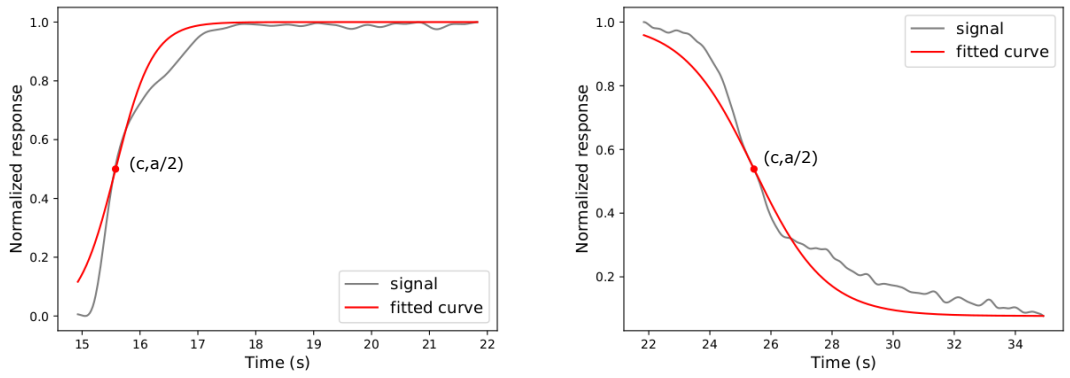
This approach had the additional benefit of being able to extract a and c directly from the signal. Thus, only the parameter b needed to be estimated.

The fact that the baseline does not always return to zero was taken into consideration and the amplitude (a) corresponds to the respective difference between the maximum and the minimum values of the cycle's part.

Table 5.1: Numeration and description of the first 16 features proposed to describe the cycles.

Number	Description
1	Time to reach the maximum value of the cycle
2	Time to reach the maximum of the first derivative of the signal
3	Time to reach the minimum of the first derivative of the signal
4	Maximum value of the first derivative of the signal
5	Minimum value of the first derivative of the signal
6	Time to reach the maximum of the first part of the second derivative of the signal
7	Time to reach the minimum of the first part of the second derivative of the signal
8	Maximum value of the first part of the second derivative of the signal
9	Minimum value of the first part of the second derivative of the signal
10	Time to reach the maximum of the second part of the second derivative of the signal
11	Time to reach the minimum of the second part of the second derivative of the signal
12	Maximum value of the second part of the second derivative of the signal
13	Minimum value of the second part of the second derivative of the signal
14	Area under the signal
15	Skewness
16	Kurtosis

With these simplifications, the curve fitting function was able to find b and an approximation of the signal if obtained. Figure 5.4 shows the fitted curves of the two parts of a cycle. Since the estimated curves are not exactly coincident with the signal, the standard error of the parameter is also taken into account by adding it as a feature.



(a) (b)

Figure 5.4: Estimated curves for (a), the first part of the cycle and for (b), the second part. The red point marks the coordinates where the signal bisects with the value of half the rise/descend.

In the end, eight more features were added to the set: *a*, *b* and *c* for each part of the signal, plus the error associated to each fitting (Table 5.2).

Table 5.2: Numeration and description of the features extracted from the curve fitting model.

Number	Description
17	Parameter a from the first part of the signal
18	Parameter b from the first part of the signal
19	Parameter c from the first part of the signal
20	Error associated with the fit from the first part of the signal
21	Parameter a from the second part of the signal
22	Parameter b from the second part of the signal
23	Parameter c from the second part of the signal
24	Error associated with the fit from the second part of the signal

A set of 24 features was proposed to describe the signals produced by the hybrid sensing gels exposed to VOCs in the e-nose. This information was organized in a table where 24 columns contain the features extracted, and each row corresponds to an example (cycle). The last column corresponds to the name of the volatile to which the sensor was exposed. This format of organization makes easier the usage of *Python* library *pandas*, designed to facilitate data manipulation.

5.3.2 Feature selection

First, the RFECV algorithm ran with the training set. In a first run, leave-one-out (LOO) cross validation was used as the cross-validation method. After that, the algorithm ran with 10-fold validation, and it ran approximately 70 times faster than LOO method. As expected, 10-fold validation was faster than leave-one-out. Moreover, the features chosen as relevant were the same for both cross validation methods. The rank of the features made by each run is shown in Table 5.3. Features ranking is given as: the higher the rank, the lower the relevance. The algorithm ranks as 1 the best features not attributing any discrimination among them. Therefore, features ranked as 1 were the ones selected.

Figure 5.5 shows the graphical representation of the results for the RFECV with 10-fold cross validation. Each dot corresponds to a feature and the mean score given by the learning algorithm. The number of optimal features determined was 12 and the global accuracy was 98.3%.

The selected morphological features were: time to reach minimum value of the first derivative (feature 3), the values of the four prominent peaks of the second derivative (features 8, 9, 12 and 13), the position of the first positive peak of the second derivative (feature 6) and area (feature 14).

The measures of asymmetry of probability distributions, kurtosis (feature 15) and skewness (feature 16) were also suited for the presented problem.

Table 5.3: Ranking of the selected features by RFECV with LOO and 10-fold cross validation.

Number	Description	LOO	10-fold
1	Time to reach the maximum value of the cycle	7	7
2	Time to reach the maximum of the first derivative of the signal	2	4
3	**Time to reach the minimum of the first derivative of the signal**	**1**	**1**
4	Maximum value of the first derivative of the signal	4	3
5	Minimum value of the first derivative of the signal	3	2
6	**Time to reach the maximum of the first part of the second derivative of the signal**	**1**	**1**
7	Time to reach the minimum of the first part of the second derivative of the signal	11	10
8	**Maximum value of the first part of the second derivative of the signal**	**1**	**1**
9	**Minimum value of the first part of the second derivative of the signal**	**1**	**1**
10	Time to reach the maximum of the second part of the second derivative of the signal	10	11
11	Time to reach the minimum of the second part of the second derivative of the signal	8	9
12	**Maximum value of the second part of the second derivative of the signal**	**1**	**1**
13	**Minimum value of the second part of the second derivative of the signal**	**1**	**1**
14	**Area under the signal**	**1**	**1**
15	**Skewness**	**1**	**1**
16	**Kurtosis**	**1**	**1**
17	Parameter a from the first half of the signal	12	13
18	**Parameter b from the first part of the signal**	**1**	**1**
19	**Parameter c from the first part of the signal**	**1**	**1**
20	Error associated with the fit from the first part of the signal	5	5
21	Parameter a from the second part of the signal	13	12
22	**Parameter b from the second part of the signal**	**1**	**1**
23	Parameter c from the second part of the signal	6	6
24	Error associated with the fit from the second part of the signal	9	8

From the fitted curves, the selected parameters were: parameters b and c, from the first part of the signal (features 18 and 19), which correspond to the slope of the rise of the signal and the position in time were it intersects half the amplitude; and the parameter b from the second part of the signal (feature 22), which corresponds to the slope of the descend of the signal.

Moreover, the worst ranked features were the amplitudes of the fitted curves, this

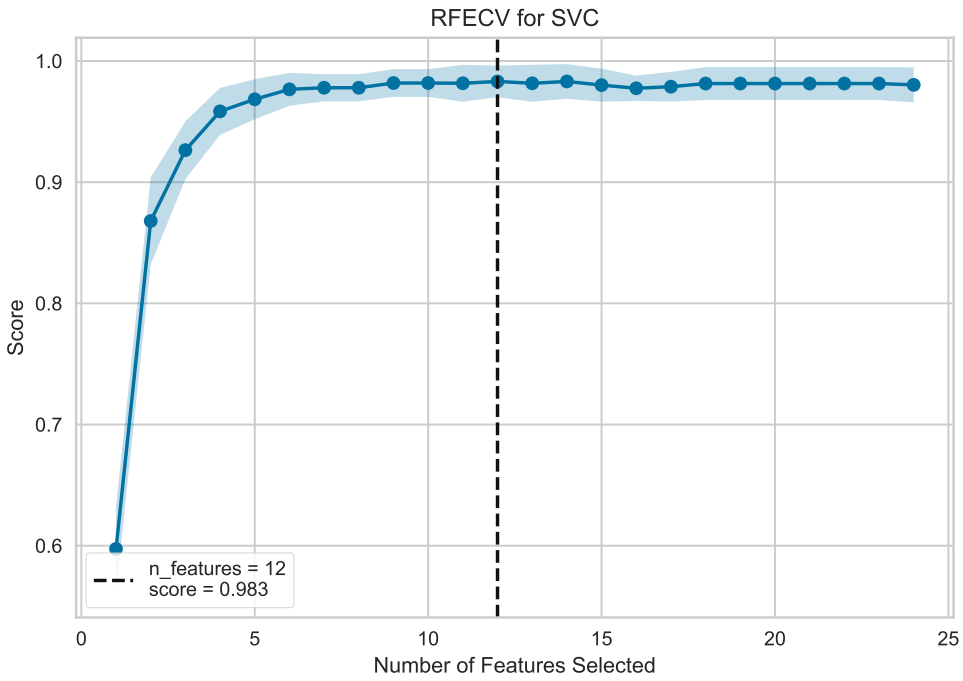


Figure 5.5: Results from the RFECV algorithm run with half dataset of Experiment 2. Dots represent the features along with their mean cross-validated test score, the standard deviation associated is represented in lighter blue (Average accuracy was 98.3%). The vertical line marks the optimal number of features.

implies, that although baseline does not always starts at or returns to zero at the end of a cycle, those variations are not significant.

5.3.3 Tuning the hyperparameters of the classifier

The grid search was performed with the training set, the best kernel was selected and parameters C and gamma were tuned accordingly.

The best performance was obtained with the radial basis function kernel and parameters C: 100 and γ: 0.1. The result of the cross validation was an accuracy of 99.4% ($\pm$ 1.9%).

To validate these results with unseen data, the validation set was used. The classifier managed to obtain an accuracy of 98.8%. Figure 5.6 shows the correspondent normalized confusion matrix.

To visualize the separation of the classes and the decision regions made by the classifier, it was necessary to resort to dimensionality reduction of the data set. For that, Principal Component Analisys (PCA) was applied. Two principal components, as well as the decision boundaries made by the trained SVC are plotted in Figure 5.7. It is noticeable that some clusters are too close together and that the volatiles within those aggregations have similar molecular structures (See Figure 5.8).

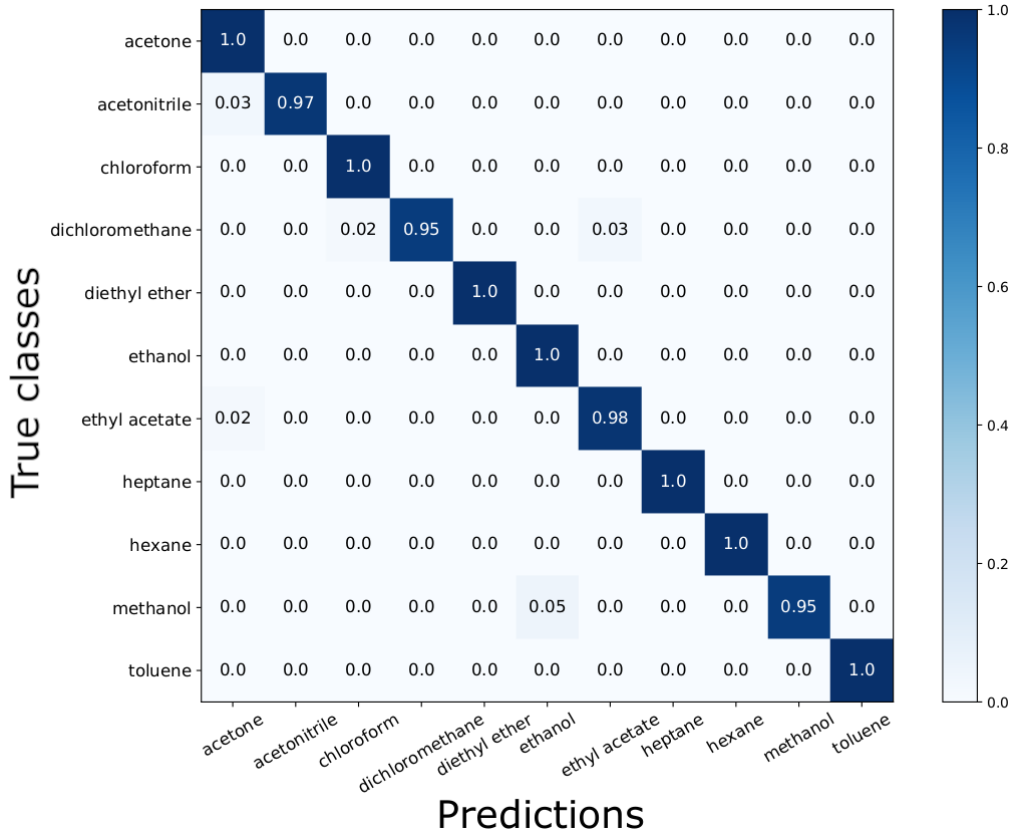


Figure 5.6: Normalized confusion matrix obtained for the validation of the RFECV algorithm. Accuracy of 98.8%

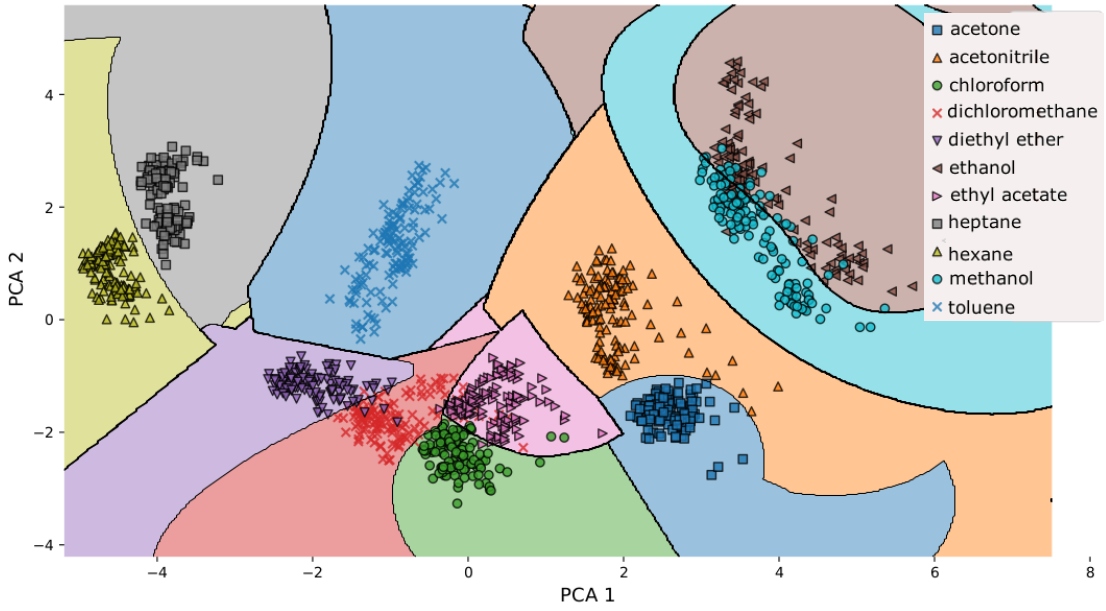


Figure 5.7: Visualization of the classifier's decision regions in two dimensions with use of the two principal components of a PCA.

Heptane and hexane are apolar compounds and their interaction with the gels was expected to be similar since they only differ in one $-CH_3$. It is believed that both these molecules could be interacting directly with 5CB due to structural resemblances (Figure 5.9). Toluene may also be interacting with the LC since it also resembles a part of the structure of 5CB, but a different one than hexane and heptane [15].

Dichloromethane and chloroform have chlorine atoms which are very electrophilic, making them more polar than the last two mentioned. Both ethyl acetate and diethyl ether have a methoxy group, so it would be expected that they would be closer in the clusters, but ethyl acetate also has a double bond. Therefore, the oxygen in the double bond attracts more electrons and makes the molecule more polar than diethyl ether. Acetone also has a double bound, but does not have the methoxy group, this may explain why its interaction with the material is different. Acetonitrile has a nitrogen atom that is less electrophilic than oxygen, but it interacts with the rest of the molecule by a triple bond making electrons more attracted to it. As expected, the effects of the different structures in one molecule are combined and the interaction mechanism resultant show differences in their responses [15].

Methanol and Ethanol have the functional group hydroxyl. Therefore, the proposed explanation for what happens when interacting with the sensors is that they perturb the gelatine by creating hydrogen bounds [15].

5.4 Conclusion

Wrapper methods are time consuming but using a cross validation method that is less demanding helped counterbalance the effects.

The results obtained by the clustering of the two principal components seem to be in agreement with what was expected to occur in relation to the way the volatiles interacted with the sensors.

Since the investigation project at Biomolecular Engineering Laboratory is still ongoing, if a new formulation of the gels is created and their responses are much different than those analysed in this work, the set of features proposed could no longer be useful and a new analysis should be made.

Finally, given that the accuracy obtained in validation was high, the selected features, the classifier and its associated parameters were accepted and used throughout the rest of this work.

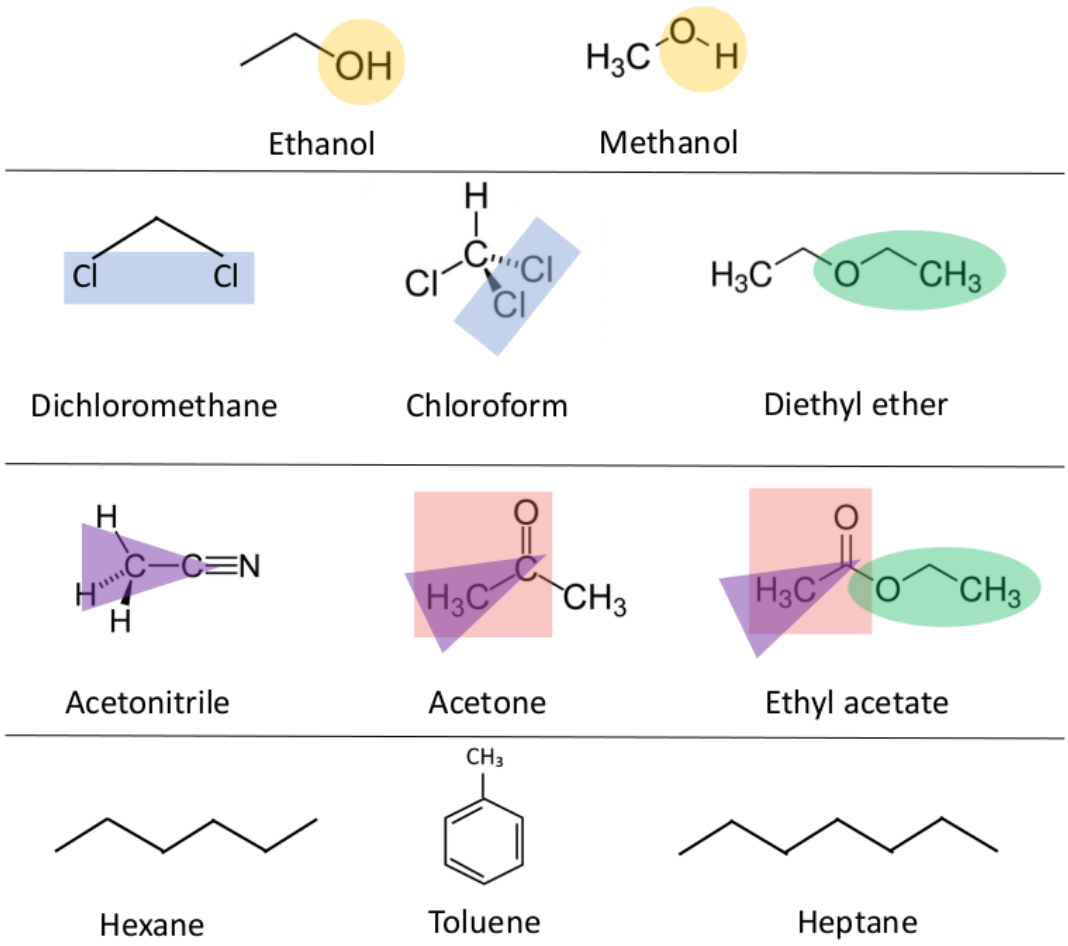


Figure 5.8: Molecular structure of the 11 VOCs used. Functional groups are represented with different shapes and colours.

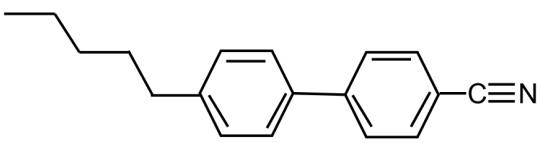


Figure 5.9: Molecular structure of 5CB.

Chapter 6

Case studies

After the definition of the classification tool, tests with different sets of sensors and gels with variations in the formulations were performed.

6.1 Introduction

Three deferent formulations were used: the standard formulation with [BMIM][DCA], 5CB and gelatine; a formulation with 1-butyl-3-methylimidazolium chloride ([BMIM][DCl]), 5CB and gelatine (different from the standard only in the IL); and a formulation with [BMIM][DCA], 5CB and starch (different from the standard only in the polymer) (Table 6.1).

Table 6.1: Different sensor formulations used. Names in bold highlight the difference from standard formulation.

Formulation	Liquid Crystal	Ionic Liquid	Polymer
Standard	5CB	[BMIM][DCA]	Gelatin
Different IL	5CB	**[BMIM][Cl]**	Gelatin
Different polymer	5CB	[BMIM][DCA]	**Starch**

Data obtained from different formulations were not combined since in this initial stage, it was important to understand the individual contribution that each type of sensor could have.

First, an attempt to mimic the commercial use of the gels was made. A classifier would be trained with data acquired in the laboratory and then new sensors would be made and sold. Differences in the gel morphology resulted in different responses, therefore this attempt was not successful. Visual comparison of the signals was made, as well as

comparing the pictures of the sensors obtained by POM, in order to find a correlation between the morphology and the signal, but conclusions where not obtained.

More data was acquired and larger datasets containing sensors with the same formulation from different batches were combined and used as input to train the classifier. This resulted in a improvement on the performance. Increasing variability in terms of gel morphology, allowed the learning model to become more generalized and overfitting was surpassed.

6.2 Methods

6.2.1 Experimental methods

Altogether, four experiments were analysed:

- Experiment 2 - 6 sensors, three made with the standard formulation and three made with a formulation that has a different IL ([BMIM][DCl]); These six sensors were exposed to a sequence of 11 VOCs on May 7 (Experiment 2.a) and again, two months later, on July 9 (Experiment 2.b). For more detail see annex III. Part of the data from this experiment (Experiment 2.a was used in the previous chapters of this thesis).

- Experiment 3 - 3 sensors, made by a different person, with the standard formulation were exposed to 4 volatiles: hexane, toluene, acetone and ethanol. For more detail see annex IV.

- Experiment 4 - 6 sensors, three made with the standard formulation that has a different IL ([BMIM][DCl]). The sensors were exposed to a sequence of 11 VOCs on July 10 (experiment 4.a) and again, one week later, on July 17 (experiment 4.b). For more detail see annex V.

- Experiment 5 - 3 sensors were made with gelatine replaced with starch were exposed to a sequence of 11 VOCs on June 12 (Experiment 5.a) and then again on June 15 (Experiment 5.b). For more detail see annex VI.

6.2.2 Computational methods

Signals from the different experiments were filtered, normalized and divided by cycles. Then, the selected features in the previous chapter (Table 6.2) were extracted. Therefore, each experiment gave origin to a dataset.

All datasets were normalized with *StandardScaler* and when there was the need to separate between training and validation set, it was made randomly but maintaining class proportion in each set.

When the signal is too constant, the algorithm is not able to extract features and it automatically ignores those responses. Sometimes, there is a small response and the algorithm is able to extract some information. For those cases, if the response is not

Table 6.2: Set of the 12 selected features to describe the cycles and to be used as input in the classifier.

Number	Description
3	Time to reach the minimum of the first derivative of the signal
6	Time to reach the maximum of the first half of the second derivative of the signal
8	Maximum value of the first part of the second derivative of the signal
9	Minimum value of the first part of the second derivative of the signal
12	Maximum value of the second part of the second derivative of the signal
13	Minimum value of the second part of the second derivative of the signal
14	Area under the signal
15	Skewness
16	Kurtosis
18	Parameter b from the first part of the signal
19	Parameter c from the first part of the signal
22	Parameter b from the second part of the signal

considered to be significant, they have to be removed manually. When extracting the data, some responses were not taken into account. It was the case for: sensor *D_69* response to acetonitrile, and sensors *C_45* and *C_48* responses for ethanol and methanol in 4.a; sensor *C_45* response for acetone, acetonitrile, ethanol, athyl acetate, dichloromethane and heptane in 4.b.

The Support Vector Classifier was used with the radial basis function kernel and parameters C: 100 and γ: 0.1. It was trained to distinguish eleven classes (hexane, heptane, toluene, diethyl ether, ethyl acetate, dichloromethane, chloroform, methanol, ethanol, acetone and acetonitrile).

6.3 Results

6.3.1 Data from sensors made with the standard formulation

In the present subsection, data extracted from Experiments 2, 3 and 4 were analysed, but only the sensors made with the standard formulation were taken into account. First, data from Experiment 2.a served as the training set (1485 examples) and data from Experiments 3 and 4.a as the testing set (873 examples). The accuracy obtained was 31.3%.

These results lead to the premise that maybe the sensors from different batches were too different, perhaps due to variability in the gel production process. Sensor pictures from before the acquisitions of the signals form Experiments 2.a and 4.a can be found in Figure 6.1. Pictures from Experiment 3 were not available.

The original pictures were processed to obtain more contrast, first the colours were inverted and then a binarization was performed. This means that the black dots are the light that passed through the cross polarizers. Sensors *D_1*, *D_3*, *D_4* from 2.a appear darker than the other three from experiment 4.b, so the possibility that the e-nose was

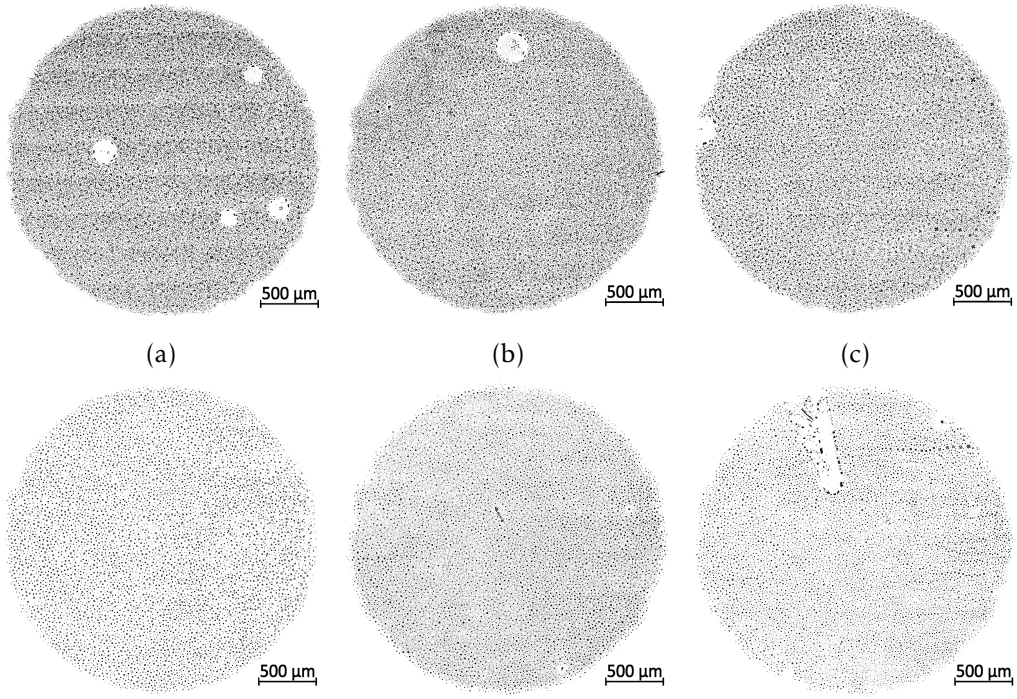


Figure 6.1: Comparison of two batches of standard sensors before they were exposed to VOC's. Sensors (a) *D_1*, (b) *D_3* and (c) *D_4* were used in Experiment 2 and sensors (d) *D_70*, (e) *D_69* and (f) *D_67* in Experiment 4.

sensible to slight differences was considered. Parallel studies were being conducted in the Laboratory to correlate the morphology of the gel with its response, namely mean gray value of the pictures correlated with baseline.

With the premise that sensors from two different batches were too different to be compared, two more acquisitions were made. The same sensors were once again exposed to the same sequence of 11 VOCs. Unfortunately, sensors from Experiment 3 were no longer available. This gave origin to Experiments 2.b and 4.b. Two tests were then made:

1. Data from Experiment 2.a as training set (1485 examples) and data from Experiment 2.b as testing set (1485 examples). The result was an accuracy of 27.9%.

2. Data from Experiment 4.a as training set (693 examples) and data from Experiment 4.b (693 examples) as testing set. The result was an accuracy of 41.1%.

There was an improvement in the accuracy regarding Experiment 4, but it was still low. Therefore, another premise was put that perhaps the VOCs were too abrasive for the gels and for that reason the way they responded to the volatiles changed after being exposed twice to the sequence of volatiles. Figure 6.2 shows an example of a before and after comparison.

Comparing the pictures, differences can be noticed, meaning that during the exposure to the volatiles, there were changes in sensors' morphology. Moreover, a parallel study

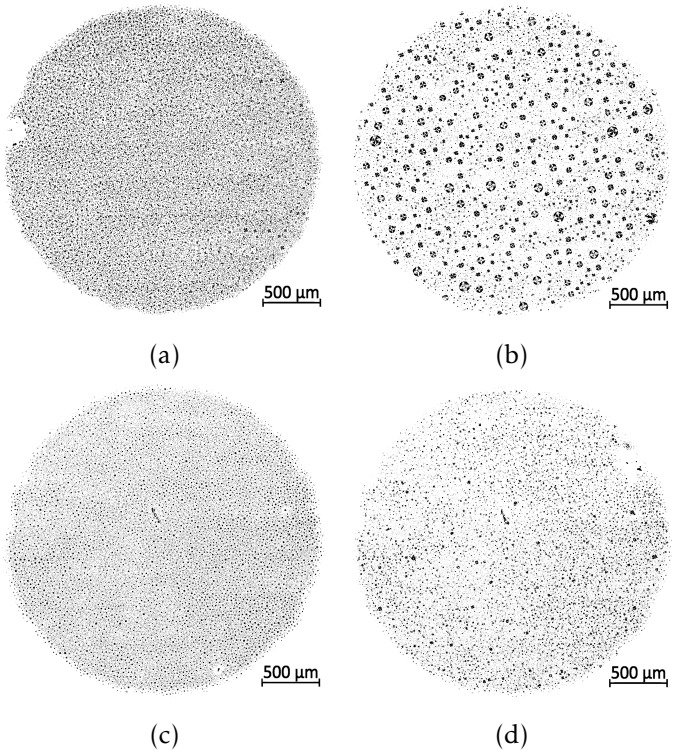


Figure 6.2: Comparison of two sensors before being exposed to any volatile and after being exposed to a sequence of 11 VOCs, twice. (a) and (b) are, respectively, the before and after of sensor *D_4*; (c) and (d) are, respectively, the before and after of sensor *D_69*.

was being conducted in the Laboratory because there was a suspicion that room conditions, such as humidity, could affect the gels morphology. Hence, it was also possible that the sensors had changed during storage.

There were also concerns about some of the sensors' responses falling out of the sensitivity range of the photodiode (0 - 2.7 V). Figure 6.3 shows an example of a signal which the sensor that may have been too dark and part of the response is lost due to saturation in the upper operating limit.

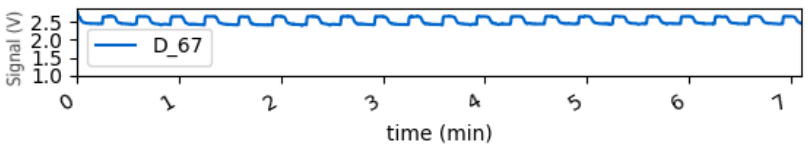


Figure 6.3: Saturated response of sensor *D_67* when exposed to acetone in Experiment 4.a.

On the other hand, it was also noted that the low accuracy could be an overfitting problem. Therefore, data from Experiments 2,3 and 4 regarding the gels made with standard formulation, was grouped together and submitted to a 10-fold validation. It was ensured that the data set had the same percentage of cycles per batch to avoid giving more weight to the responses of one batch. The resulting data set had 2143 examples,

approximately 194 per VOC. Cross validation result had an average accuracy of 89.3% with a standard deviation of 4.7%.

To be able to see what was being correctly classified, the confusion matrix was plotted (Figure 6.4).

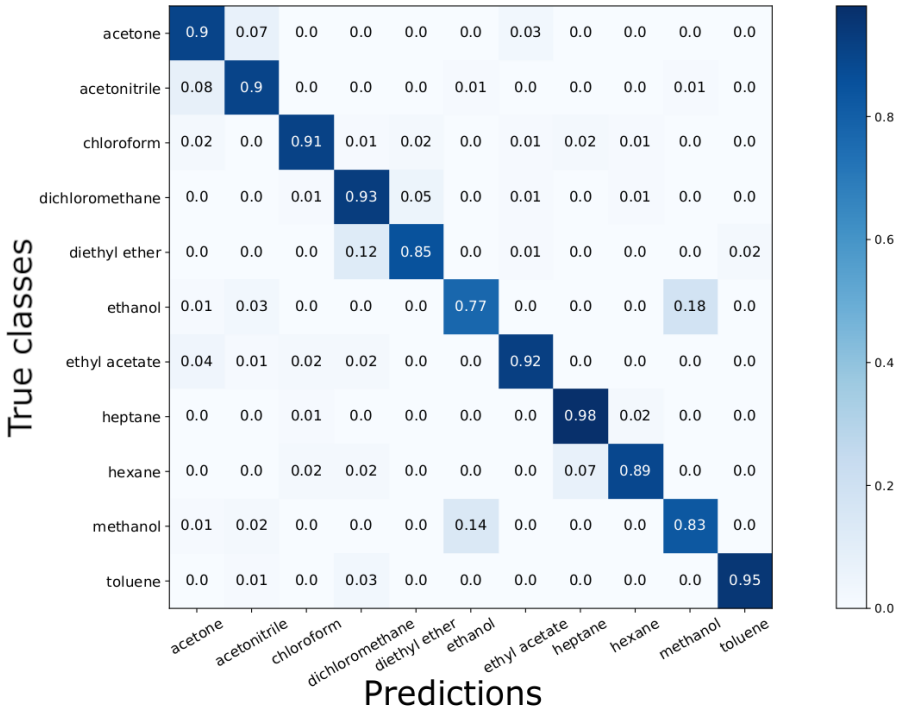


Figure 6.4: Normalized confusion matrix for the test with standard sensors. Average accuracy of 89.3%, standard deviation of 4.7%.

6.3.2 Data from sensors made with the Ionic Liquid [BMIM][Cl]

The same analysis was made for the sensors with a formulation variation in the Ionic Liquid. There was a possibility that this formulation did not present as much variability as standard sensors. Data extracted from 2 and 4 were analysed, but only the sensors made with the IL [BMIM][Cl] were taken into account.

The same tests as the previous section were performed:

1. Data from experiment 2.a as training set and data from Experiment 4.a as testing set. The result was an accuracy of 34.8%.

2. Data from experiment 2.a as training set and data from Experiment 2.b as testing set. The result was an accuracy of 21.8%.

3. Data from experiment 4.a as training set and data from Experiment 4.b as testing set. The result was an accuracy of 31.0%.

Due to low the accuracies, saturation was also a concern. Figure 6.5 shows the response from sensor *C_46* exposed to acetonitrile in experiment 4.b, the signal reaches both values of the operation limits. This could mean that information may have been lost.

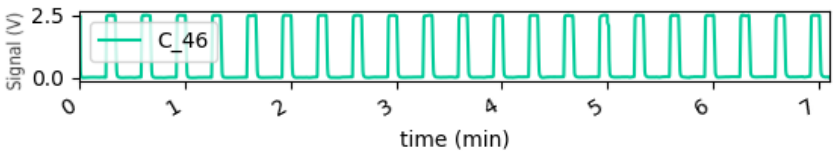


Figure 6.5: Saturated response of sensor *C_45* when exposed to acetone in Experiment 4.b.

The results also indicate that changes between batches and also between usages of the sensors were present. Figure 6.6 compares the sensors from the two batches.

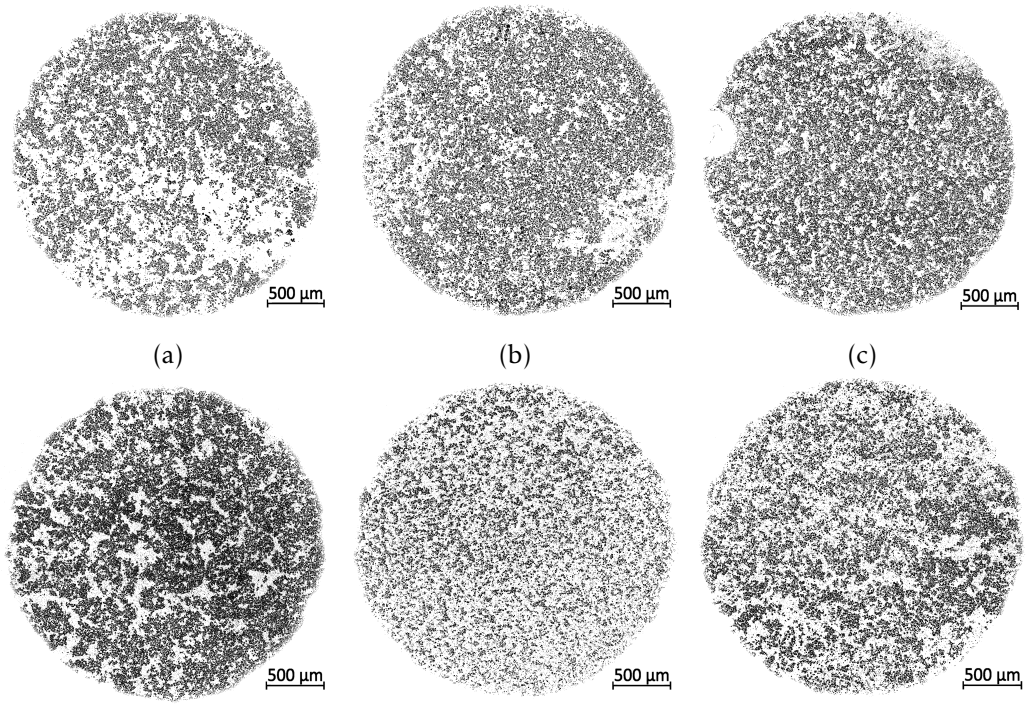
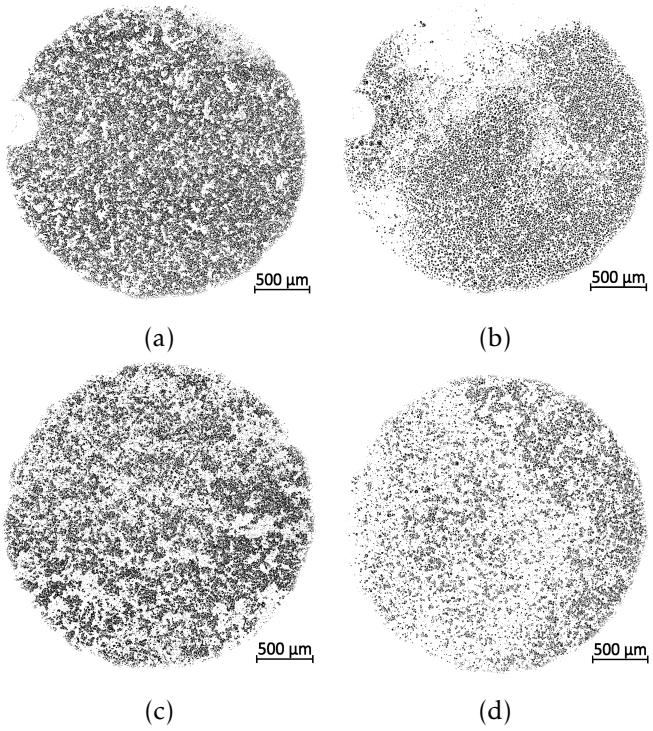


Figure 6.6: Comparison of the two batches of sensors that differ the IL from the standard formulation before they were exposed to VOC's. Sensors (a) *C_1*, (c) *C_2* and (c) *C_4* were used in Experiment 2 and sensors (d) *C_48*, (e) *C_46* and (f) *C_45* in Experiment 4.

Figure 6.7 shows the differences of two sensors before any exposure and after being exposed twice to a sequence of 11 VOCs. One of the sensors is from the batch from Experiment 2 and the other from experiment 4. With resemblance with what happened with the standard sensors, differences between the before and after are noticeable.

Still, when training the classifier with data from only one acquisition, overfitting occurs. Therefore, data regarding the sensors with the IL [BMIM][Cl] from Experiments



(a) (b)

(c) (d)

Figure 6.7: Comparison of two sensors before being exposed to any volatile and after being exposed to a sequence of 11 VOCs, twice. (a) and (b) are the before and after of sensor *C_4*; (c) and (d) are the before and after of sensor *C_45*.

2 and 4 was aggregated and used in the same way as in the previous section. In this case a total of 2434 examples were used, approximately 221 per VOC. The result had a mean accuracy score of 87.1% with a standard deviation of 6.9%. The confusion matrix is presented in Figure 6.8.

6.3.3 Data from sensors made with starch instead of gelatine

Lastly, the responses of a formulation with starch substituting the gelatine were analysed. Two test were performed. Data from experiment 5.a was used as the training set and data from experiment 5.b as testing set. The result was an accuracy of 65.9%.

The result was satisfactory, but once again data from both experiments was mixed (264 examples per VOC) and a cross validation was performed. This resulted in an average accuracy of 91.2% with standard deviation of 1.5%. The normalized confusion matrix is presented in Figure 6.9.

6.4 Discussion of results

The test made with different batches (Table 6.3) show that there is some variability even if the batches are made by the same person.

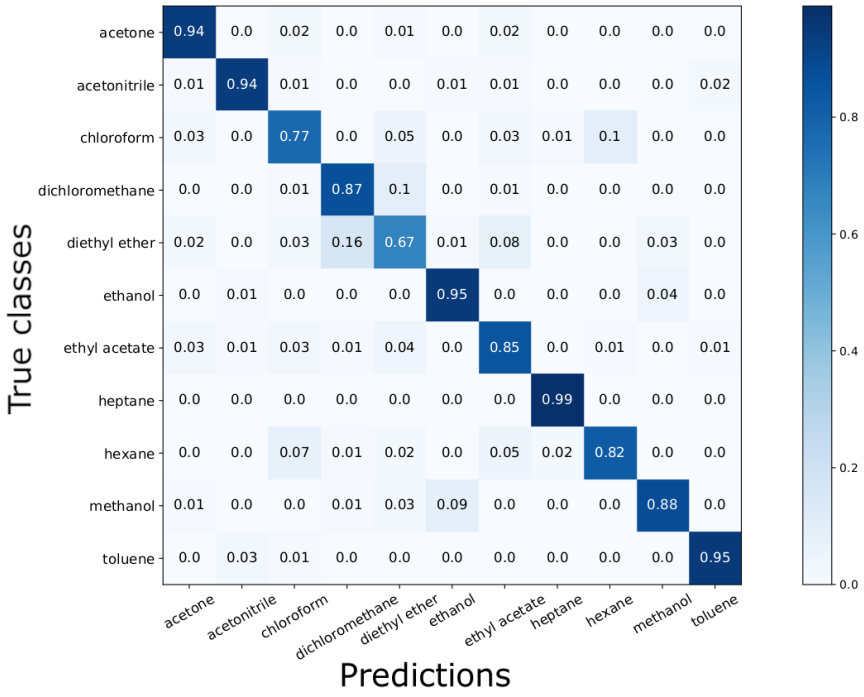


Figure 6.8: Normalized confusion matrix made for the sensors that differ from the standard formulation in the IL. Average accuracy of 87.1%, standard deviation of 6.9%.

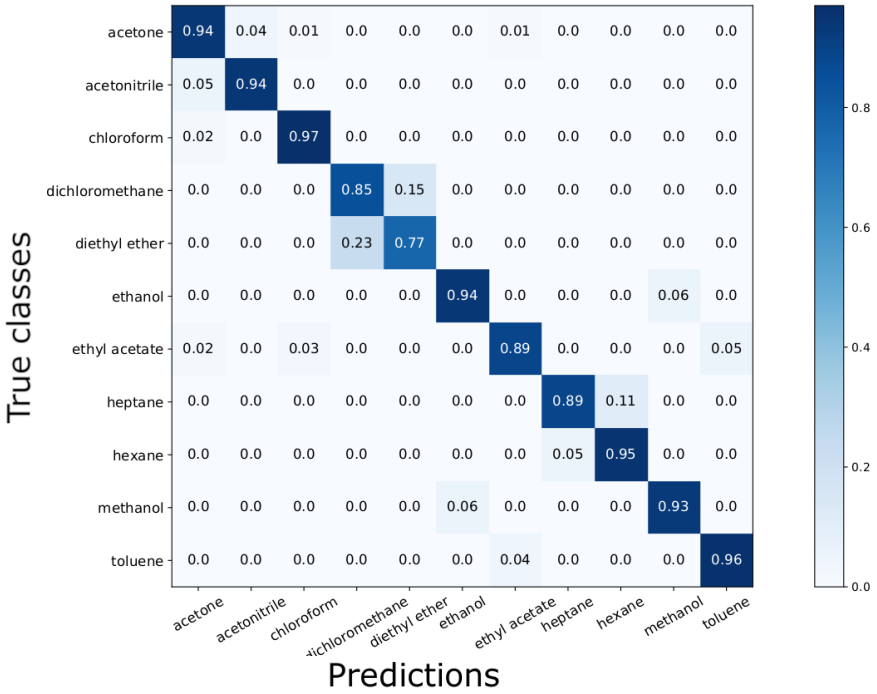


Figure 6.9: Normalized confusion matrix made for the sensors that have starch instead of gelatine. Average accuracy of 91.2%, standard deviation of 1.5%.

The first test where features extracted from Experiment 2.a were used as training and features extracted from Experiments 3 and 4.a as testing, was an attempt to mimic

Table 6.3: Scores obtained by comparing to different batches of sensors.

Train set	Test set	Score (%)	Formulation
Experiment 2.a	Experiments 3 and 4.a	31.3	Standard
Experiment 2.a	Experiment 4.a	34.8	Different IL

commercial use. New sensors were expected to respond the same way, and it is true that the responses obtained from Experiment 3 are similar to the ones from Experiment 2.a. However, signals from Experiment 4.a were saturated, therefore the results were not as good as desired.

Moreover, the experiments made with the same batch of sensors exposed twice to the sequence of VOCs in different point of time (Table 6.4), evidence that changes occur in the sensors during storage.

Table 6.4: Scores obtained from the same batch of sensors.

Storage time	Exposure time	Train set	Test set	Score (%)	Formulation
2 months	15 min	Experiment 2.a	Experiment 2.b	27.8	Standard
2 months	15 min	Experiment 2.a	Experiment 2.b	21.8	Different IL
1 week	7 min	Experiment 4.a	Experiment 4.b	41.1	Standard
1 week	7 min	Experiment 4.a	Experiment 4.b	31.0	Different IL
3 days	15 min	Experiment 5.a	Experiment 5.b	65.9	Different polymer

It was not strange to obtain low accuracies in relation to these tests due to differences in the signals when analysing them visually. For example, from Experiment 2.a to Experiment 2.b (See annex III) the responses from hexane are completely different, but the responses of the other VOCs are very similar. In Experiment 2.b it can be noticed that after the exposure to diethyl ether there is a saturation on the upper operating limit; in experiment 4 (See annex V), responses from heptane, hexane, and ethanol are also very distinct from Experiment 4.a to Experiment 4.b, either in terms of shape and also their baseline. Also, while in Experiment 4.a signals seem to saturate in the upper limit, in Experiment 4.b the baseline is lower, making some signals from gels with IL [BMIM][Cl] saturate in the lower limit; finally, comparing signals from Experiment 5.a and 5.b (See annex VI), what is observed is that baseline and wave format are almost the same.

Due to the fact that the gels from different experiments were stored and exposed to volatiles during different periods of time, conclusions about the effect of time and VOC abrasion on gel morphology could not be withdrawn from this work. Sensors from Experiments 2 were stored for two months and exposed to each VOC for 15 minutes, the ones from Experiment 4 were stored for one week and exposed to each VOC for 7

minutes, and the sensors made with starch (Experiment 5) were stored during three days and exposed to each VOC for 15 minutes.

Moreover, it also worth mentioning that the researchers stated that during the experiments, strong smells could be scented, meaning that the system may have had leaks. Changes in the concentration that arrives to the chamber could also have significant effects on the responses. Since there are too many variables, the reasons behind the responses' differences could not be concluded from this work.

Although saturation issues occurred and that morphological modifications due to storage conditions and exposure to the VOCs were noted, good results were obtained when data from different batches of sensors was combined (Table 6.5). Meaning that although the e-nose is sensitive to changes in the morphology, with enough data, machine learning is able to overcome this and find patterns that distinguish the 11 VOCs.

Table 6.5: Cross valitadion scores of datasets obtained from different experiments.

Validation method	Gathered data	Score (%)	Formulation
10-fold cross validation	Experiments 2, 3 and 4	89.3	Standard
10-fold cross validation	Experiment 2 and 4	87.1	Different IL
10-fold cross validation	Experiment 5	91.2	Different polymer

All types of sensors seem to be able to distinguish all different samples. Moreover, the VOCs that have similar ways of interacting with the gels, are the ones that are more frequently misclassified, as one can see from the confusion matrices of the three types of sensors (Figures 6.4, 6.8 and 6.9). Methanol and ethanol get mistaken from one another, and dichloromethane from diethyl ether as well. On the other hand, heptane, hexane and toluene are never mistaken for ethanol or methanol.

No conclusion about which type of sensor could be the best to distinguish VOCs can be inferred since the datasets did not have all the same variability in terms of sensor batches. Classification for the standard gels were made with 3 different batches, for the gels with the IL [BMIM][Cl] 2 batches were used and for the gels made with starch only one batch was used to acquire signals.

Starch, such as Liquid Crystal, also has birefringence, but when exposed to the volatiles it does not respond. This means that this type of sensor never gets completely dark, thus making saturation in the upper limit of the scale less probable. The analysis for this type of gel is only interesting as preliminary work. The sensors were able to keep the same response (See annex VI) but they were only stored for three days.

6.5 Conclusions

The attempt to mimic commercial use failed because the classifier was not trained with data with enough variability. Overfitting was the main problem when trying to obtain a good classification score. Although morphological differences inherent to sensing gels can

result in different responses, a more generalized classifier can be obtained when adding some variability in the training set.

On the other hand, while comparing different batches of sensors, differences in their responses was evidenced. This confirms the need to put bigger efforts in the understanding of the dynamics of hybrid gas-sensing gels. Mainly, in terms of response saturation to avoid losing information. Also, what could be improved in terms of the production method reproducibility.

It would also be preferable to have experiments made under the same conditions for the result relative to storage and morphology to be conclusive.

In relation to the different types of formulations, all of them have the ability to respond differently to each volatile even when their chemical structure interacts with the gel in a similar way.

Chapter 7

Concluding Remarks and Future Work

Establishing methods for signal pre-processing is essential for data analysis. In this work, a set of functions were proposed to treat data acquired with the electronic nose that is being developed in the Biomolecular Engineering Laboratory. The tools allowed to confirm that sensors are able to maintain a stable response for long periods of time. In addition, the developed tools will allow that in the future, e-nose users in the laboratory use the same standard filtering and normalization conditions when treating their data. This together with the established signal acquisition protocol will allow data acquired by different team members to be comparable.

In relation to the classification tools, feature extraction and feature selection are usually seen as the key factors to obtain a good performance. In the present work, the proposed methods for feature extraction were a combination of features that can be directly extracted from the signal (morphological features) and parameters that are estimated trough curve fitting. However, not all features extracted were used. A recursive feature elimination algorithm with 10-fold cross validation was used to select the subset that best characterized the signals' cycles.

After that, the hyperparameters of the chosen classifier (Support Vector Classifier) were tuned and the best combination was validated with unseen data.

Then, established the preprocessing and classification tools were used in the analysis of data obtained from different experiments. At first, a good performance was not being achieved, and an attempt was made to relate the differences in the signals with the differences in gel morphology that could be caused by ambient conditions or abrasion of the VOCs. No conclusion could be drawn due to the fact that there were too many variables to consider, namely different storage and exposure periods. What was found was that the main problem was overfitting.

The effects of overfitting were surpassed when more data was considered. The results

improved, meaning that acquiring data from more batches of gels, allowed the generalization of the model and it became more prepared for variability.

It was clear that although some VOC's were frequently mistaken, there were others which were efficiently separated. This had to do with their different polarities and subsequent ways of interacting with the hybrid sensing gels. The goal to distinguish 11 volatile organic compounds was successfully fulfilled.

Counterbalancing all the factors to avoid overfitting and achieve a good performance is important. Efforts were made to meet a balance that satisfied those demands. Besides the fact that the algorithm is able to classify all the classes, other important input was given regarding the current state of the development of the sensors. Efforts are being made in order to find a way to produce gels that are more stable. Either in the production process or in the component's selection. Different ionic liquids, liquid crystals and biopolymers are being tested. The acquisition device is also being subjected to improvements.

Before reaching the final goal of identifying bacteria, every aspect of the electronic nose must be tuned. Since it is a project where things are being developed in parallel, improvements are always being made and the computational tools, resulting from this work, can be used to analyse data from future experiments.

7.1 Future Work

Given that data sets used in this work were not polished, some outliers could be adding noise into the classifier. More data should be acquired under equal conditions and tuned to create a solid database for future training.

In future works, a frequency-domain method could be used to extract features and compare the results to the ones obtained by the proposed method in this chapter.

There could be a better classifier to distinguish the patterns made by the volatiles. Therefore, other learning models, such as neural networks or decision trees, should also be tested and results compared to the obtained in this thesis.

Normally, electronic noses are composed by an array of sensors, with different specificities. In this work, only one type of sensor was considered in each analysis. The combination of different types of sensors should be made to increase selectivity.

Another thing to have into consideration is that an interface should be developed in order to allow data analysis by users that are not as familiar with programming.

Bibliography

[1] World Health Organization. *Every infection prevented is an antibiotic treatment avoided*. 2017. URL: http://www.euro.who.int/en/health-topics/disease-prevention/antimicrobial-resistance/news/news/2017/11/every-infection-prevented-is-an-antibiotic-treatment-avoided (visited on 01/18/2018).

[2] C. Morel et al. *Ensuring innovation in diagnostics for bacterial infection*. Vol. 44. European Observatory on Health Systems and Policies, 2016, pp. 1–40. ISBN: 978 92 890 50364.

[3] T. Wasilewski et al. "Bioelectronic nose: Current status and perspectives." In: *Biosensors and Bioelectronics* 87 (2017), pp. 480–494. ISSN: 18734235. DOI: 10.1016/j.bios.2016.08.080.

[4] F. Röck et al. "Electronic nose: Current status and future trends." In: *Chemical Reviews* 108.2 (2008), pp. 705–725. ISSN: 00092665. DOI: 10.1021/cr068121q.

[5] J. W. G. Tim C. Pearce, Susan S. Schiffman, H. Troy Nagle. *Handbook of Machine Olfaction: Electronic Nose Technology*. First. Wiley-VHC, 2003. ISBN: 3-527-30358-8.

[6] S. Qiu et al. "The prediction of food additives in the fruit juice based on electronic nose with chemometrics." In: *Food Chemistry* 230 (2017), pp. 208–214. ISSN: 0308-8146. DOI: 10.1016/J.FOODCHEM.2017.03.011.

[7] H. L. Ramírez et al. "Evaluation of the Food Sniffer electronic nose for assessing the shelf life of fresh pork meat compared to physicochemical measurements of meat quality." In: *European Food Research and Technology* 244.6 (2017), pp. 1047–1055. ISSN: 14382385. DOI: 10.1007/s00217-017-3021-0.

[8] L. Capelli et al. "Monitoring odour emisssions from an oil & gas plant: Electronic nose performance testing in the field." In: *ISOCS/IEEE International Symposium on Olfaction and Electronic Nose, Proceedings* (2017), pp. 1–3. DOI: 10.1109/ISOEN.2017.7968862.

[9] S. Cociorva et al. "Indoor Air Quality Evaluation in Intelligent Building." In: *Energy Procedia* 112 (2017), pp. 261–268. ISSN: 18766102. DOI: 10.1016/j.egypro.2017.03.1095.

[10] M. K. Nakhleh et al. "Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules." In: *ACS Nano* 11.1 (2016), pp. 112–125. ISSN: 1936086X. DOI: 10.1021/acsnano.6b04930.

[11] R. Coronel Teixeira et al. "The potential of a portable, point-of-care electronic nose to diagnose tuberculosis." In: *Journal of Infection* 75.5 (2017), pp. 441–447. ISSN: 15322742. DOI: 10.1016/j.jinf.2017.08.003.

[12] T. Saidi et al. "Exhaled breath analysis using electronic nose and gas chromatography–mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects." In: *Sensors and Actuators, B: Chemical* 257 (2018), pp. 178–188. ISSN: 09254005. DOI: 10.1016/j.snb.2017.10.178.

[13] H. Cruz et al. "Development of e-nose biosensors based on organic semiconductors towards low-cost health care diagnosis in gynecological diseases." In: *Materials Today: Proceedings* 4.11 (2017), pp. 11544–11553. ISSN: 22147853. DOI: 10.1016/j.matpr.2017.09.065.

[14] T. Seesaard et al. "Self-screening for diabetes by sniffing urine samples based on a hand-held electronic nose." In: *BMEiCON 2016 - 9th Biomedical Engineering International Conference* (2017). DOI: 10.1109/BMEiCON.2016.7859586.

[15] A. Hussain et al. "Tunable Gas Sensing Gels by Cooperative Assembly." In: *Advanced Functional Materials* 27.27 (2017), pp. 1–9. ISSN: 16163028. DOI: 10.1002/adfm.201700803.

[16] ISO. *Terms and definitions - volatile organic compound*. 2011. URL: https://www.iso.org/obp/ui/{\#}iso:std:iso:12219:-2:ed-1:v1:en:term:3.4 (visited on 01/18/2018).

[17] S. I.C. J. Palma et al. "Machine learning for the meta-analyses of microbial pathogens´ volatile signatures." In: *Scientific Reports* 8.3360 (2018), pp. 1–15. ISSN: 20452322. DOI: 10.1038/s41598-018-21544-1.

[18] M. Trincavelli et al. "Direct Identification of Bacteria in Blood Culture Samples Using an Electronic Nose." In: *IEEE Transactions on Biomedical Engineering* 57.12 (2010), pp. 2884–2890. DOI: 10.1109/TBME.2010.2049492.

[19] R. Dutta et al. "Identification of Staphylococcus aureus infections in hospital environment: Electronic nose based approach." In: *Sensors and Actuators, B: Chemical* 109.2 (2005), pp. 355–362. ISSN: 09254005. DOI: 10.1016/j.snb.2005.01.013.

[20] L. D. J. Bos et al. "Volatile Metabolites of Pathogens: A Systematic Review." In: *PLoS Pathogens* 9.5 (2013), pp. 1–8. ISSN: 15537366. DOI: 10.1371/journal.ppat.1003311.

[21] A. C. Pádua et al. "Design and Evolution of an Opto-electronic Device for VOCs Detection." In: *Biomedical engineering systems and technologies, international joint conference, BIOSTEC* (2018), pp. 48 –55. DOI: 10.5220/0006558100480055.

[22] J. P. P. G. de Gennes. "Liquid Crystals: main types and properties." In: *The Physics of Liquid Cristal*. Second Edi. Oxford University Press, 1972. Chap. Chapter 1. ISBN: 0198520247.

[23] J.-G. An et al. "Characterization of Liquid Crystals: A Literature Review." In: *Advanced Materials Science* 44.4 (2016), pp. 398–406. ISSN: 16058127.

[24] M. Widom et al. *Liquid crystal*. 2013. URL: https://www.britannica.com/science/liquid-crystal (visited on 01/01/2018).

[25] D. S. Miller et al. "Introduction to optical methods for characterizing liquid crystals at interfaces." In: *Langmuir : the ACS journal of surfaces and colloids* 29.10 (2013), pp. 3154–69. ISSN: 1520-5827. DOI: 10.1021/la304679f. arXiv: NIHMS150003.

[26] M. Watanabe. "Design and Materialization of Ionic Liquids Based on an Understanding of Their Fundamental Properties." In: *Electrochemistry* 84.9 (2016), pp. 642–653. ISSN: 1344-3542. DOI: 10.5796/electrochemistry.84.642.

[27] E. Alpaydin. *Introduction to Machine Learning*. Second Edi. The MIT Press, 2010. ISBN: 978-0-262-01243-0.

[28] R. Sebastian. *Python Machine Learning*. Packt Publishing Ltd., 2015. ISBN: 978-1-78355-513-0.

[29] S. G. Andreas C. Müller. "Introduction to Machine Learning with Python." In: *Introduction to Machine learning with Python*. First Edi. O'Reilly Media, Inc, 2016. ISBN: 978-1-449-36941-5.

[30] J. Yan et al. "Electronic Nose Feature Extraction Methods: A Review." In: *Sensors* 15.11 (2015), pp. 27804–27831. ISSN: 1424-8220. DOI: 10.3390/s151127804.

[31] D. Chicco. "Ten quick tips for machine learning in computational biology." In: *BioData Mining* 10.1 (2017), p. 35. ISSN: 1756-0381. DOI: 10.1186/s13040-017-0155-3.

[32] Y. Saeys et al. "A review of feature selection techniques in bioinformatics." In: *Bioinformatics* 23.19 (2007), pp. 2507–2517. ISSN: 13674803. DOI: 10.1093/bioinformatics/btm344. arXiv: bioinformatics/btm344 [10.1093].

[33] A. Jovic et al. "A review of feature selection methods with applications." In: *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2015, pp. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.

[34] P. Ramanathan et al. "Smart controller for conical tank system using reinforcement learning algorithm." In: *Measurement* 116 (2017), pp. 422–428. ISSN: 02632241. DOI: 10.1016/j.measurement.2017.11.007.

[35] T. Fawcett. "An introduction to ROC analysis." In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010. eprint: /dx.doi.org/10.1016/j.patrec.200.

[36] A. D. Wilson et al. "Advances in electronic-nose technologies developed for biomedical applications." In: *Sensors* 11.1 (2011), pp. 1105–1176. ISSN: 14248220. DOI: 10.3390/s110101105.

[37] M. Moens et al. "Fast identification of ten clinically important micro-organisms using an electronic nose." In: *Letters in Applied Microbiology* 42.2 (2006), pp. 121–126. ISSN: 02668254. DOI: 10.1111/j.1472-765X.2005.01822.x.

[38] W. H. van Geffen et al. "Diagnosing viral and bacterial respiratory infections in acute COPD exacerbations by an electronic nose: a pilot study." In: *Journal of Breath Research* 10.3 (2016), p. 036001. ISSN: 1752-7163. DOI: 10.1088/1752-7155/10/3/036001.

[39] J. Ritaban Dutta*, Evor L Hines et al. "Bacteria classification using Cyranose 320 electronic nose." In: *BioMedical Engineering OnLine* 7.1 (2002). DOI: 10.1186/1475-925X-1-4.

[40] J. W. T. Yates et al. "Data reduction in headspace analysis of blood and urine samples for robust bacterial identification." In: *Computer Methods and Programs in Biomedicine* 79 (2005), pp. 259–271. DOI: 10.1016/j.cmpb.2005.04.003.

[41] V. S. Kodogiannis et al. "Artificial Odor Discrimination System Using Electronic Nose and Neural Networks for the Identification of Urinary Tract Infection." In: *IEEE Transactions on Information Technology in Biomedicine* 12.6 (2008), pp. 707–713. DOI: 10.1109/TITB.2008.917928.

[42] S Aathithan et al. "Diagnosis of Bacteriuria by Detection of Volatile Organic Compounds in Urine Using an Automated Headspace Analyzer with Multiple Conducting Polymer Sensors." In: *Journal of Clinical Microbiology* 39.7 (2001), pp. 2590–2593. DOI: 10.1128/JCM.39.7.2590.

[43] A. Roine et al. "Rapid and Accurate Detection of Urinary Pathogens by Mobile IMS-Based Electronic Nose : A Proof-of-Principle Study." In: *PLOS ONE* 9.12 (2014), pp. 1–11. DOI: 10.1371/journal.pone.0114279.

[44] K. S. Suslick et al. "Seeing Smells: Development of an Optoelectronic Nose." In: *Quimica Nova* 30.3 (2007), pp. 677–681. DOI: 10.1590/S0100-40422007000300029.

[45] S.-w. Chiu et al. "A Fully Integrated Nose-on-a-Chip for Rapid Diagnosis of Ventilator-Associated Pneumonia." In: *IEEE Transactions on Biomedical Circuits and Systems* 8.6 (2015), pp. 765–778. DOI: 10.1109/TBCAS.2014.2377754.

[46] H. Sun et al. "Sensor Array Optimization of Electronic Nose for Detection of Bacteria in Wound Infection." In: *IEEE Transactions on Industrial Electronics* 64.9 (2017), pp. 7350–7358. ISSN: 02780046. DOI: 10.1109/TIE.2017.2694353.

[47] B. H. Aksebzeci et al. "Classification of root canal microorganisms using electronic-nose and discriminant analysis." In: *BioMedical Engineering Online* 9.77 (2010). ISSN: 1475925X. DOI: 10.1186/1475-925X-9-77.

[48] Z. Liang et al. "A correlated information removing based interference suppression technique in electronic nose for detection of bacteria." In: *Analytica Chimica Acta* 986 (2017), pp. 145–152. ISSN: 18734324. DOI: 10.1016/j.aca.2017.07.028.

[49] R. Dutta et al. "Intelligent Bayes Classifier (IBC) for ENT infection classification in hospital environment." In: *BioMedical Engineering Online* 5 (2006). ISSN: 1475925X. DOI: 10.1186/1475-925X-5-65.

[50] J. Feng et al. "A background elimination method based on wavelet transform in wound infection detection by electronic nose." In: *Sensors and Actuators, B: Chemical* 157.2 (2011), pp. 395–400. ISSN: 09254005. DOI: 10.1016/j.snb.2011.04.069.

[51] Q. He et al. "Classification of Electronic Nose Data in Wound Infection Detection Based on PSO-SVM Combined with Wavelet Transform." In: *Intelligent Automation and Soft Computing* 18.7 (2012), pp. 967–979. ISSN: 10798587. DOI: 10.1080/10798587.2012.10643302.

[52] J. Serrà et al. "An Empirical Evaluation of Similarity Measures for Time Series Classification." In: *Knowledge-Based Systems* 67 (2014), pp. 305–314. DOI: 10.1016/j.knosys.2014.04.035.

[53] M. R. Berthold et al. "On Clustering Time Series Using Euclidean Distance and Pearson Correlation." In: *CoRR* abs/1601.02213 (2016). arXiv: 1601.02213. URL: http://arxiv.org/abs/1601.02213.

[54] C. A. R. Eamonn Keogh. "Exact indexing of dynamic time warping." In: *Knowledge and Information Systems* 7.February 2003 (2005), pp. 358–386. ISSN: 14451336. DOI: 10.1007/s10115-004-0154-9. arXiv: 1201.2969.

[55] Y. Ye et al. "A Shape Based Similarity Measure for Time Series Classification with Weighted Dynamic Time Warping Algorithm." In: *4th International Conference on Information Science and Control Engineering (ICISCE)*. 2017, pp. 104–109. DOI: 10.1109/ICISCE.2017.32.

[56] B. Bengfort et al. *Yellowbrick*. Version 0.6. Mar. 17, 2018. DOI: 10.5281/zenodo.1206264. URL: http://www.scikit-yb.org/en/latest/.

[57] Scikit-learn. *Support Vector Classification*. URL: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (visited on 09/19/2018).

[58] T. Hastie et al. *The Elements of Statistical Learning*. Second Edi. Springer New York Inc., 2001.

[59] NIST/SEMATECH. *Engineering statistics handbook, Least Squares*. URL: http://www.itl.nist.gov/div898/handbook/ (visited on 08/15/2018).

[60] M. Holmberg et al. "Bacteria classification based on feature extraction from sensor data." In: *Biotechnology Techniques* 12.4 (1998), pp. 319–324. ISSN: 1573-6784. DOI: 10.1023/A:1008862617082. URL: https://doi.org/10.1023/A:1008862617082.

ANNEX I

Experiment 1

This experiment was conducted by PhD student Carolina Pádua. Three sensors were exposed to a VOC during a period of 100 minutes. This was repeated six times for six different VOCs. Figures I.1 to I.12 correspond to the plot of the responses acquired. Figures I.7 to I.12 show the overlaped cycles of the sensors. The sensors of this batch are labeled *cX*, where *X* is the unique identifier.

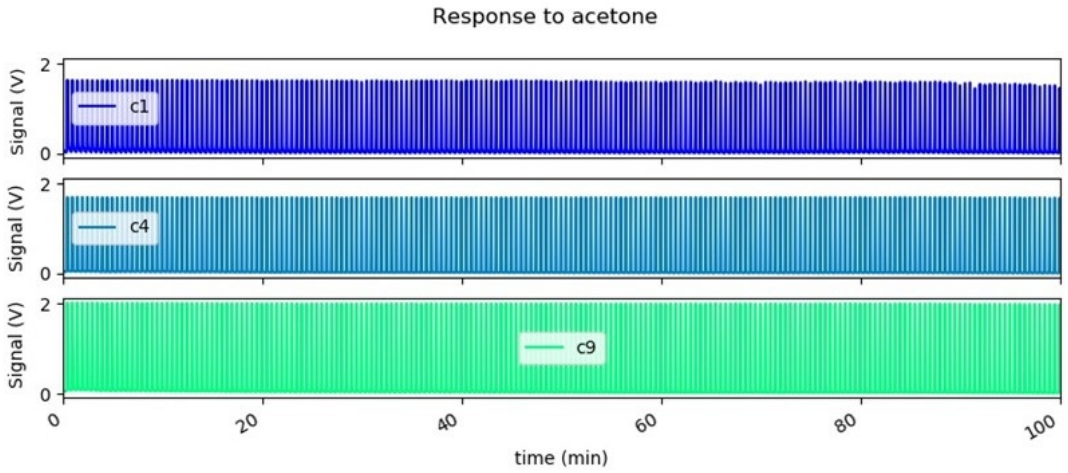


Figure I.1: Responses of all sensors exposed to acetone: (a) Experiment 1.a; (b) Experiment 1.b

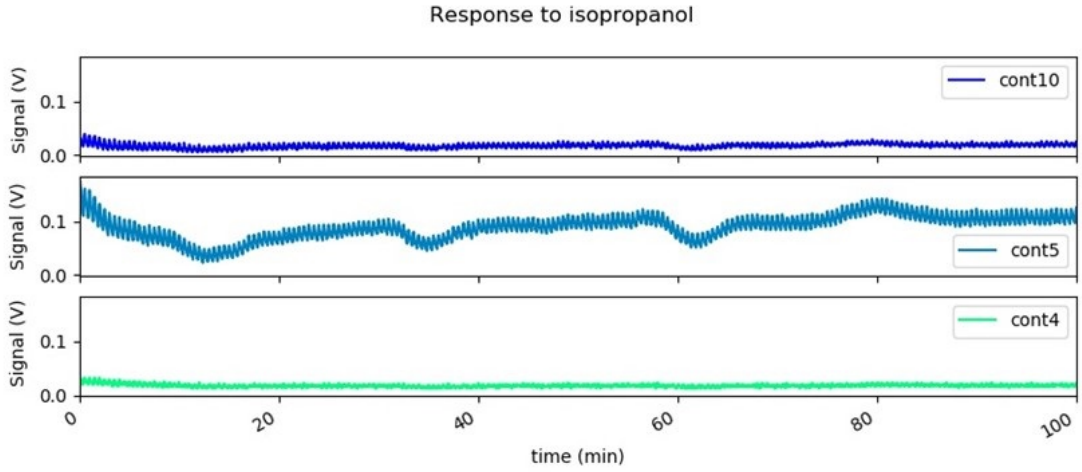


Figure I.2: Responses of all sensors exposed to isopropanol: (a) Experiment 1.a; (b) Experiment 1.b

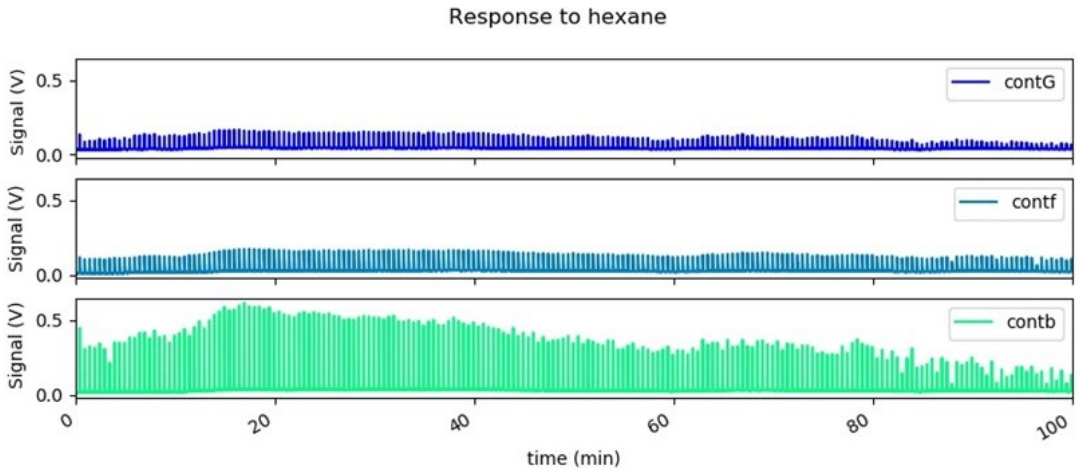


Figure I.3: Responses of all sensors exposed to hexane: (a) Experiment 1.a; (b) Experiment 1.b

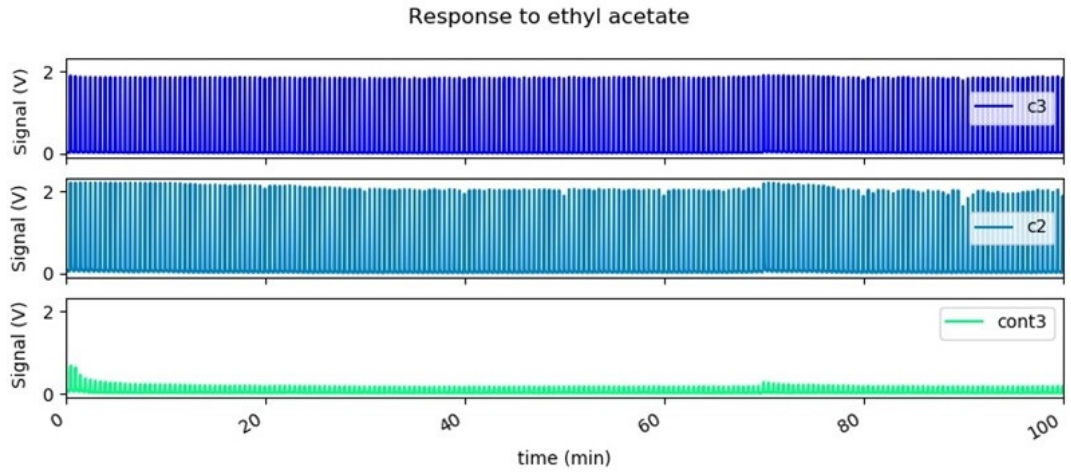


Figure I.4: Responses of all sensors exposed to ethyl acetate: (a) Experiment 1.a; (b) Experiment 1.b

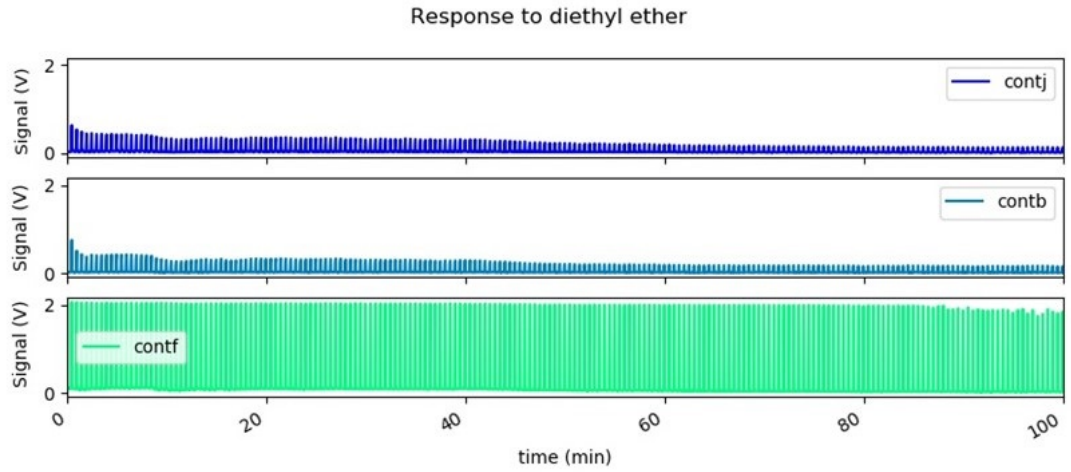


Figure I.5: Responses of all sensors exposed to diethyl ether: (a) Experiment 1.a; (b) Experiment 1.b

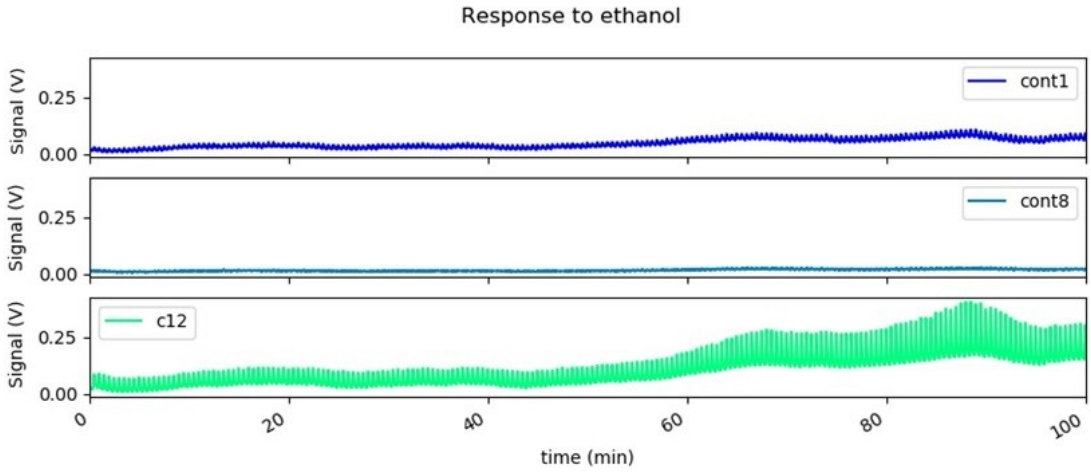


Figure I.6: Responses of all sensors exposed to ethanol: (a) Experiment 1.a; (b) Experiment 1.b

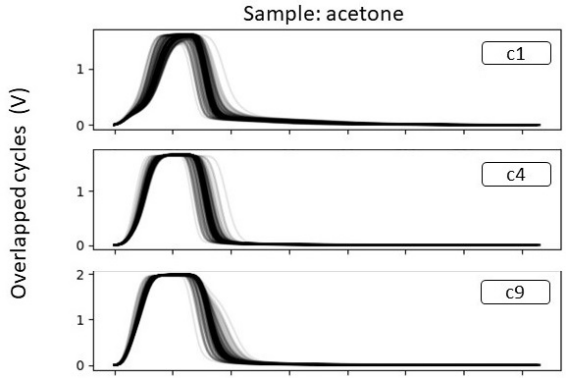


Figure I.7: Experiment A - Overlaped cycles of all sensors exposed to acetone.

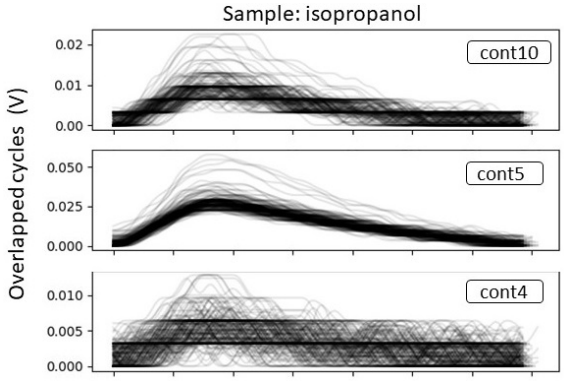


Figure I.8: Experiment A - Overlaped cycles of all sensors exposed to isopropanol.

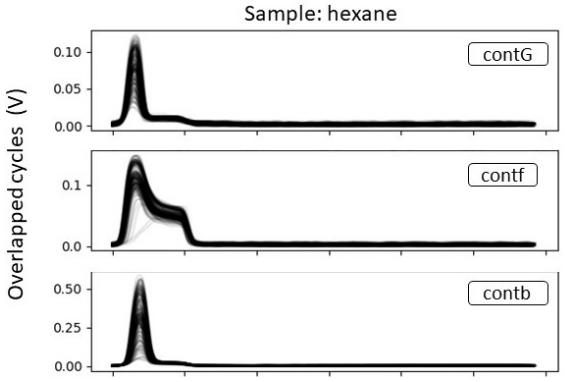


Figure I.9: Experiment A - Overlaped cycles of all sensors exposed to hexane.

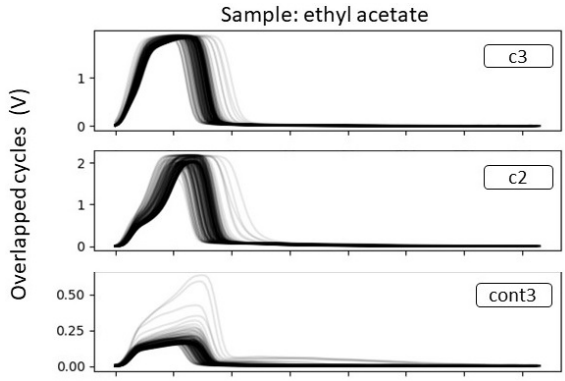


Figure I.10: Experiment A - Overlaped cycles of all sensors exposed to ethyl acetate.

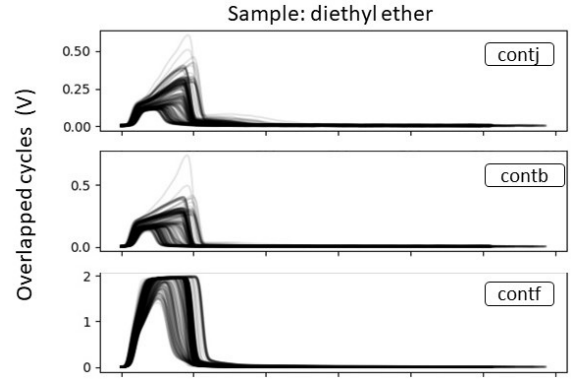


Figure I.11: Experiment A - Overlaped cycles of all sensors exposed to diethyl ether.

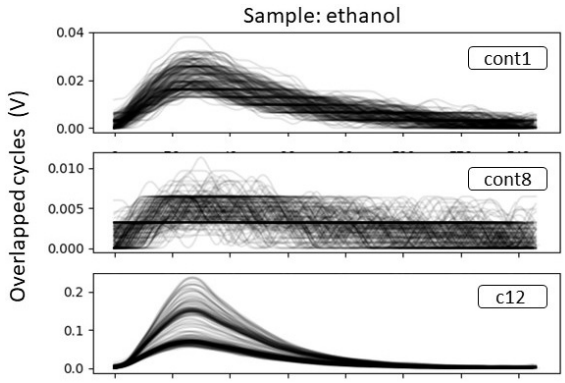


Figure I.12: Experiment A - Overlaped cycles of all sensors exposed to ethanol.

ANNEX II

Experimental protocol

Firstly, a batch of sensors must be prepared. Normally, the resultant quantity of the standard formulation is enough for it to be distributed into 6 slides. Their baseline is measured in the electronic nose to choose the ones that are good to use. The baseline must not be close to the operating limits of the sensor (0 - 2.7 V). This is made to try to ensure that when the sensor is exposed to the samples, its response lies between operating limits of the system.

After the sensors are chosen, the acquisition can begin. The quantity of the samples must be 15 mL. It must be heated at 37 degrees Celsius during 15 minutes before starting the experiment, then it must be maintained at that temperature till the end of acquisition. This quantity of volatile ensures that saturation concentrations are achieved. The relative concentration of the samples in the chamber were estimated by team member Henrique Costa (Table II.1).

Table II.1: Relative concentrations of sample in the e-nose chamber.

VOC	Average Concentration (% (v/v))
Diethyl Ether	22
Hexane	20
Dichloromethane	17
Ethyl Acetate	13
Acetone	15
Ethanol	13
Toluene	15
Acetonitrile	13
Chloroform	20
Heptane	15
Methanol	20

Acquisition time is made with cycles of five seconds of exposure and fifteen of recovery. The sampling rate used is 90Hz.

Through videos made with with POM by team member João Filho it was possible to see that some volatiles changed the gel. For this reason it was decided that when a sensor is exposed to more than one VOC, a sequence must be followed: heptane, hexane, toluene, chloroform, dichloromethane, diethyl ether, ethyl acetate, acetone, acetonitrile, ethanol, methanol. The compounds are ordered from the least abrasive to the most.

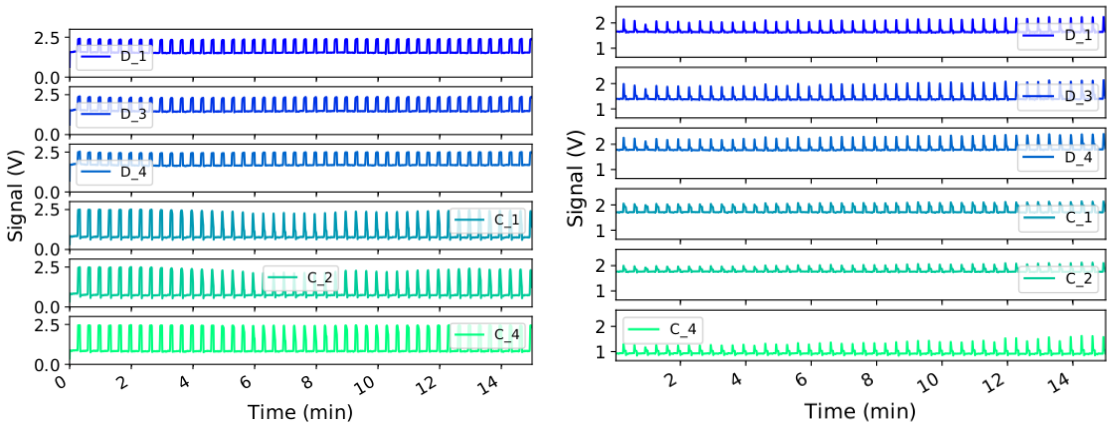
Pictures of the sensors are taken before and after the experiments with Polarized light microscopy (POM).

ANNEX III

EXPERIMENT 2

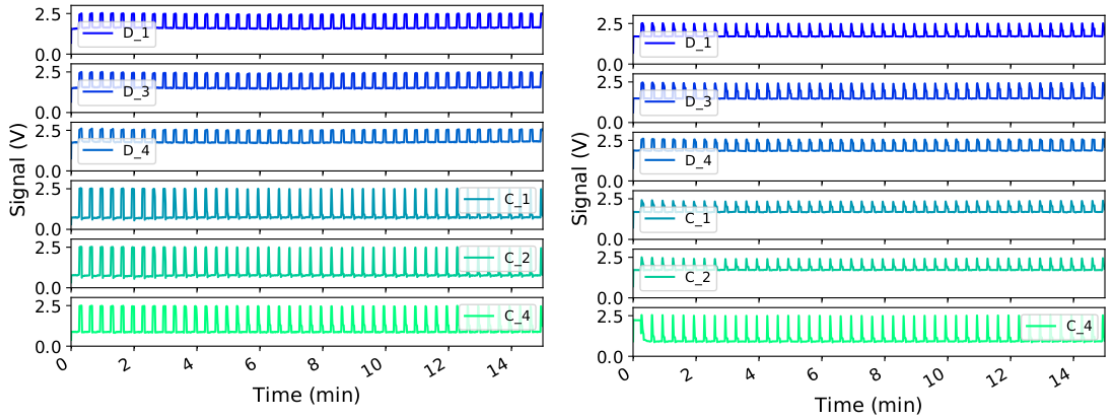
This experiment follows the established protocol (annex II) and was conducted by Doctor Susana Palma. Two different formulations of the sensors were made. Sensors nominated *D_X* correspond to the standard formulation and the ones nominated *C_X* correspond a the formulation that substitutes the IL [BMIM][DCA] with the IL [BMIM][Cl].

This batch of sensors were exposed to a sequence of 11 VOCs (15 minutes per VOC) on May 7 (nominated Experiment 2.a) and again, two months later, on July 9 (nominated Experiment 2.b). Figures III.1 to III.11, correspond to the plot of their responses. The images on the left are from Experiment 2.a and on the right are from Experiment 2.b.



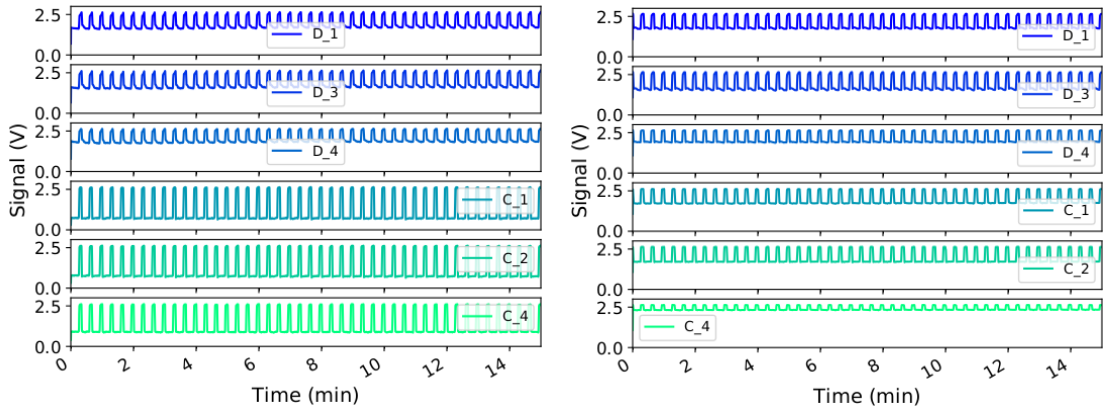
(a) (b)

Figure III.1: Responses of all sensors exposed to heptane: (a) Experiment 2.a; (b) Experiment 2.b



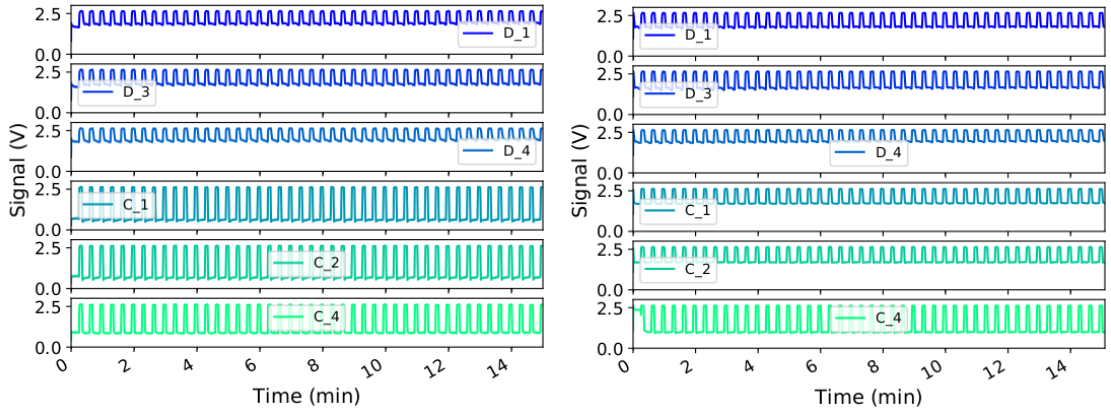
(a) (b)

Figure III.2: Responses of all sensors exposed to hexane: (a) Experiment 2.a; (b) Experiment 2.b



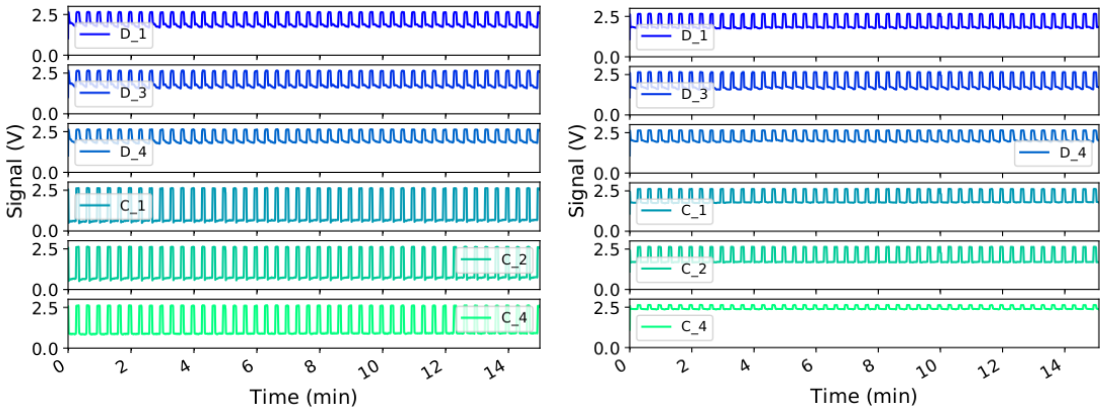
(a) (b)

Figure III.3: Responses of all sensors exposed to toluene: (a) Experiment 2.a; (b) Experiment 2.b



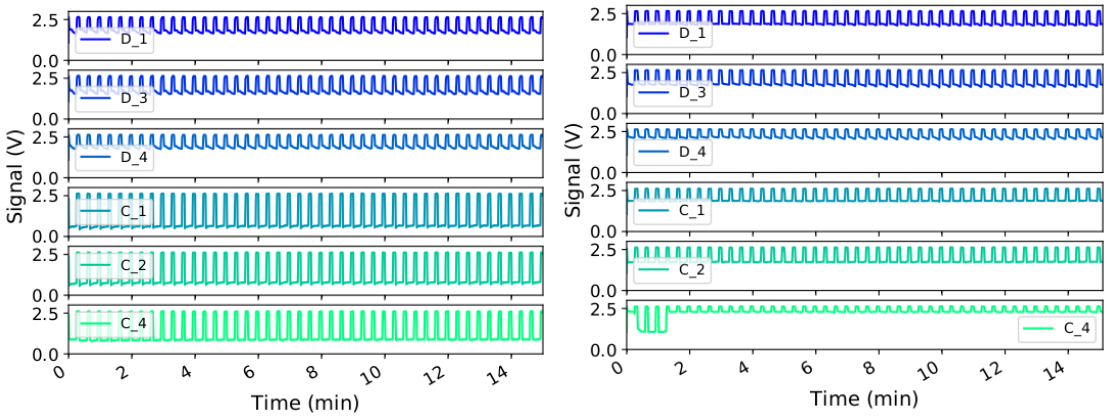
(a) (b)

Figure III.4: Responses of all sensors exposed to chloroform: (a) Experiment 2.a; (b) Experiment 2.b



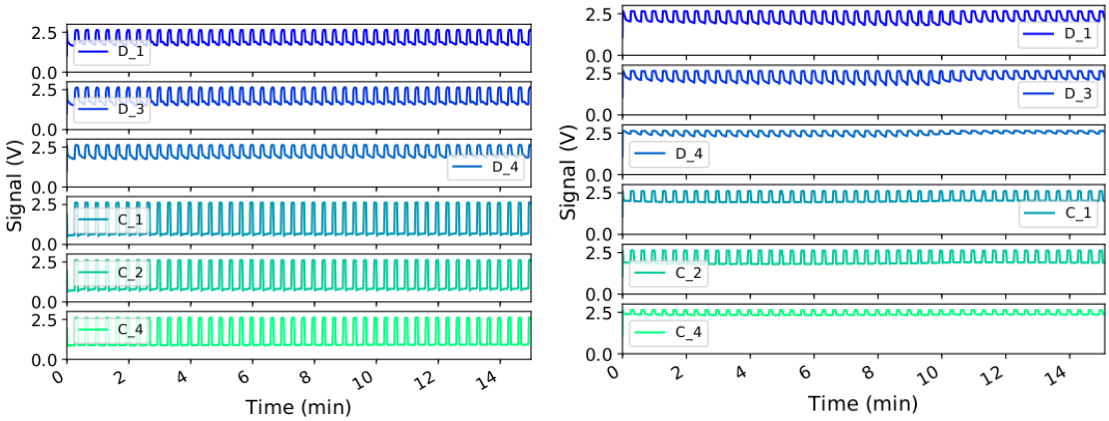
(a) (b)

Figure III.5: Responses of all sensors exposed to dichloromethane: (a) Experiment 2.a; (b) Experiment 2.b.



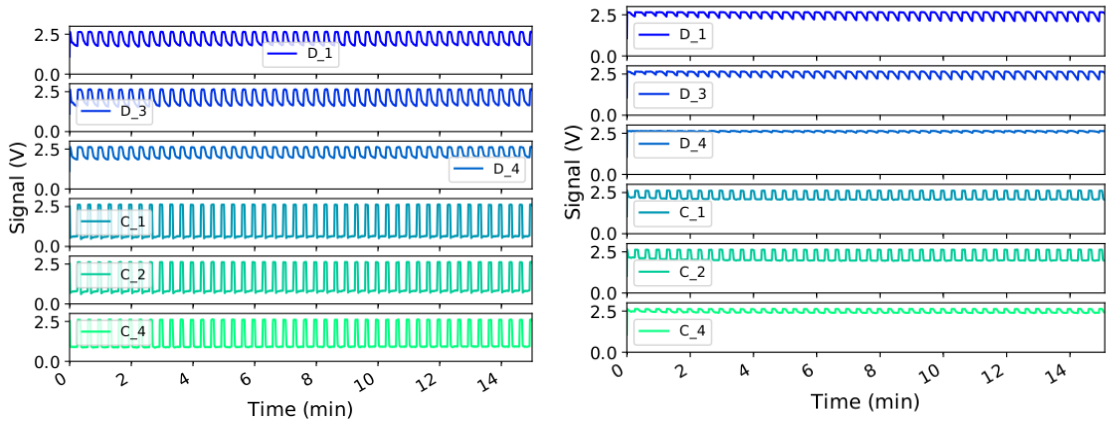
(a) (b)

Figure III.6: Responses of all sensors exposed to diethyl ether: (a) Experiment 2.a; (b) Experiment 2.b



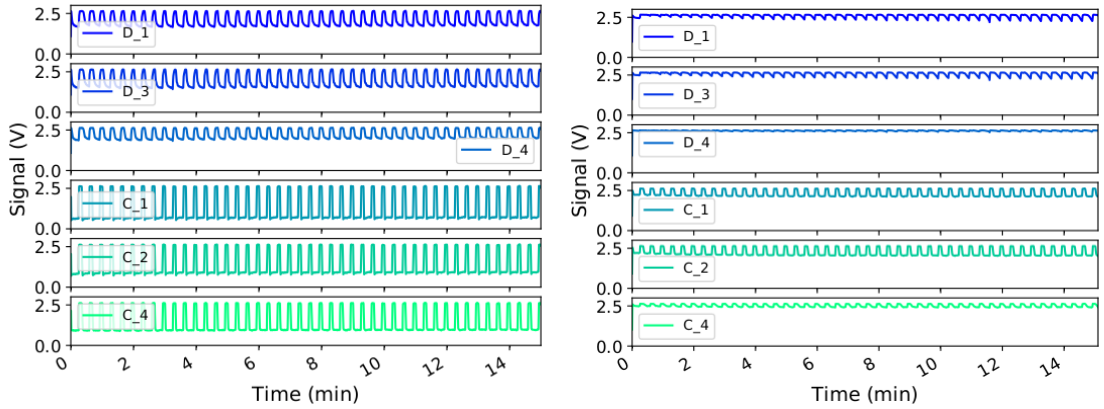
(a) (b)

Figure III.7: Responses of all sensors exposed to ethyl acetate: (a) Experiment 2.a; (b) Experiment 2.b



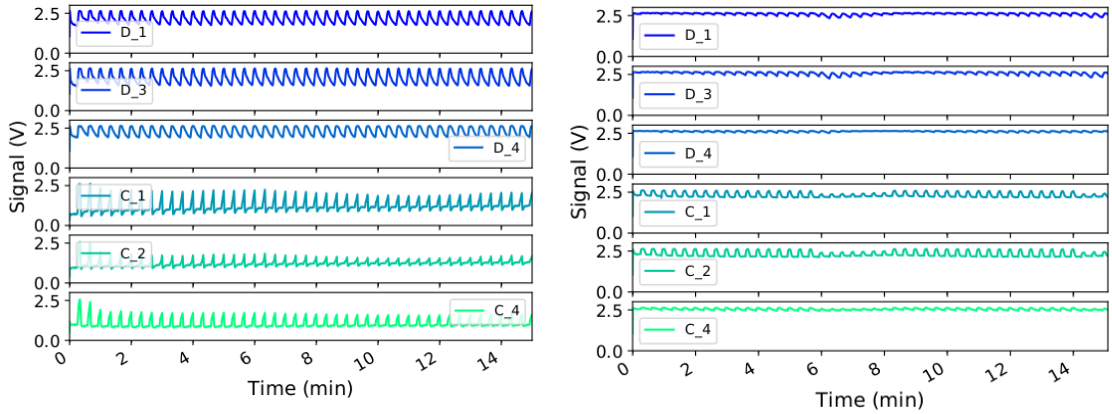
(a) (b)

Figure III.8: Responses of all sensors exposed to acetone: (a) Experiment 2.a; (b) Experiment 2.b



(a) (b)

Figure III.9: Responses of all sensors exposed to acetonitrile: (a) Experiment 2.a; (b) Experiment 2.b



(a) (b)

Figure III.10: Responses of all sensors exposed to ethanol: (a) Experiment 2.a; (b) Experiment 2.b

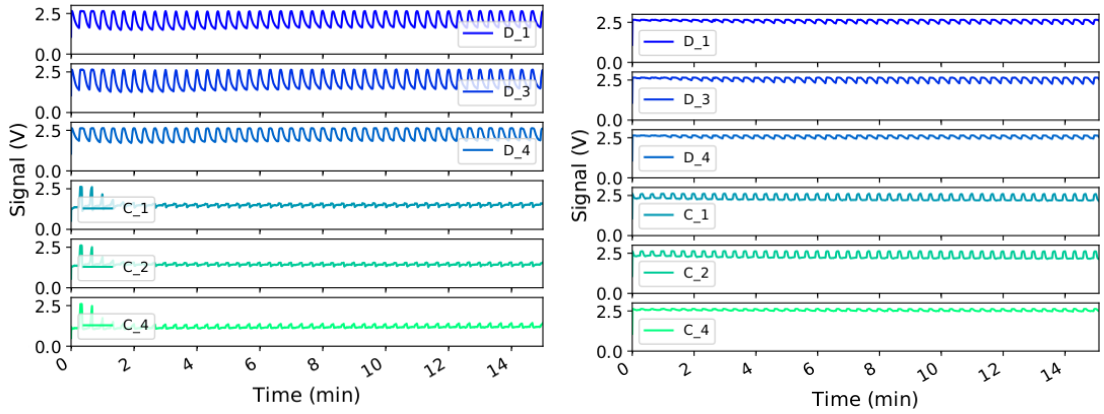


Figure III.11: Responses of all sensors exposed to methanol: (a) Experiment 2.a; (b) Experiment 2.b

ANNEX IV

Experiment 3

This experiment was conducted on April by PhD student Gonçalo Teixeira in the scope of his research. It follows the established protocol, the gas samples were used during five minutes. The sensors were made with the standard formulation. Sensors are nominated *G0X*, *X* corresponds to the unique identifier of the sensor. Only four volatiles were used: hexane, toluene, acetone and ethanol.

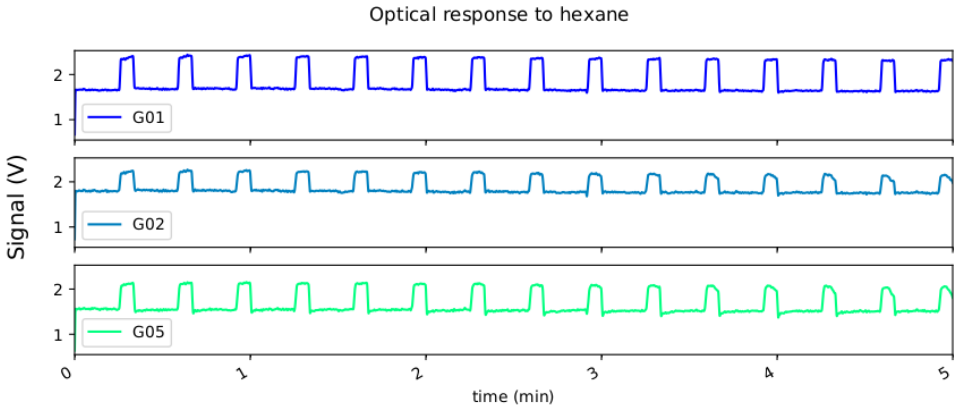


Figure IV.1: Responses of all sensors exposed to hexane: (a) Experiment 3.a; (b) Experiment 3.b

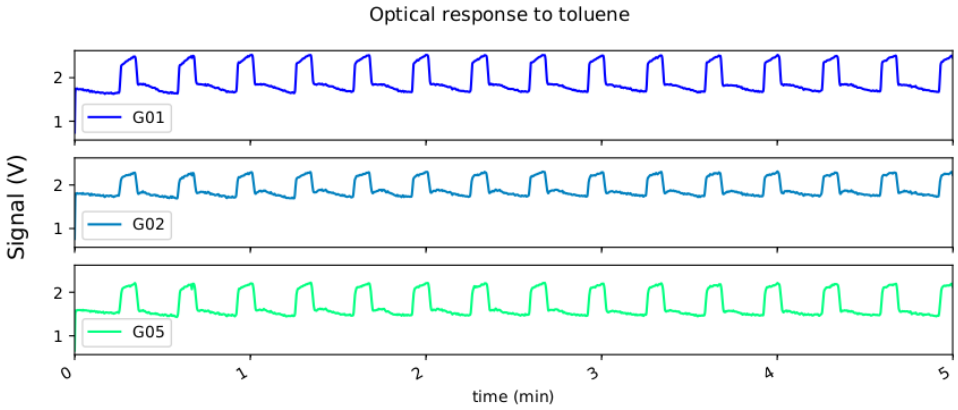


Figure IV.2: Responses of all sensors exposed to toluene: (a) Experiment 3.a; (b) Experiment 3.b

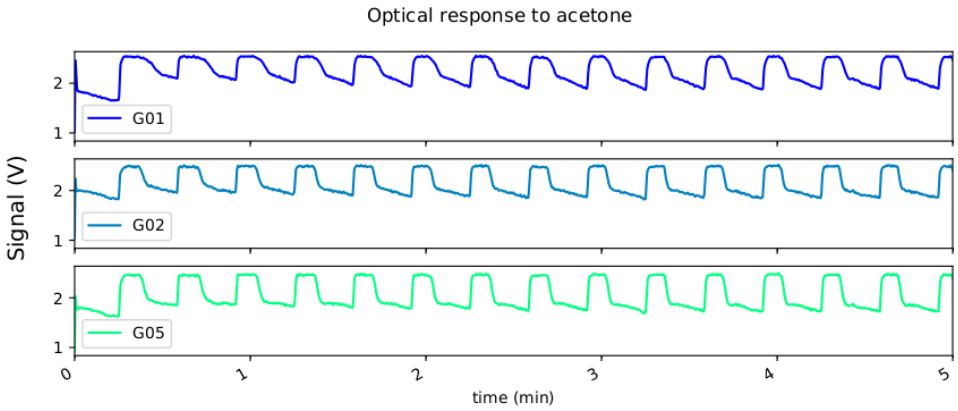


Figure IV.3: Responses of all sensors exposed to acetone: (a) Experiment 3.a; (b) Experiment 3.b

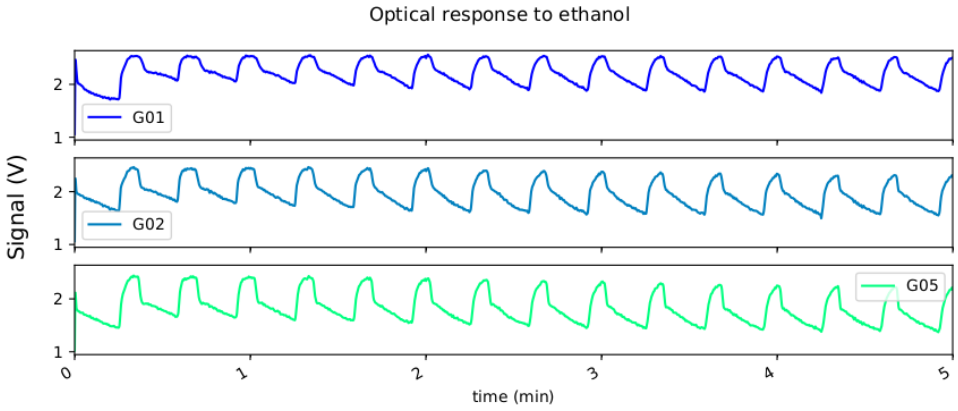
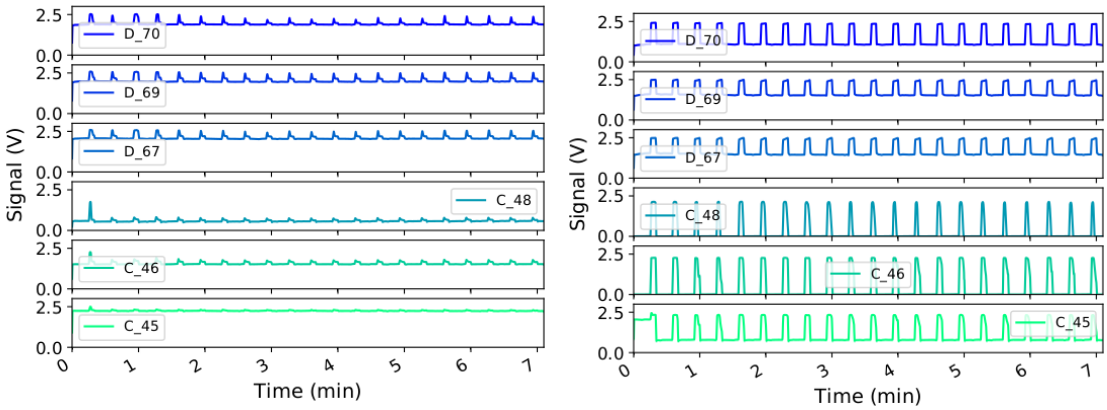


Figure IV.4: Responses of all sensors exposed to ethanol: (a) Experiment 3.a; (b) Experiment 3.b

ANNEX V

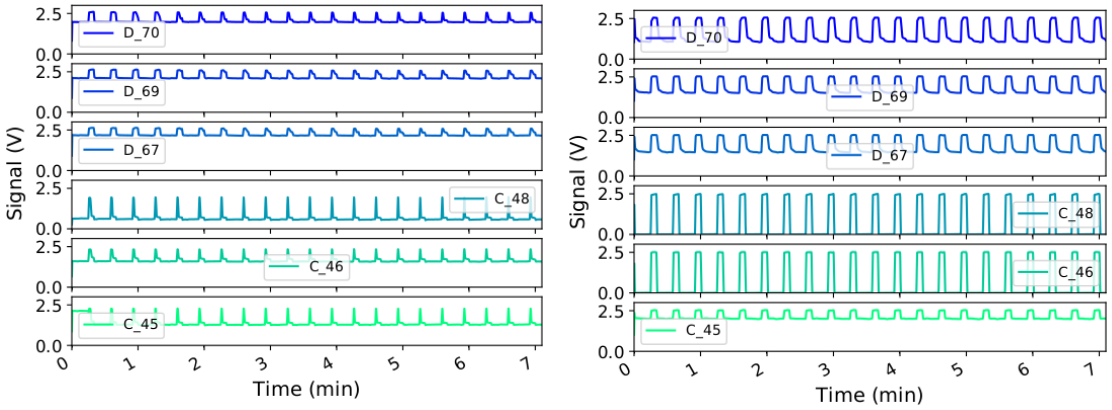
Experiment 4

This experiment was conducted by Doctor Susana Palma. It follows the similar conditions as Experiment 2 with the exception that the samples were only used during 7 minutes. It also has two parts, Experiment 4.a was performed on July 10 and Experiment 4.b was performed on July 17. Figures V.1 to V.11, correspond to the plot of the responses. The images on the left are from Experiment 4.a and on the right are from Experiment 4.b.



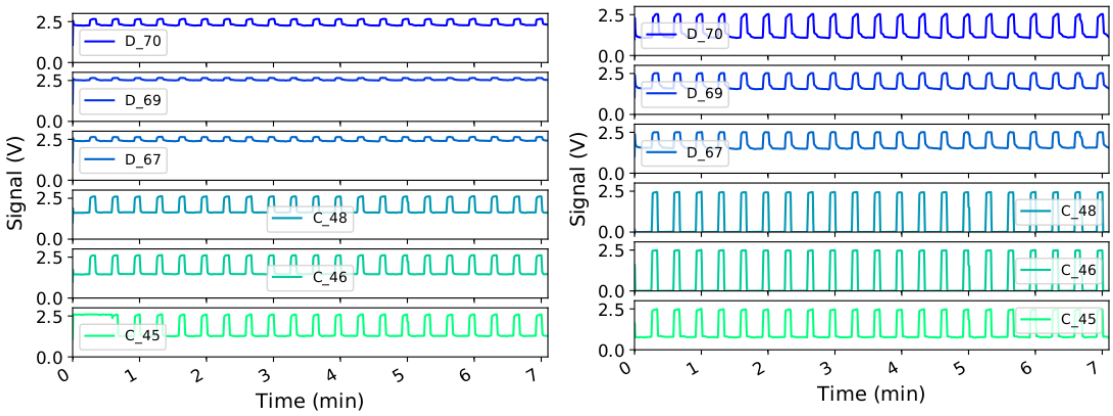
(a) (b)

Figure V.1: Responses of all sensors exposed to heptane: (a) Experiment 4.a; (b) Experiment 4.b



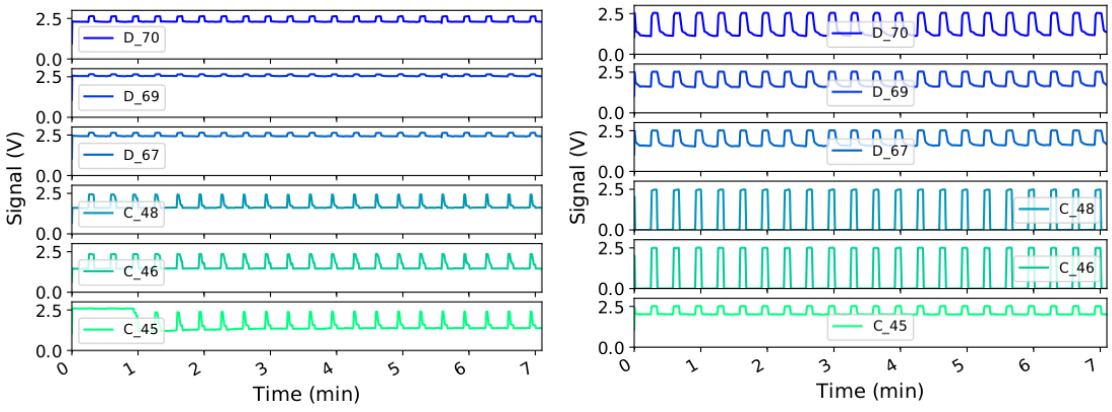
(a) (b)

Figure V.2: Responses of all sensors exposed to hexane: (a) Experiment 4.a; (b) Experiment 4.b



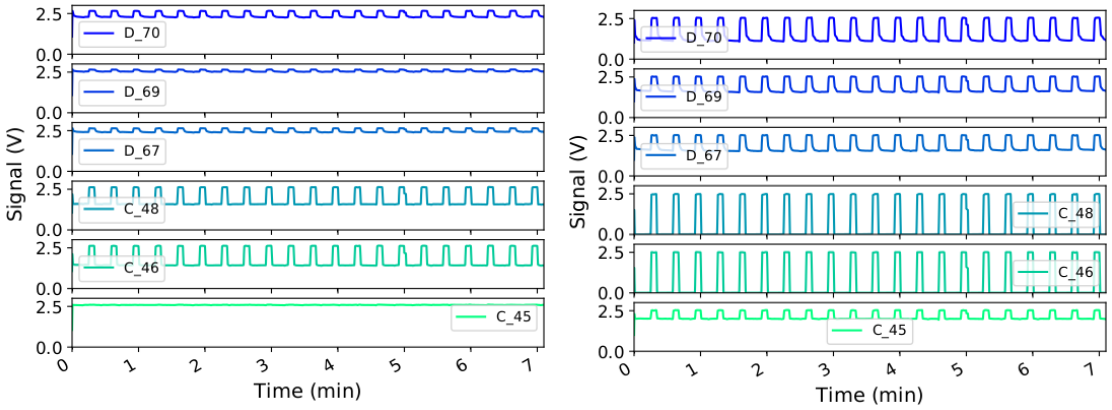
(a) (b)

Figure V.3: Responses of all sensors exposed to toluene: (a) Experiment 4.a; (b) Experiment 4.b



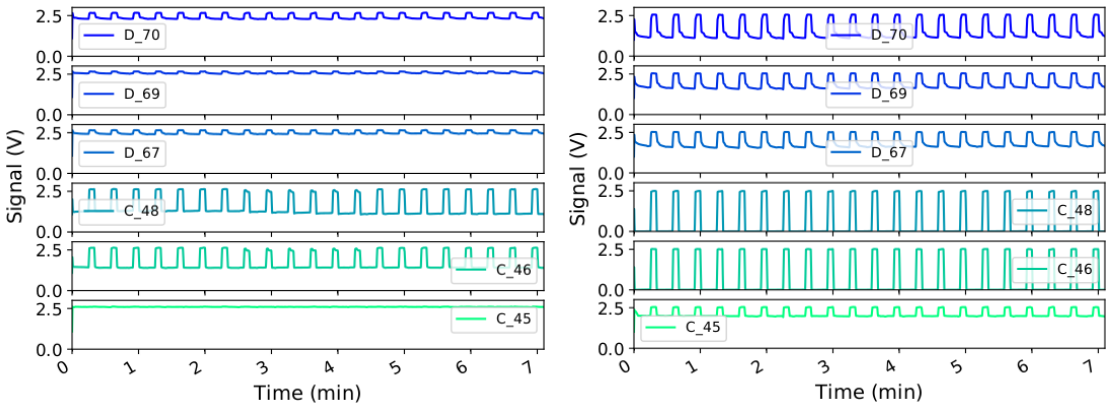
(a) (b)

Figure V.4: Responses of all sensors exposed to chloroform: (a) Experiment 4.a; (b) Experiment 4.b



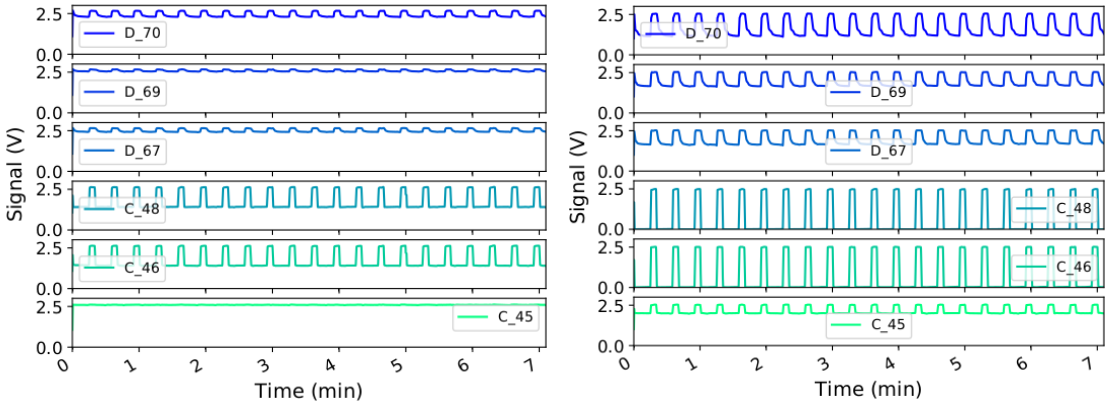
(a) (b)

Figure V.5: Responses of all sensors exposed to dichloromethane: (a) Experiment 4.a; (b) Experiment 4.b.



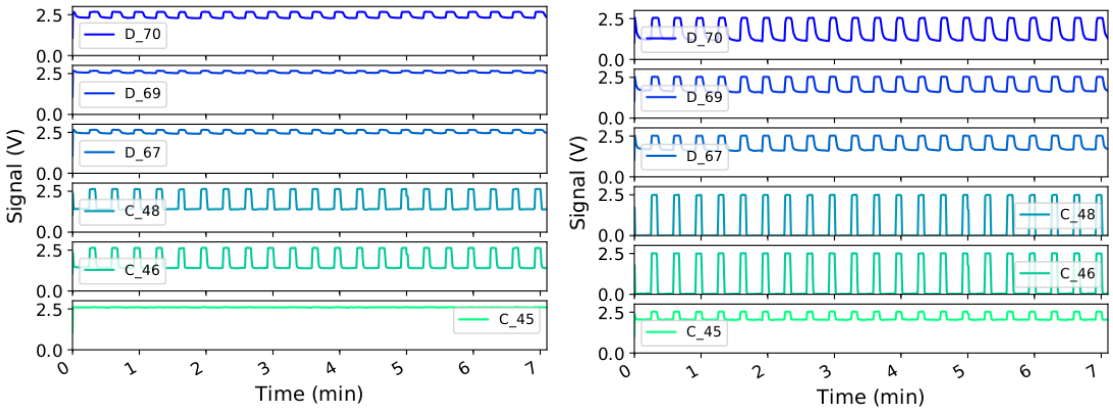
(a) (b)

Figure V.6: Responses of all sensors exposed to diethyl ether: (a) Experiment 4.a; (b) Experiment 4.b



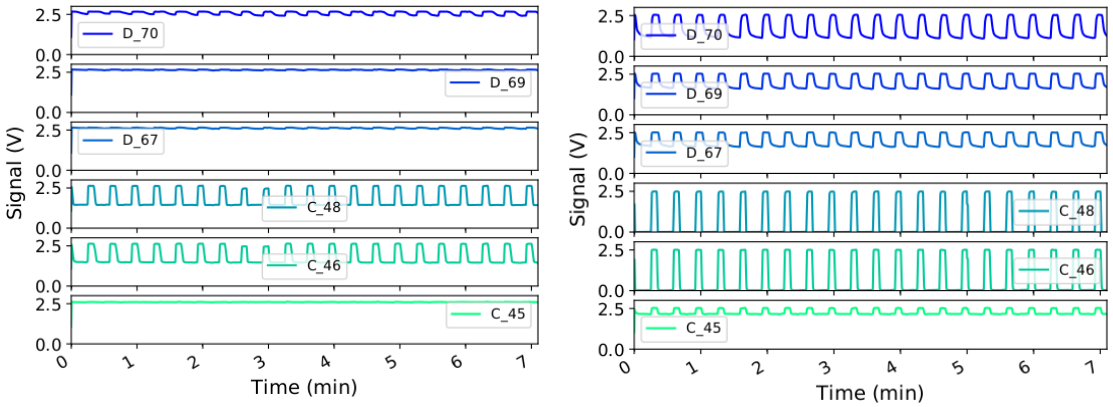
(a) (b)

Figure V.7: Responses of all sensors exposed to ethyl acetate: (a) Experiment 4.a; (b) Experiment 4.b



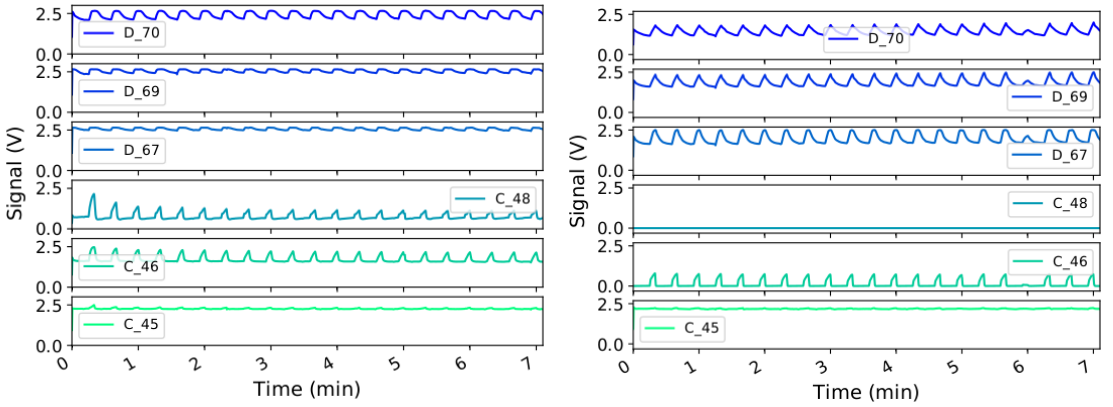
(a) (b)

Figure V.8: Responses of all sensors exposed to acetone: (a) Experiment 4.a; (b) Experiment 4.b



(a) (b)

Figure V.9: Responses of all sensors exposed to acetonitrile: (a) Experiment 4.a; (b) Experiment 4.b



(a) (b)

Figure V.10: Responses of all sensors exposed to ethanol: (a) Experiment 4.a; (b) Experiment 4.b

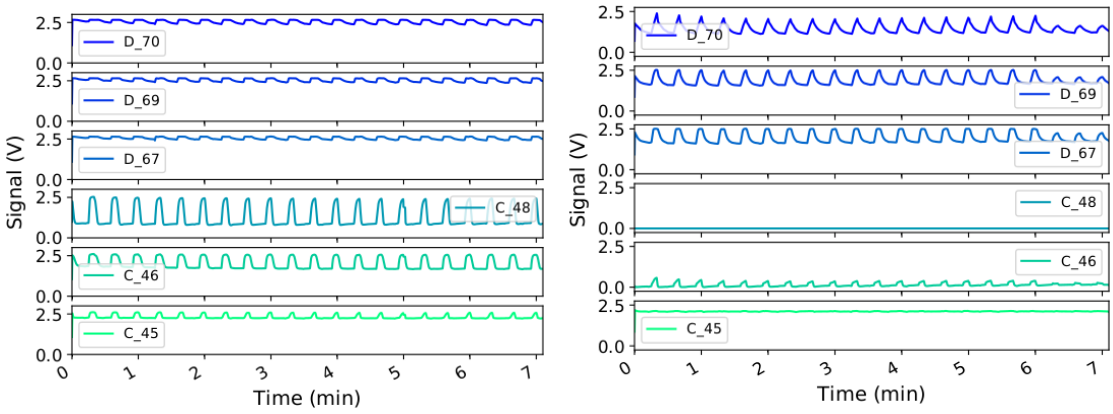


Figure V.11: Responses of all sensors exposed to methanol: (a) Experiment 4.a; (b) Experiment 4.b

ANNEX VI

EXPERIMENT 5

This experiment was conducted by Doctor Inês in the scope of her Post-Doctoral research, it follows the established protocol, with the difference that the gelatin was substituted with starch. This batch of sensors is nominated with *SX*, in which *X* corresponds to the unique identifier of the sensors. Three sensors where were exposed to a sequence of 11 VOCs on June 12 (nominated Experiment 5.a) and again, two months later, on June 15 (nominated Experiment 5.b). Figures VI.1 to VI.11, correspond to the plot of the responses.

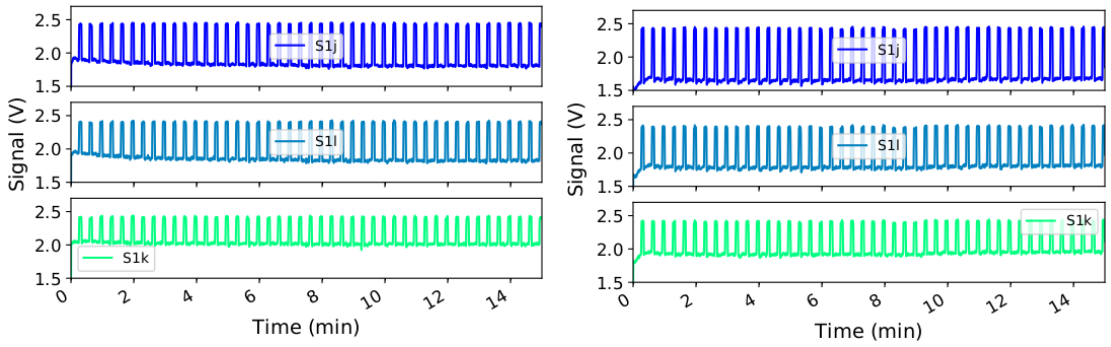
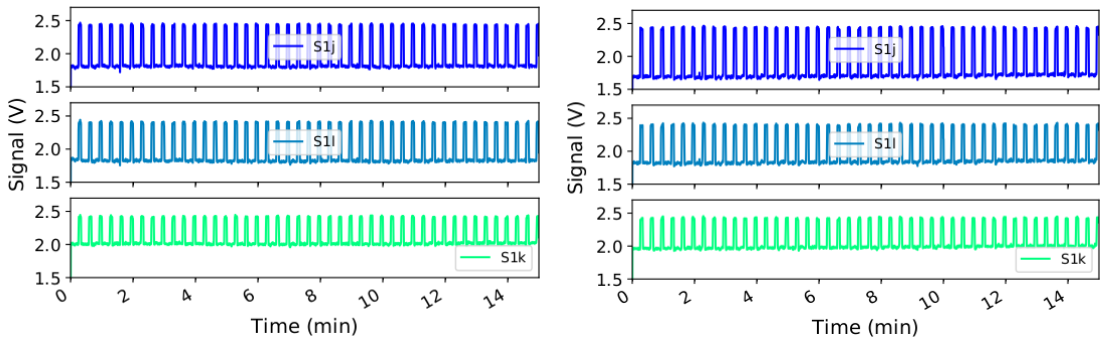
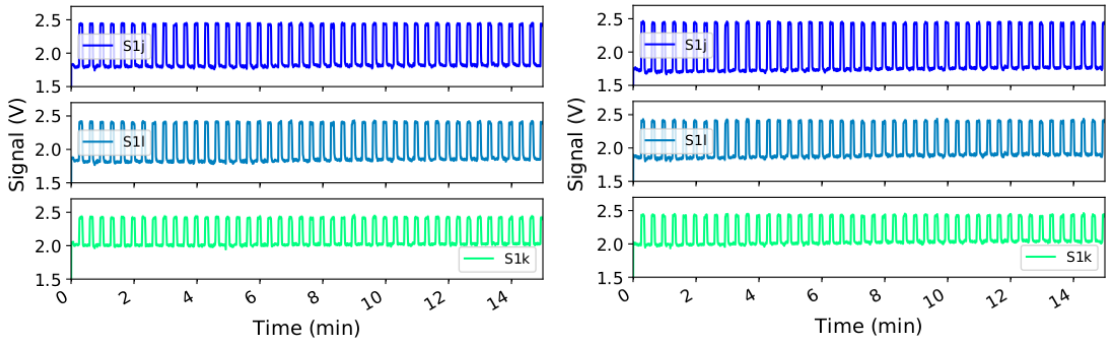


Figure VI.1: Responses of all sensors exposed to heptane: (a) Experiment 5.a; (b) Experiment 5.b



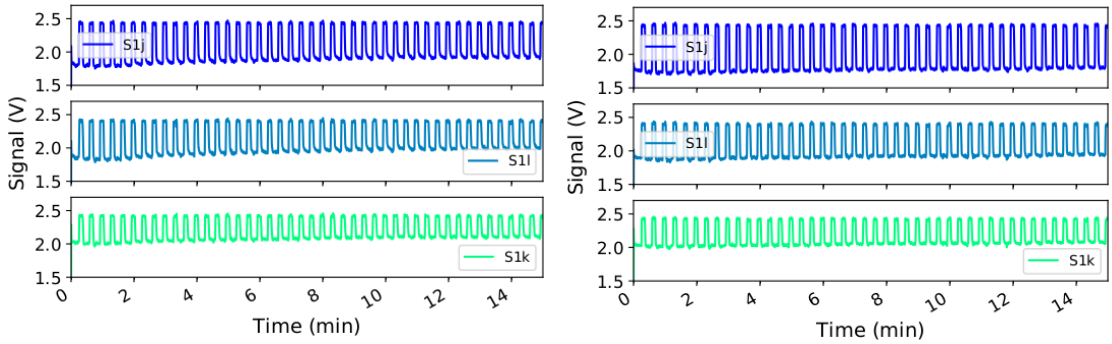
(a) (b)

Figure VI.2: Responses of all sensors exposed to hexane: (a) Experiment 5.a; (b) Experiment 5.b



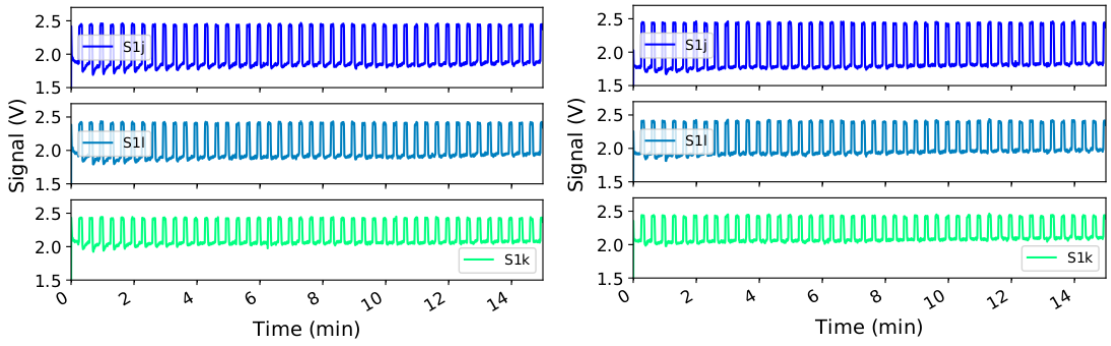
(a) (b)

Figure VI.3: Responses of all sensors exposed to toluene: (a) Experiment 5.a; (b) Experiment 5.b



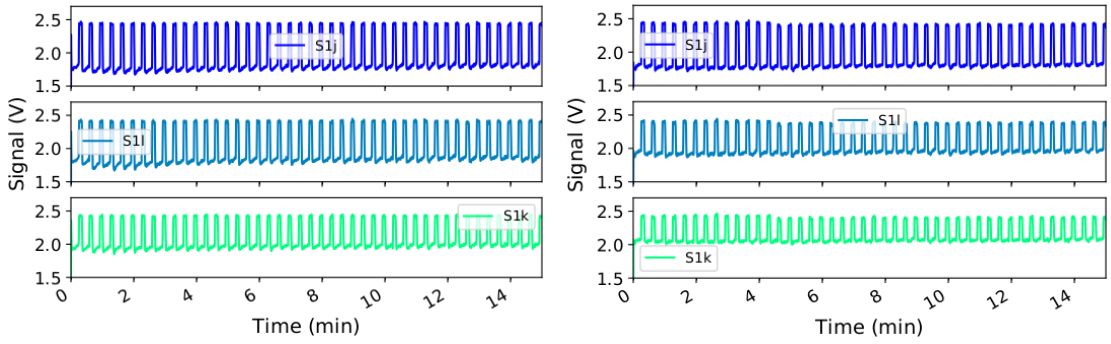
(a) (b)

Figure VI.4: Responses of all sensors exposed to chloroform: (a) Experiment 5.a; (b) Experiment 5.b



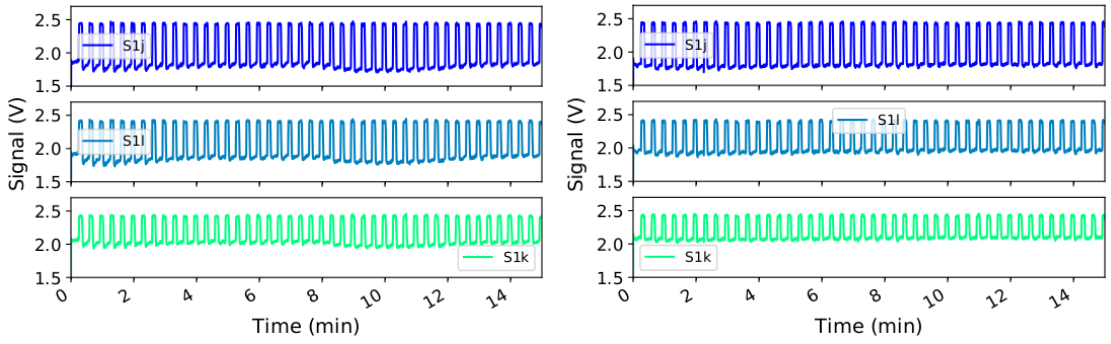
(a) (b)

Figure VI.5: Responses of all sensors exposed to dichloromethane: (a) Experiment 5.a; (b) Experiment 5.b.



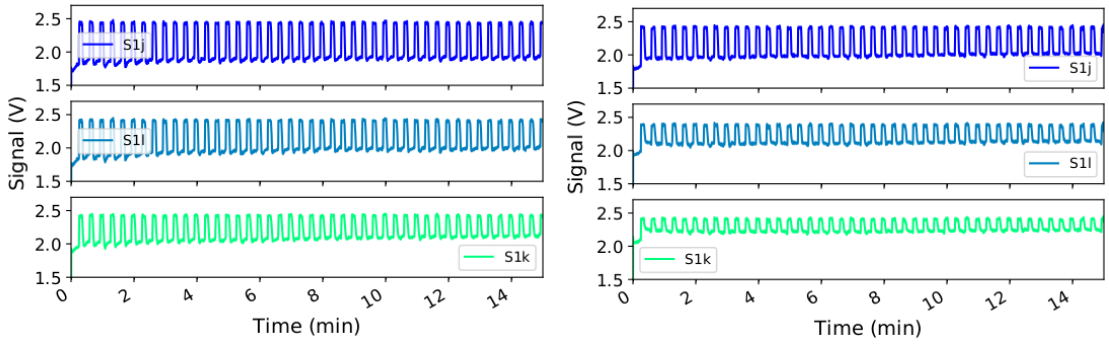
(a) (b)

Figure VI.6: Responses of all sensors exposed to diethyl ether: (a) Experiment 5.a; (b) Experiment 5.b



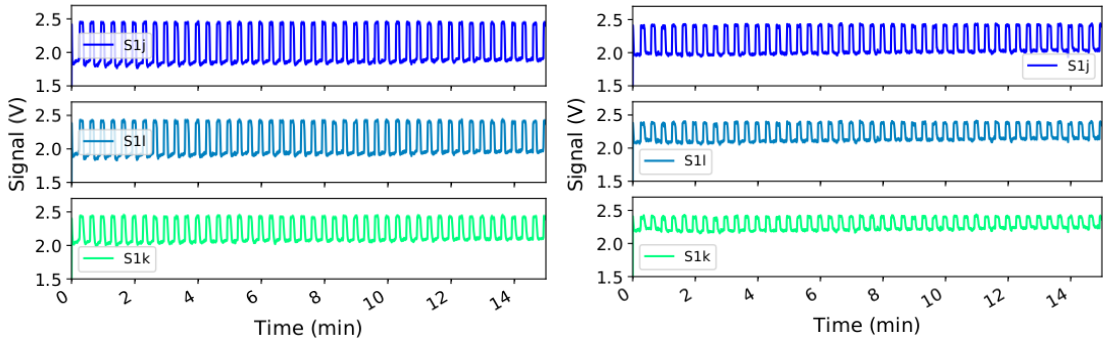
(a) (b)

Figure VI.7: Responses of all sensors exposed to ethyl acetate: (a) Experiment 5.a; (b) Experiment 5.b



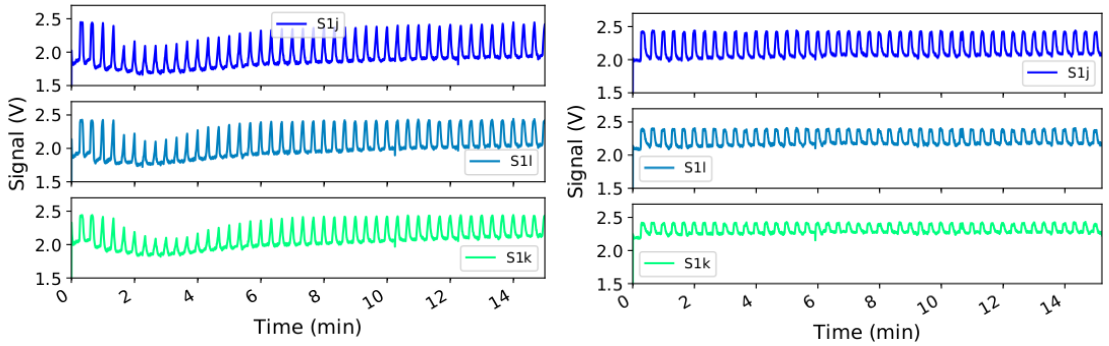
(a) (b)

Figure VI.8: Responses of all sensors exposed to acetone: (a) Experiment 5.a; (b) Experiment 5.b



(a) (b)

Figure VI.9: Responses of all sensors exposed to acetonitrile: (a) Experiment 5.a; (b) Experiment 5.b



(a) (b)

Figure VI.10: Responses of all sensors exposed to ethanol: (a) Experiment 5.a; (b) Experiment 5.b

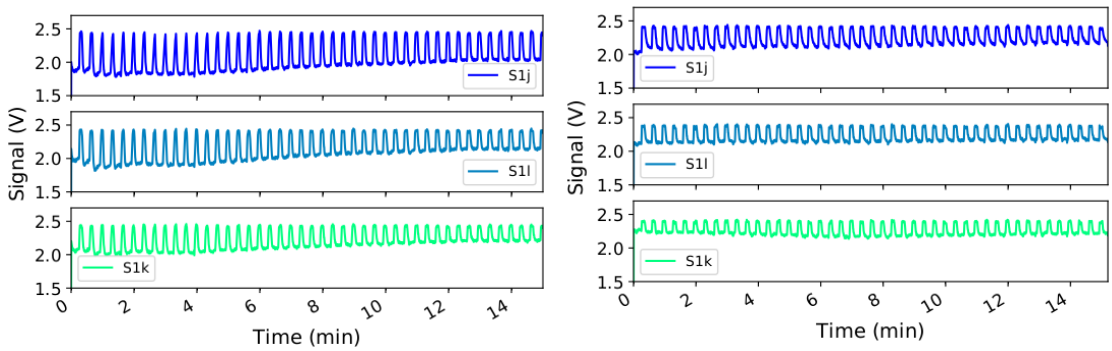


Figure VI.11: Responses of all sensors exposed to methanol: (a) Experiment 5.a; (b) Experiment 5.b