FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Fernando Manuel Ribeiro Pereira

Classification of daily life odours with an electronic nose

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Lino Marques (DEEC-ISR)

Coimbra, 2017

This work was developped in collaboration with:

Instituto de Sistemas e Robótica

Acknowledgments

First of all, I would like to express my sincere gratitude and admiration for my advisor Dr. Lino Marques, for the continuous support during the development of this thesis as well as his patience, suggestions and his overall vast knowledge. I would like to thank as well my ISR laboratory colleagues, namely Jonatas, João and Sedat for their readily help and support in this endeavour. To all my friends that supported me and were always open for a chat my wholehearted appreciation. Lastly, but perhaps the most important, to my father and sister that fully supported me since my inception and to their acceptance and comprehension during a problematic phase of my life.

Resumo

Com o potencial desenvolvimento de um dispositivo portátil capaz e robusto o suficiente para detetar odores variados em mente e para ajudar aqueles que não são capazes de cheirar, esta tese focou-se no desenvolvimento de um pequeno nariz eletrónico composto por um conjunto de quatro sensores diferentes com alvos e sensibilidades heterogéneas. O nariz eletrónico construído contém um grupo de quatro sensores de gás dentro de uma câmara de gás. O ar é extraído de dentro da câmara através de uma bomba de ar. As leituras dos sensores passam por um conversor analógico-digital e seguem para um computador de placa única onde são recebidos e armazenados. A metodologia de amostragem é controlada pela placa referida. Uma alargada coleção de odores do dia-a-dia foi selecionada para expor esta tecnologia a um complexo conjunto de odores bem como para desafiar as capacidades discriminativas do nariz eletrónico desenvolvido. Duas técnicas diferentes de escalamento de dados foram aplicadas à informação recolhida: normalização de 0 a 1 e normalização logarítmica. Redução de dimensionalidade foi aplicada pela PCA(Análise de Componentes Principais) e LDA(Análise de Discriminantes Lineares). Das vinte e cinco amostras para cada odor adquiridas, a metodologia de validação cruzada de 75-25% foi aplicada. Vinte amostras foram escolhidas aleatoriamente para servirem como grupo de treino e cinco como grupo de teste. A variância cumulativa das cinco primeiras componentes do PCA é 96.50% para dados não processados, 98.20% para dados escalados de 0 a 1 e quase 100.00% para dados normalizados logaritmicamente. Devido a processos internos do algoritmo, os primeiros cinco discriminantes do LDA atingem 98.30% da informação do grupo de dados, independentemente do estado dos dados. Rede Neuronal de Retropropagaçao(BPNN e Máquina de suporte de vetores(SVM) são os classificadores utilizados nesta tese.) Dos resultados do SVM, foi obtida uma exatidão de 92.50% com dados não processados e 75.00% com dados escalados de 0 a 1. Dos components do PCA, uma exatidão de 85.00% nos dados não processados foi o melhor resultado. Dados escalados logaritmicamente conseguiram uma exatidão de 34.90% para ambos os estados de projeção.

Os discriminantes LDA alcançaram 92.50% de exatidão. Considerando agora os resultados de exatidão da BPNN, classificação com dados não processados obteve 53.12% e 82.50% com dados normalizados de 0 a 1. Os componentes PCA dos dados não processados conseguiram uma exatidão de 82.50%, 67.50% com dados escalados de 0 a 1 e 80.00% com dados escalados logaritmicamente. Discriminantes da LDA conseguiram 97.50% de exatidão.

Abstract

With the potential development of a portable device capable and robust enough to detect different odours in mind and to help those who are not able to smell, this thesis focused on the development of a small electronic nose comprised of an array of four different sensors with dissimilar targets and sensitivities. The eletronic nose assembled contemplates an array of four gas sensors inside a gas chamber. Air is extracted from inside with an air pump. Sensor readings pass through and analog-to-digital converter and follow towards a single-board computer where they are collected and stored. Sampling methodology is controlled by the latter. A broad collection of daily life odours was selected to expose the technology to a quite complex set of odours as well as to challenge the discriminative capabilities of the electronic nose developed. Two different data scaling techniques were applied to the collected data: 0 to 1 normalization and logarithmic normalization. Dimensionality reduction was enabled by both PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis). From twenty-five samples from each odour collected, a 75-25% cross-validation methodology was applied. Twenty samples were picked randomly to serve as the training set and five as the testing set. PCA cumulative variance of the first five components is 96.50% for raw data, 98.20% for data scaled from 0 to 1 and almost 100.00% for logarithmic normalized data. Due to intrinsics of the algorithm, LDA first five discriminants amount to 99.80% of the dataset information, regardless of data status. Backpropagation Neural Network(BPNN) and Support Vector Machines(SVM) are the classifiers applied in this thesis. From the SVM results, an accuracy of 92.50% was obtained with raw data and 75.00% with data scaled 0 to 1. From PCA components, 85.00% accuracy on raw data was the best result. Logarithmic scaled data accomplished an accuracy of 34.90% for both data projection status. LDA discriminants achieved 92.50% of accuracy. Regarding BPNN accuracy results, classification with raw data achieved 40.00% and 82.50% with normalized data from 0 to 1. PCA components from raw data had an accuracy of 82.50%, 67.50% from data scaled 0 to 1 and 80.00% from logarithmic scaled data.

Discriminants from LDA reached 97.50% of accuracy.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The olfactory system is one of the main functions of the brain. This sensory system was and still is essential to mammals along their evolution and survival. With the odoriferous information gathered in our environment cooperatively with other senses we are able to react emotionally in accordance towards our olfactory memory. Considering the human nose is able to detect around 400000 volatile compounds with different chemical structures may they be monomolecular odours or multimolecular ones [1], the task to identify what is the logic behind the neuronal pathways used to be able to perceive a certain behaviour becomes highly complex. Recent progress in molecular biology showed that there are around 400 separate types of olfactory receptors genes in the olfactory human system [2] . It's highly believed that the connection between odour molecules and the correspondent receptors does not follow the lock-key model, considering that those just recognize partial structures of the odours they connect to.

1.1 Motivation

1.1.1 Portrayal of odours visually

In our daily life we experience a large amount of different smells, volatile organic compounds emanated from organic entities around us and due to the olfactory sense we are able to detect them. Usually odours spread around in the environment as a plume like form that is subject to air and wind actions. This constant movement gives them a really complex structure to analyse because the concentration of compounds in the air mixture is constantly changing. Spatiotemporal detection

of this information is a real necessity for both animals and humans alike in their environment contextualization. Different types of hints about food or eminent dangers is information too valuable to be ignored by the other senses. In order to face these challenges, the scientific community has been developing different types of chemical sensors to be able to detect and even visualize the odoriferous space. Despite their weak spatiotemporal resolution, those sensors can be incorporated in both immobile and mobile robots so that they can, to a certain extent, behave like an animal searching for food, detecting potential mating partners and predators. Considering this technological asset, those smart robots may be deployed to reveal gas leaks, dangerous chemicals and diverse types of pollutant sources. They may be also used in a healthcare context, namely in the detection of chemical compounds that are preponderant indicators of a certain diagnosis. Unfortunately, in contrast to audition and vision, olfactory information communication is particularly hard considering there is not a standardized way of interpreting the same odours and the individual memory in the odour caption is a factor that differentiates how the olfactory stimulus is processed by the brain. Therefore, if we able to develop ways of depict visually the odoriferous space, the most intuitive procedure to transmit information about the olfactory perception is through its visual portrayal.

1.1.2 Impact of odours in daily routine

According to recent questionnaires measuring changes related to olfactory loss, there are a large different number of daily life problems reported by participants in those studies. The areas of daily life included on those reports range from the decreased enjoyment of food, greater risk of failure to perceive a gas, unawareness of spoiled food and burnt smells, difficulties achieving what is asked for them on their professional lives as well as lacking awareness of their personal hygiene. In Figure 1.1 eight different studies are condensed. [3–10]. Suffering from a type of olfaction malfunction disables a sensitive layer of experience of the daily life so this together with age might lead to depressive symptoms and a poor quality of life [11] .

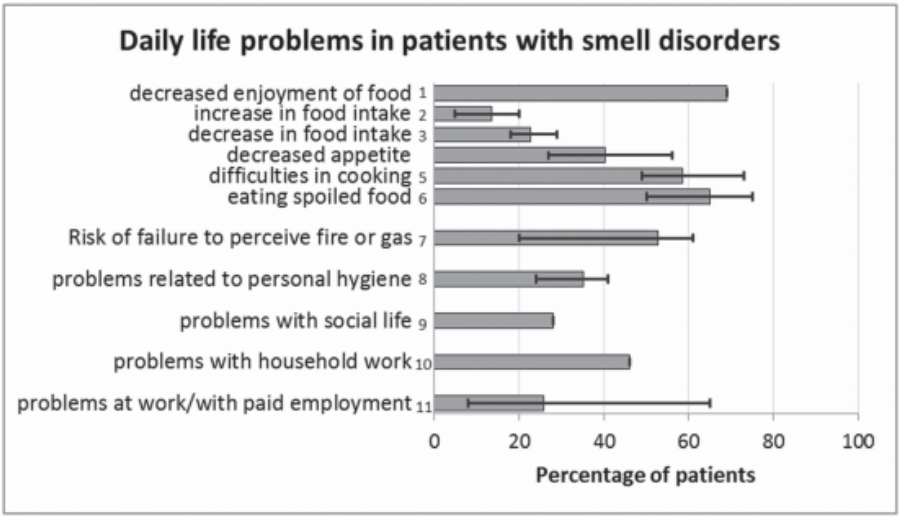


Figure 1.1: Daily life problems of patients with smell disorders from eight different studies referenced above. The error bar represents the lowest and highest reported percentage [12].

1.2 *Sniff* Rhythm

Olfactory perception mechanisms start with a discrete respiration, often called *sniff*. A sniff is defined as a reflex provoked by chemicals – either irritants or odours. In Figure 1.2, a sniff response in the olfactory receptor neurons of a rat is displayed. According to recent behavioural studies in rats, each sniff might contain enough information to provide a complete snapshot of the local olfactory environment [13]. After a sniff, odorants captured are drawn, together with air, into the nasal cavity promoting the activation of olfactory sensory neurons in the nose.

1.3 Olfaction Disorders

A significant fraction of the populations experiences some sort of olfactory malfunction, although this is many times neglected. There are two different classifications that can be applied to olfactory disorders, quantitative and qualitative. From the quantitative group, the disorder that stands out is anosmia, which can be a total ineptitude of detecting smells and hyposmia, a partial unawareness specific to certain odours. This lack of ability to smell is due to a physiological phenomenon that can be caused by a variety of reasons described in the next section. In the qualitative side, parosmia is a olfactory dysfunction that is characterized by identifying odours incorrectly often describing them as unpleasant aromas. Phantosmia, which is also

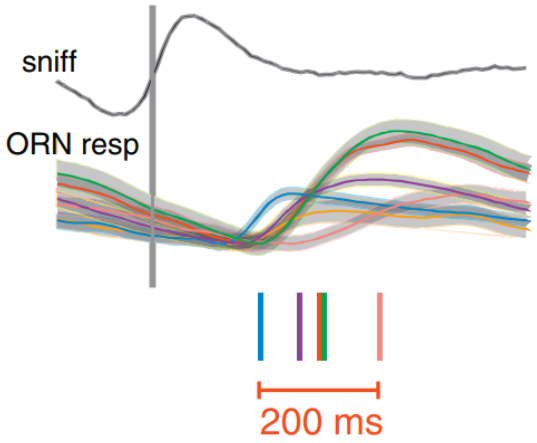


Figure 1.2: Olfactory receptor neurons inputs responses promoted by an inhalation of an odorant in five different glomeruli, imaged in an awake rat. Traces are "sniff-triggered averages" of presynaptic calcium signals (shaded areas indicate variance around the mean response). Vertical lines below trace indicate the time stamp in which the responses reached half the maximum in respect towards their respective glomeruli [14].

considered a type of parosmia, is an olfactory hallucination that consists of smelling an odour when there is none present at all [15].

1.3.1 Impact

Population based studies pinpoint an olfactory loss of 22% in individuals with ages between 25 and 75 years [16]. Lee et. al [17] reported a prevalence of olfactory dysfunction of 4.5%. In other study, there was found that this percentage achieves 24.8% in individuals with ages between 60 and 90 years old, free of neurodegenerative disease [18]. Rawal et al. [19] outlined a self-reported smell alteration of 23% for adult aged 40+ years. Lafreniere et. al [20] estimated prevalence rates of olfactory disorders of more than 50% for people aged above 80 years old. Unawareness of olfactory loss is one major problem that researchers face analysing population based studies and unfortunately, it is quite common [21–23]. One reason that might explain this issue is the fact that olfactory stimuli and interactions are processed not in a conscious way but more like an experience after-effect. Considering this there is no surprise in the fact that the self-reported smell deficiency oscillates between 1,4% and 15% [24].

There are several potential causes that may lead to smell loss. The most impactful one is post viral upper respiratory infection that is responsible for 18-45% of the

clinical population. Nasal/sinus diseases, where conditions like rhinitis [25], chronic rhinosinusitis with polyps [26] and sinonasal tumors are included albeit the latter low presence [27], comes second with percentages between 7 and 56%, followed by head trauma (8-20%). Exposure to toxins/drugs (2-6%) and congenital anosmia (0-4%) are the last less prevalent known smell loss causes [28]. It is also worth noting that, olfactory disorders caused by neurodegenerative diseases like Alzheimer's are considered one of the first clinical symptoms [29] and Hori et. al [30] found that Alzheimer's disease patients score significantly lower on odour tests than an age-matched group with healthy individuals.

Considering smell loss among different medical conditions, the percentage of patients with clinically proven smell loss is rather high: 76-95 % in post viral upper respiratory infection, 72-98 % in nasal/sinus disease, 86-94 % in head trauma, 67 % in exposure to toxins/drugs and 100 % in congenital cases [28]. Loss due to post viral upper respiratory infection, head trauma and exposure to toxins/drugs is to some degree reversible [31], whereas many cases of nasal/sinus disease can be treated with medication or with a combination of conservative and surgical treatments [32]. Congenital anosmia prevalence is estimated at 1:5000 - 10000 [33]. Patients that carry this disorder have their final diagnosis approximately 13 years after the first time they noticed it, at about 10 years of age according to Bojanowski et al. [34].

Although quantitative disorders are the most widespread, qualitative ones are still quite significant. In the general population the prevalence of phantosmia(odour perceptions in an absence of odour) [35] is estimated between 0.8 % and 2.1 % [36] and parosmia(distorted odour perception of an odour) is estimated to be around 4% [37]. Among patients with olfactory disorders, parosmia oscillates between 10-60 % [38–41].

1.4 Objectives

The main objective of this thesis is to classify and represent visually a series of odours that might be useful to a individual that is not capable of detecting them due to some dysfunction in its olfactory system or to detect odours that have a dangerous presence in concentrations that a human cannot detect naturally.

If possible, this solution should be integrated in a smartphone application or connected to a monitor to guarantee its coverage everywhere as long as the device is available.

1.5 Electronic Nose

An electronic nose is an electronic instrument that is able to detect certain chemical characteristics of gases around it and is able to discriminate those gases with an appropriate pattern recognition system. The properties of those characteristics depend on which sensors are used in its conception.

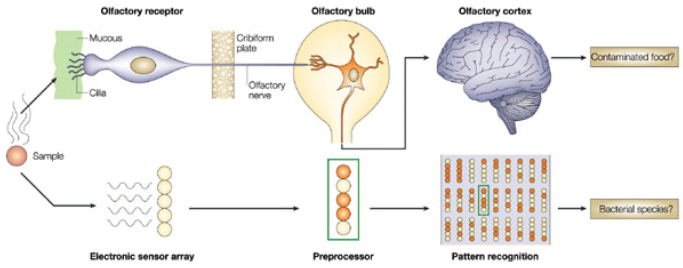


Figure 1.3: Electronic nose mimicking the human olfactory system. (Adapted from: [42])

1.5.1 Metal Oxide sensors

The principle of operation of metal oxide sensors is based on the change in conductance of the oxide in its composition on interaction with a gas and the change is usually proportional to its concentration. When in the presence of a gas, there is a vast amount of metal oxides that change their conductivity. What differentiates them is their diversified electronic structures. Pre-transition metal oxides are usually inert because they have large electronic band gaps. The difficulty they present in promoting electrical conductivity makes them one of the gas sensor materials often used. Transition metal oxides have an unstable structure and non-optimality of important parameters for conductometric gas sensors because their electronic band gaps are really small [43] so they are not applied in gas sensing systems very often [44] . Only two transition-metal oxide electronic configurations allows them to be a realistic gas sensor operating material. One of them is the d^0 configuration that contemplates metal oxides such as TiO_2 and WO_3. The other one is the d^{10} configuration that can be found in oxides such as ZnO and SnO_2 [45].

There are two types of metal oxide sensors; n-type (such as zinc oxide, tin dioxide or titanium dioxide) which response to reducing gases and p-type(nickel oxide, cobalt oxide) which respond to oxidising gases.

The n-type sensor operates as follows: oxygen in the air reacts with the surface of the

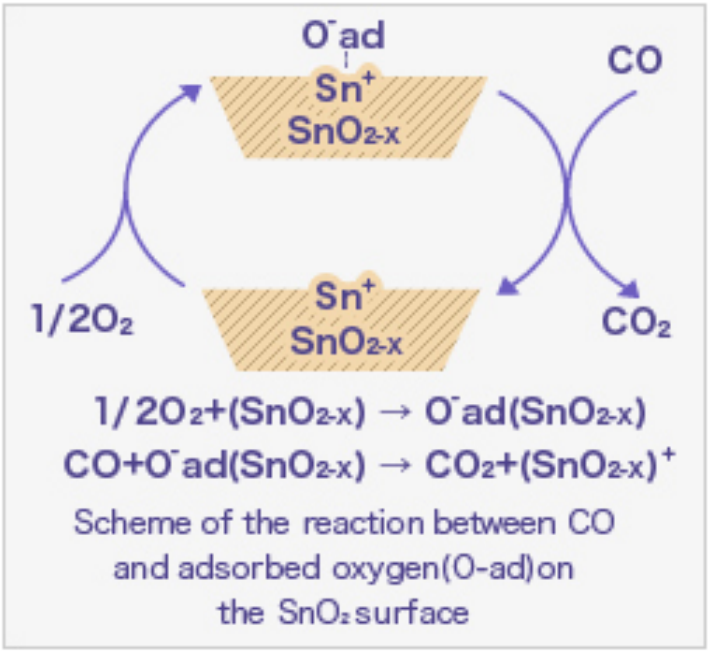


Figure 1.4: Scheme of the reaction of the surface of a SnO_2 sensor and a reducing gas species (CO in this case). (Adapted from: [46])

sensor and traps any free electrons on the surface or at the grain boundaries of the oxide grains. This produces large resistance in these areas due to the lack of carriers and the resulting potential barriers produced between the grains inhibit carrier's mobility preventing electric current flow. However, if the sensor is introduced to a reducing gas such as H_2, CH_4, C_2H_5 or H_2S the resistance drops because the gas reacts with the oxygen and releases an electron. This lowers the potential barrier and allows the electrons to flow, thereby increasing the conductivity [47]. This flow of electrons promotes the sensor resistance decrease. Gas concentration around the sensor can be inferred by measuring its resistance change, which drop is similar to a linear decrease in log space with the increase of concentration of a reducing gas [48]. An approximation of this process is shown in Equation 1.1.

$$\frac{R_s}{R_0} = KC^{\alpha} \tag{1.1}$$

Where:

- R_s represents the sensor resistance
- R_0 is the air resistance in clean air
- C is the concentration of the reducing gas

- α represents the sensitivity of the sensor to that particular gas.

Gas concentration around the sensor can be inferred by measuring its resistance change. Both reactivity of sensing materials, temperature and humidity influence the sensor performance. 1.4 P-type sensors will not be described and are out of the scope of this thesis. Usually, these type of sensors have a sensitive layer deposited over a substrate provided with electrodes for the measurement of electrical characteristics. The device has a embedded heater separated from the sensing layer and the electrodes by an electrical insulating layer. [49]. A typical metal oxide sensor response is shown in Figure 1.5. Starting from a constant response in ambient conditions, there is a sensor voltage increase in response to an odour corresponding to the internal resistance decrease followed by an odour recovery time when the odourants slowly leave the sensor surface until the baseline is achieved once more.

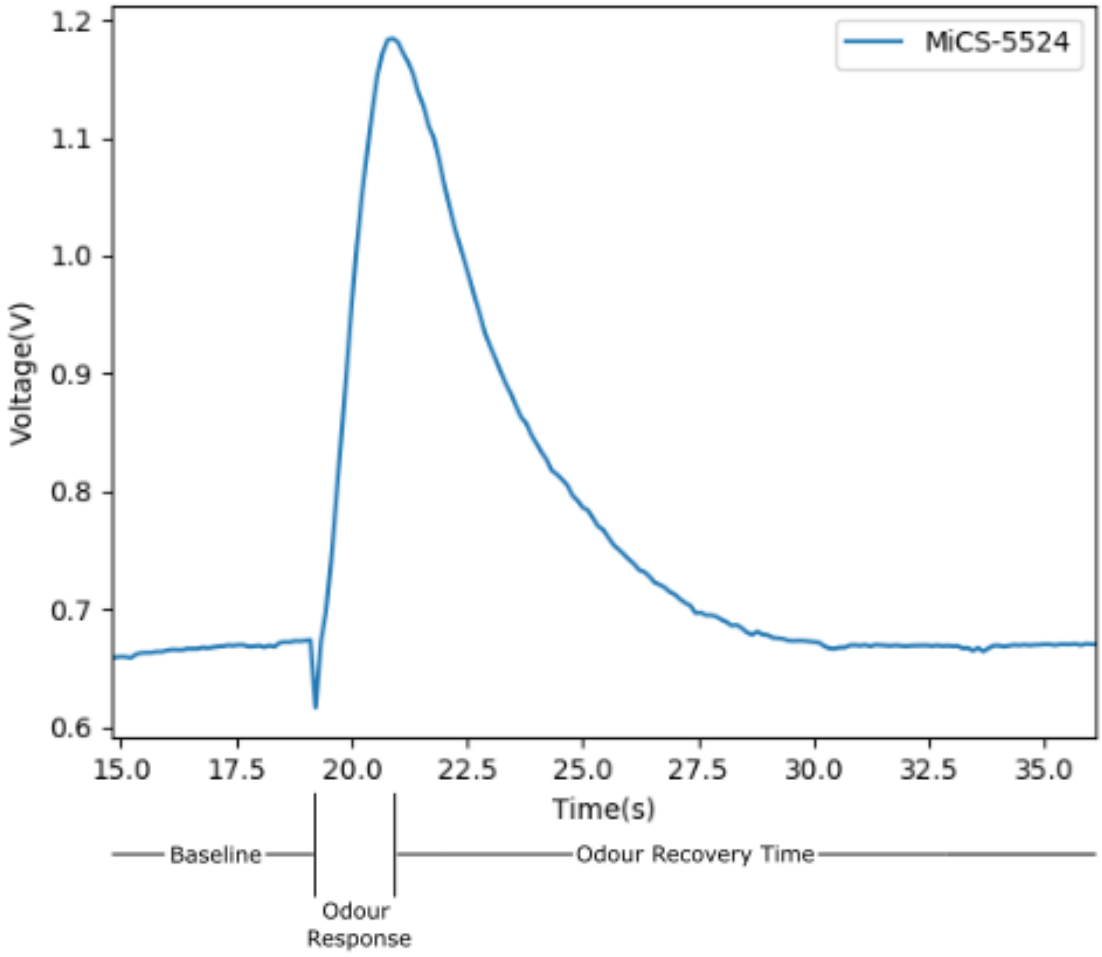


Figure 1.5: Characteristic response of a metal oxide sensor. This specific case is the response of a MiCS-5524 sensor to ethanol exposure.

1.5.2 Applications

1.5.2.1 Environmental control

One of the main incentives to the development of electronic noses was environmental control. Before this type of technology was around, this task was mainly performed by trained personnel that had the responsibility to smell a specific area and report their findings in order to evaluate the odours present. One of the problems of the field was that there are many environmental pollutants that are odourless like carbon monoxide or nitrogen monoxide. Those two are seriously toxic to human beings and not being able to detect them is a real issue. [50] The development of gas sensors sensible to these type of pollutants was a groundbreaking change in monitoring field. Nowadays there are environmental monitoring stations spread around the world equipped with this devices that are monitoring the environmental 24/7 and reporting that data to governmental agencies that can analyse it and act towards a safer environment to everyone and sanction those that pollute it.

1.5.2.2 Food quality control

There are several food related studies that use electronic noses to register product odours fingerprints and to differentiate between different types of wines for example, distinct types of tea (variations of the seeds), quality of ice stored meat and fish products, type and quality of coffee beans and other types of fruits [51].

Those references prove that the usage of this device to perform quality control is only going to have a wider range of applications in this area. It is a relatively affordable way of doing quality control with minimum participation of human labour in the process. It might not dispense professional evaluation and reporting but it might take a more centric and powerful presence in the industry's quality assessment work flow [52].

1.5.2.3 Disease diagnosis

There are a couple of applications of electronic noses in disease diagnosis. To name a few: urine evaluation [53,54], breath evaluation in order to diagnose diabetes [55,56], respiratory problems that are detectable in the expired air like asthma [57], hepatic diseases and several cancer types: lung [58–60], urinary track [61], prostate [62] and colorectal [63]). This technology might be useful in helping robots to assist the most

vulnerable aged segments of the population considering several other diseases show a different volatile organic compound profile and thus can be characterized for a diagnosis [64].

Chapter 2

Materials and Methods

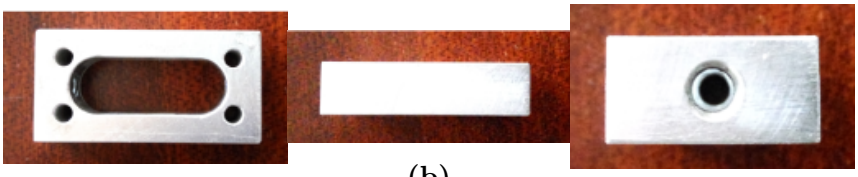
Several materials had to be gather in order to develop an experimental setup capable of achieving the goals of this thesis. A series of data handling methodologies were selected with the objective of accomplish classification tasks. In the section 2.1, the group of elements needed for the experiment are presented along with the reasons behind their selection. In the succeeding section, section 2.2, data processing methods and algorithms are shown and explained.

2.1 Hardware Elements

In subsection 2.1.1 the process of design and inception of a gas chamber is presented, 2.1.2 goes in depth about sensor picking and designs, 2.1.4 reflects on the reasoning behind the assortment of the odours chosen and 2.1.5 describes the odour containers used. Subsection 2.1.6 shows the analog-to-digital converter used in this experiment and 2.1.7 gives a detailed report on the specifics and properties of the Orange Pi, a single-board computer selected to orchestrate the system.

2.1.1 Gas exposure chamber

In order to develop an experimental electronic nose, a low dimensions chamber was designed to take the role of a artificial nostril. The chamber was designed via 3d software design tool SolidWorks®. A stable and odourless aluminium alloy was picked as the constituent of the chamber to neglect major interferences with odours to be tested. Avoiding gas leaks from inside was a crucial issue in the design process. A opening in the middle of the chamber, from top to bottom, was



(a) (b) (c)

Figure 2.1: Three different views of the aluminium chamber: a) top view , b) front view and c) side view.

included to facilitate the sensors attachment afterwards. The 3d designs are present in Appendix C and chamber views are presented in 2.1.

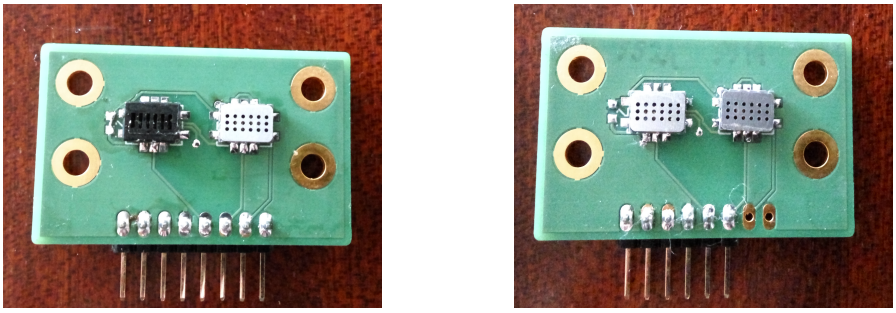
This aluminium chamber was designed in order to accommodate the sensors chosen and to have a reasonable area of contact between them and the odour molecules that would be pumped inwards via an air pump coupled with it. Chamber dimensions were chosen to be as small as possible in order to facilitate increased odour detection. The material itself does not partake a substantial role in the exposition process and the air flow and concentration of odours inside the chamber are not perturbed by it. Each side of the chamber has a circuit plate with two different sensors attached and they are presented in Figure 2.2.

2.1.2 Sensor boards

Four different metal oxide gas sensors were chosen to partake in this experiment: MiCS 5524, 5914, 2614 and 4514. Their printed circuit boards schematics can be consulted in Appendix B. The selection of these specific type of sensors was made in order to maximize the scope of the volatile spectre covered taking into account laboratory resources availability. Their target analytes are presented in Table 2.1 and their volatile detection capability seemed robust enough to the desired experiment. These sensors were then soldered to the circuit boards accordingly as depicted in Figure 2.2.

Table 2.1: Characteristics of the sensors chosen developed by SGX Sensortech®.

Sensor types	Sensor number	Target analytes
MiCS-5524	S1	Carbon monoxide, ethanol, hydrocarbons, VOCs
MiCS-5914	S2	Ammonia
MiCS-2614	S3	Ozone
MiCS-4514	S4	Carbon monoxide, nitrogen dioxide



(a) (b)

Figure 2.2: The two different printed circuit boards attached to the aluminium chamber. On a) MiCS 4514 on the left and MiCS 2614 on the right and on b) MiCS 5524 on the left and MiCS 5914 on the right.

Load resistors for each of the 4 sensors are placed externally in a breadboard. This way it was possible to change and adapt the resistor values towards our experimental goals and limitations. Aside from a initial tuning of the values those parameters were not changed during the sampling phase.

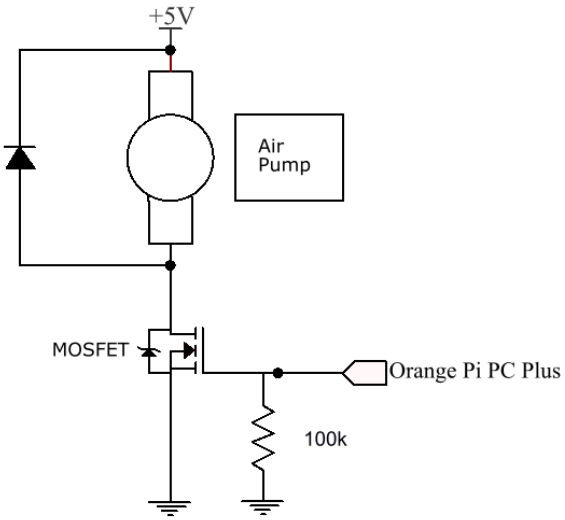


Figure 2.3: Handling circuit for the air pump.

2.1.3 Air Pump

A small motor coupled with a diaphragm serve as an air pump to effectively withdraw air from inside the gas chamber during the odour injection and to not overexpose the gas sensors. This motor is powered and controlled by one of the generic pins of the Orange Pi that powers both ADC and sensor boards as well. The flow which

this pump is able to withdraw air from the chamber is the same throughout the sampling procedure. The circuit present in figure 2.3 is the handling circuit needed to make sure that voltage that flows through the pump does not affect the normal functioning of the single-board computer used. The resistor is a security measure in a case of a sudden voltage turn off. MOSFET works as an interrupter that is controlled by the Orange Pi as mentioned earlier. The flyback diode in parallel with the air pump has the function to eliminate flyback voltage spikes that may happen due to an eventual pump's motor shutdown or interruption. This could potentially damage the circuitry involved.

2.1.4 Odours

In order to include several areas of daily life experiences as presented in Figure 1.1, the nature of the group of odours chosen is very broad. Specific reasons to each one are given in Table 2.2.

Table 2.2: Description of the odour sources and their main odorant components.

No	Odour	Major odour components	Reasoning
1	Ethanol 96%	Ethanol	Toxic and inflammable
2	Garlic	Aliin, Allicin	Strong aroma, often used in food preparation
3	Window cleaner	Ammonia, 2-Butoxyethanol	Toxic
4	Gasoline	Benzene, Xylene	Inflammable and toxic
5	Lemon peels	Limonene	Pleasantness, food preparation
6	Bay leaf	Myrcene	Medium strong aroma, food preparation
7	Vinegar	Acetic acid	Strong aroma and pleasantness
8	Mothballs	Naphtalene	Toxic

2.1.5 Odour containers

Odour containers are bottles with odour inside that serve as the odour source for the experimental setup. The set of odour containers selected is portrayed in Figure 2.4. Each lid has one perforated hole where a plastic tube is attached to and serves as a pathway towards the sensors nostril. Although there is a way to volatiles emanated from the odour source to leave the bottle, the majority surface of the container is sealed from the outside environment. This near complete isolation promotes an accumulation of odour compounds in the headspace until there is a mass equilibrium between the odour source and the air that surrounds it. This group of volatiles are responsible for the odour that sensors are exposed to. It is also worth noting that

organic odour sources do not emanate the same volatiles due to the action of internal and external oxidation.

Figure 2.4: Group of odour containers and respective odours used in the experiment. From right to left: ethanol, garlic, ammonia, gasoline, lemon peels, bay leaf, vinegar and mothballs.

2.1.6 ADS1115

ADS1115 is a small, medium precision, four channel analog-to-digital converter. It is well suited for power and space constrained sensor measurement applications. It functions as a versatile acquisition module that is capable of connecting to any computerized device such as microcontrollers or single-board computers. This device works as a converter of analog signals that are fed to it via its four channels, transforms those signals into a digital state and, communicate this information to a device afterwards.

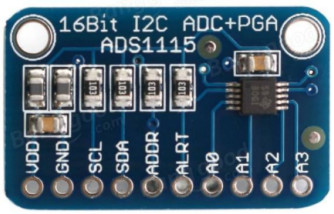


Figure 2.5: ADS1115 analog-to-digital converter developed by Texas Instruments®.

2.1.7 Orange Pi

Orange Pi PC Plus is a single-board computer capable of achieving what is desired from a mid to low end computer. Its price and availability makes it a good choice to control and managing experiments like the electronic nose developed. This technology substitutes in part the usage of a microcontroller, avoiding low level software development and configuration. It is also worth noting that the board is a high level software platform and using a microcontroller would require deeper low level knowledge that it was beyond this experiment. Data storage, sampling collection as well as circuitry power supply and handling are its main functions in the experimental setup. It is worth noting that, as can be seen in Figure 2.6, this board supports wireless communicability and can take part of a potential future application of this sensing technology in a real life context. HDMI port allows a graphic interface usage of this computer via a monitor. It is also possible to access and configure it via SSH protocol over a common shared network with security.

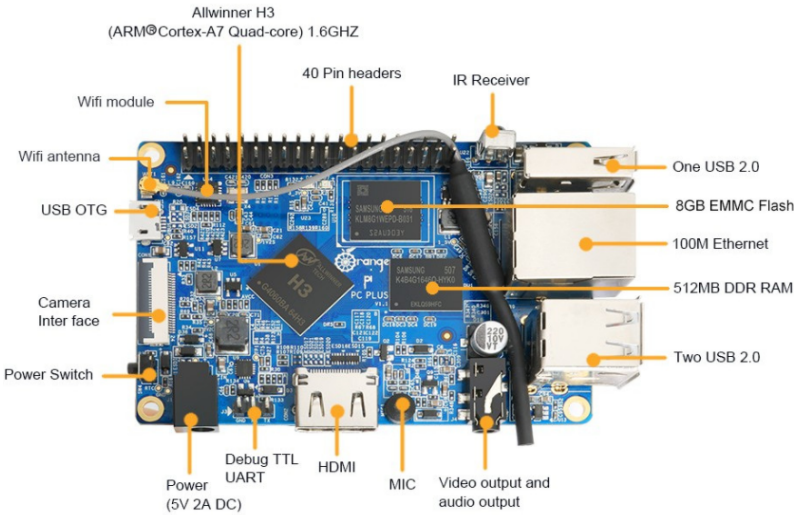


Figure 2.6: Orange Pi PC Plus single-board computer developed by Shenzhen Xunlong Software CO., Limited ®.

2.2 Data Processing

Data samples collected from this assembled electronic nose need to be processed before a search for insights is initialized. In order to implement a pattern recognition system in a way it is possible to classify correctly an unseen odour sample, a similar pipeline to the one presented in Figure 2.7 with those particular different steps needs

to be implemented. In the following subsections, the strategy planned for each of those steps is described and its fundamentals stated.

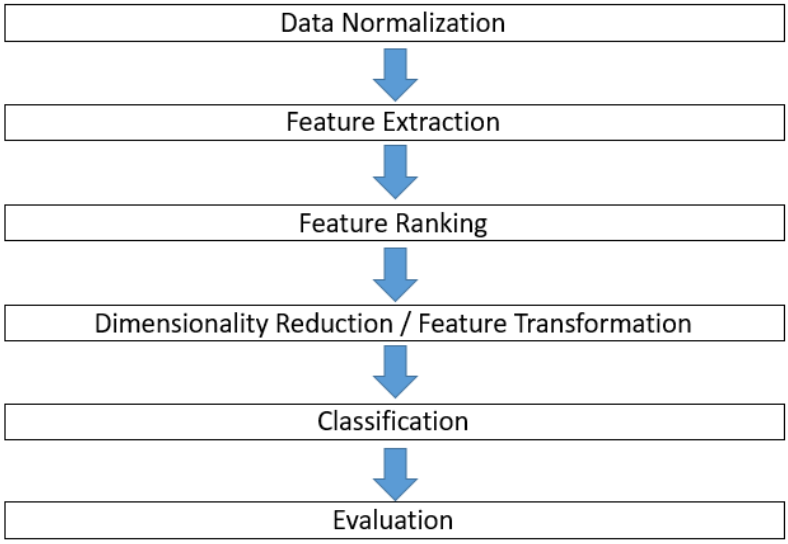


Figure 2.7: Data processing pipeline different steps.

2.2.1 Data Normalization

In order to better compare data from the four different gas sensors between themselves, two different data normalization strategies were chosen to be applied in the dataset collected for different reasons explained in the subsections above.

2.2.1.1 Normalization 0 to 1

This type of normalization, also referred as min-max normalization, is applied to datasets in order to transform their values maintaining their variance. Value dimensions alongside the features were very sparse and so normalization was applied to ease that difference and place all the values between 0 and 1. Equation 2.1 mathematically describes this data transformation.

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad for\ i\ \epsilon\ [1,3612] \tag{2.1}$$

2.2.1.2 Logarithmic Normalization

Logarithmic normalization excels in the case of datasets with variables highly skewed smoothing them. This methodology was chosen considering sensor responses in this experiment are generally skewed towards the left due to the nature of the sampling method and it is applied with the objective of counterbalancing this phenomenon. It is also applicable to data that present high spikes of magnitude. Equation 2.2 describes the logarithmic transformation of base 10 applied to the data.

$$z_i = \log_{10}(x_i) \quad for\ i\ \epsilon\ [1,3612] \tag{2.2}$$

2.2.2 Feature Extraction

Feature extraction takes a significant role on the performance of an electronic nose. The aim of this procedure is to gather a non redundant and robust set of information from the sensors response in order to improve the effectiveness of the ulterior classification/pattern recognition tasks.

From the different types of transient features that can be extracted from an electronic nose response signal, parameters directly extracted from the original response curve are usually the first to be focused on considering its simplicity and they are usually a decent start in order to ease the task of differentiating between different odours. One sample example is shown in Figure 4.1e.

For the analysis of the data collected, several features were extracted from the data. Some of these features are directly computed from the curve parameters of the sensors response, others have their origin in the time or in the frequency domains relating to them. The group of features selected are described in the following subsections. Those were selected with basis on several literature approaches that analysed related signals.

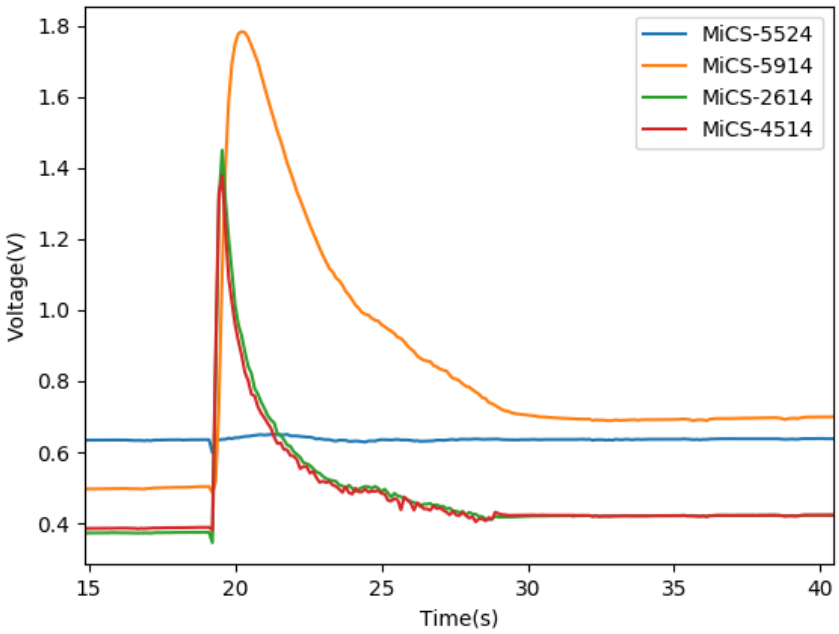


Figure 2.8: Example of a lemon odour response from the sensors utilized.

2.2.2.1 Sample Mean

The mean of each data sample is obtained via Equation 2.3.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}, \quad n = 3612 \tag{2.3}$$

2.2.2.2 Upslope Mean

The upslope mean obtains the mean of the data samples between the end of the baseline section of the sensor response, empirically defined as the 165th data point due to the sampling method procedure intrinsics, till the maximum data point achieved. Equation 2.4 describes this method.

$$\bar{x}_{upslope} = \frac{x_{165} + x_{166} + ... + x_{max}}{i_{max} - 165} \quad i \, \epsilon \, [165, i_{max}] \tag{2.4}$$

2.2.2.3 Downslope mean

The downslope mean, as opposed to the upslope, is a mean characterization of the slope of the data from the maximum data point until the 300th one. This number is defined empirically as well from data inspection taking into account sensor desorption and stabilization.

$$\bar{x}_{downslope} = \frac{x_{max} + ... + x_{299} + x_{300}}{300 - i_{max}} \quad i \in [i_{max},300] \tag{2.5}$$

2.2.2.4 Upslope Derivatives Mean

Similarly to the mean described in the subsection 2.2.2.2, the same methodology is applied but computing the mean of the derivatives in the same dominion, as is described in equation 2.6.

$$\overline{(x_{iupslope})'} = \frac{(x_{i+1})' - (x_i)'}{x_{i+1} - x_i}, \quad i \in [165, i_{max}] \tag{2.6}$$

2.2.2.5 Downslope Derivates Mean

The mean of the derivatives of the data points in the downslope section considered is described in equation 2.7.

$$\overline{(x_{idownslope})'} = \frac{(x_{i+1})' - (x_i)'}{x_{i+1} - x_i}, \quad i \in [i_{max}, 300] \tag{2.7}$$

2.2.2.6 Derivatives Mean

The mean of the derivatives from the whole sample is computed as well, as presented in equation 2.8.

$$\overline{(x_i)'} = \frac{(x_{i+1})' - (x_i)'}{x_{i+1} - x_i}, \quad i \in [1, 3612] \tag{2.8}$$

2.2.2.7 Discrete Wavelet Transform

Wavelet transform [65] is an extension of Fourier transform. It maps the signal into a new space with basis functions quite localizable in time and frequency space. They are usually of compact support, orthogonal/biorthogonal and have proper Lipschitz regularity and vanishing moment. The wavelet transform decomposes the original response into the approximation (low frequencies) and details(high frequencies). Using the wavelet coefficients of certain sub-bands as features might improve further pattern recognition techniques considering this discrete transform has reasonable anti-interference ability [66].

Mathematically speaking, a wavelet is a type of waveform with limited duration and has a average value of zero. A wavelet $\Psi(t)$ is defined in equation 2.9 and its intrinsic characteristics are show in equation 2.10.

$$\Psi \, \epsilon \, L^2(\mathbb{R}) \quad and \quad \int_{-\infty}^{\infty} \Psi(t)dt \; = \; 0 \tag{2.9}$$

$$\Psi_{j,k} = \frac{1}{\sqrt{j}}\Psi(\frac{t-k}{j}) \tag{2.10}$$

Where:

- $L^2(\mathbb{R})$ is finite energy function $\int_{-\infty}^{\infty} |f(t)|^2 dt < +\infty$
- j is a scaling factor
- k is a shifting factor

Discrete wavelet transform emerged with the objective of tackling the heavy cost of computing this type of waveform. It is implemented by selecting a group of restricted values for the scaling and translation parameters. In this type of analysis, a wavelet system is created by scaling (j) and shifting(k) a scaling function and a wavelet function as shown below in Equations 2.11 and 2.12.

$$\phi_{j,k} = 2^{1/2}\phi(2^j t - \; k) \tag{2.11}$$

$$\psi_{j,k} = 2^{1/2}\psi(2^j t - \ k) \tag{2.12}$$

Those functions are then used to generate a set of basis function such that any signal or function $f(t)\epsilon L^2$ can be expressed by equation 2.13.

$$\begin{aligned} f(t) &= \sum_k a_{j_0}(k)\phi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\psi_{j,k}(t) \\ &= \sum_k a_{j_0}(k)2^{j_0/2}\phi(2^{j_0}t - k) \\ &\quad + \sum_k \sum_{j=j_0}^{\infty} d_j(k)2^{j/2}\psi(2^j t - k) \end{aligned} \tag{2.13}$$

(2.14)

Where:

- $a_j(k)$ is designated a scaling/approximation coefficient
- $d_j(k)$ is designated a wavelet/detail coefficient
- $j,k\ \epsilon\ \mathbb{Z}$ (integer number)

In a multi-resolution analysis, the scaling and wavelet functions can be defined as presented in equation 2.15 and 2.16.

$$\phi(t) = \sum_n h_0[n]\sqrt{2}\phi(2t - n) \tag{2.15}$$

$$\psi(t) = \sum_n h_0[n]\sqrt{2}\psi(2t - n) \tag{2.16}$$

Where:

- $h_0[n]$ correspond to the coefficients for creating the scaling/approximation function
- $h_1[n]$ correspond to the coefficients for creating the wavelet/detail function

Solving *f(t)* for a_j and d_j,

$$a_j(k) = \sum_m h_0[m - 2k]a_{j+1}(m) \tag{2.17}$$

$$d_j(k) = \sum_m h_1[m - 2k]d_{j+1}(m) \tag{2.18}$$

where $m = 2k + n$.

The decompositions presented in equations 2.17 and 2.18 can be computed by applying a low pass and a high pass filter, also called filter bank algorithm. Consecutive decompositions processes on approximation coefficients can be applied in order to obtain the various multilevel DWT decompositions. This decomposition process is depicted in Figure 2.9.

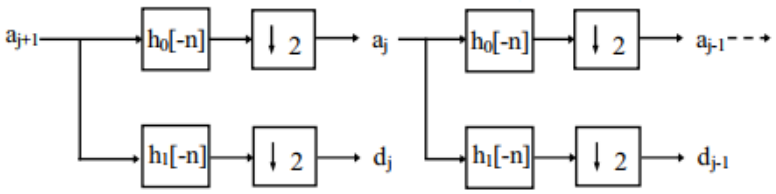


Figure 2.9: Multilevel wavelet decomposition, where a_j correspond to the approximation coefficients and d_j to the details coefficients.

According to [67], the best wavelet family type to apply in this electronic nose context is Daubechies 6(db6), portrayed in figure 2.10 considering gas sensors signals are often affected by various type of noise, which appears mostly in high frequency components.

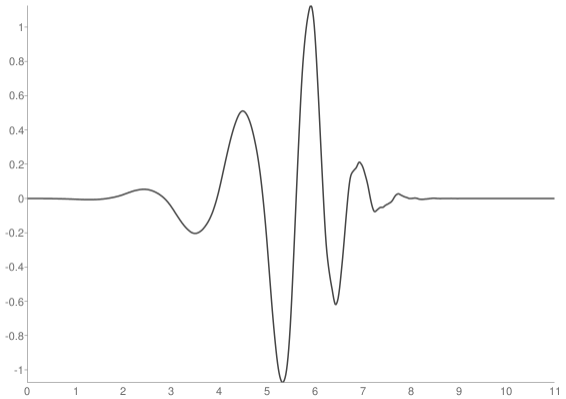


Figure 2.10: The Daubechies 6 wavelet function.

Taking this into consideration, approximation coefficients give better transient feature extraction and focus on low frequencies. A reasonable decomposition level is required as well to avoid high frequency presence in the coefficients obtained, so a wavelet decomposition of 8 levels is applied, the maximum possible taking into account the size of the data samples collected.

As an overview, Table 2.3 describes the features extracted with an abbreviation associated to facilitate references to a certain feature in specific. To this end, sensors order from 1 to 4 is identical to the one present in Table 2.1: MiCS-5524, MiCS-5914, MiCS-2614 and MiCS-4514.

Table 2.3: Overview of the features extracted.

Abbreviation	Description
mx	Mean of the response of sensor x.
um$x$	Mean of the upslope of sensor x from the end of baseline until the maximum of the response.
dm$x$	Mean of the downslope of sensor x from the maximum of the response until the following 300 points.
dvm$x$	Derivatives mean of response of the sensor x.
udm$x$	Derivatives mean of the upslope of sensor x from the end of baseline until the maximum of the response.
ddm$x$	Derivatives mean of the downslope of sensor x from the maximum of the response until the following 300 points.
wvxy	Five most discriminative approximation coefficients from Daubechies 6 (db6) discrete wavelet transform from sensor x and corresponding index y.

2.2.3 Feature ranking/transformation

2.2.3.1 Recursive Feature Elimination

Recursive Feature Elimination is a feature ranking method that takes an external estimator that has weights assigned to the features and recursively selects them considering smaller and smaller sets. In each iteration, the trained estimator ranks each feature in the data according to their importance and cuts a set of features only to repeat this process until a desire number of features selected is reached. Selecting one feature only enables the algorithm to construct an organized ranked list of feature importance [68].

2.2.4 Dimensionality reduction

Dimensionality reduction is often applied to datasets with a large amount of features in order to remove redundant information and to decrease the computational hassle of processing several dozens of features. Principal Component Analysis and Linear Discriminant Analysis are both described in the following subsections.

2.2.4.1 Principal Component Analysis

Principal component analysis is a simple and non-parametric dimensionality reduction technique often used when the data considered is sparse and multidimensional with minimal loss of information. The main objective of this algorithm is to order the data according to their maximum variance. This analytic method calculates which two of the features (dimensions) of the data are most discriminative (contain more variety, the largest variance possible) of the information contained in it. Each principal component has the maximum variance possible considering they must be orthogonal towards the preceding components. One example of a two dimensional PCA projection is shown in Figure 2.11. Taking in consideration that data can be hard to find in high dimensional data, PCA is a powerful tool in order to circumvent this problem and analyse the data [69].

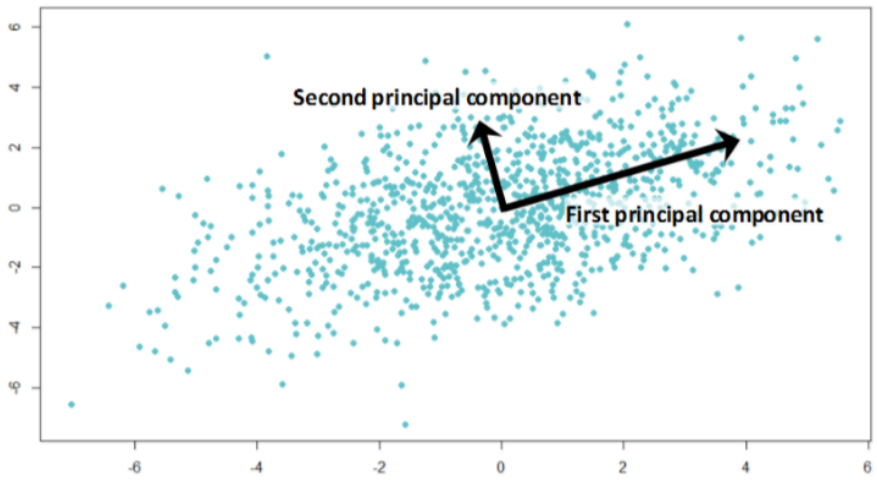


Figure 2.11: Data plot example where it is clear the directions of the two main principle components of this dataset.

Assuming the goal of its application is to reduce d-dimensional dataset by projecting it into a k-dimensional subspace (where $d<k$), this mathematical transformation starts by computing the d-dimensional mean vector out of the initial dataset. Af-

terwards the scatter matrix of this dataset is computed. This process is show in Figure 2.19.

$$S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^T \tag{2.19}$$

Where $\boldsymbol{m}$ is the mean vector

$$m = \frac{1}{n} \sum_{k=1}^{n} x_k \tag{2.20}$$

In the following step, eigenvectors and corresponding eigenvalues are computed. Eigenvectors are sorted by decreasing eigenvalues and $\boldsymbol{k}$ eigenvectors with the largest eigenvalues are chosen to form a $\boldsymbol{d}$ x $\boldsymbol{k}$ dimensional matrix $\boldsymbol{w}$. At last, the dataset samples are projected onto the new subspace via Equation 2.21.

$$y = w^T \times x \tag{2.21}$$

Where:

- w is dxk dimensional matrix. d is the dataset dimensions and k is the desired subspace dimensions.

2.2.4.2 Linear Discriminant Analysis

Linear discriminant analysis is a method used in statistics and in pattern recognition tasks to find a linear combination of features that characterizes or separates two or more classes of the problem considered searching for the in-between-class maximum separation. The result of this method can be used as a linear classifier (more precisely in binary classification problems) or as a dimensionality reduction technique. Linear discriminant analysis is related to PCA considering they both look for linear combinations of variables which best explain the data [70]. LDA attempts to model the difference between the classes of the data while PCA does not take into account any difference in the classes.

2.2.5 Classification

There are several possible classification algorithms available and there are new algorithms being developed and polished everyday. Some of the most used and more reliable will be briefly presented below and are applied in the next section of this thesis.

2.2.5.1 Support Vector Machine

Support vector machine is a classification method that tries to find a maximum-margin hyperplane that lies in a transformed input space and splits the classes, while maximizing the distance to the nearest cleanly split examples.

Considering a set of training data (x_i,y_i), i=1,..., n, $y_i \epsilon$ (-1,1), $x_i \epsilon \mathbb{R}^d$, where x_i are the feature values and y_i the class labels. Hypothesising there is a hyperplane that separates the two classes, the points x which lie on the hyperplane satisfy Equation 2.22.

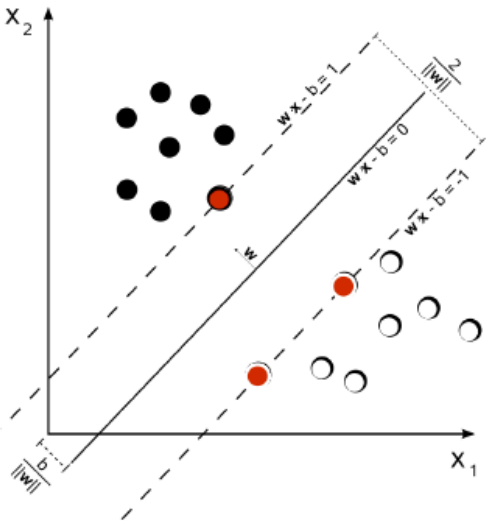


Figure 2.12: Schematics of SVM applied to a linear separable binary class problem.

It is also known that the distance from the hyperplane to the origin is $|b|/||w||$. An example of this type of hyperplane is shown in Figure 2.12. The distance of any data point x_i to the hyperplane is $|w'x_i + b|/||w||$. The vectors corresponding to the closest points are called support vectors and their distance from the hyperplane is $1/||w||$.

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.22}$$

Where:

- w is the normal vector to the hyperplane

The principle of this algorithm is to maximize the separation margin $2/||w||$, which is described by the minimisation criterion $\Psi(x)$, given by the following equation.

$$\Psi(w) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \quad y_i(w'x_i + b) \ \geq 1 - \xi_i, \ i = 1, ..., n \tag{2.23}$$

Where C is a constant that is highly related with the accuracy on the training set and the margin width. This variable plays the role of regularization constant. It is only used in non-linearly separable classes. ξ represents the separation margins. In a case that this value is less than zero, the data points could be in separation region between the hyperplane and the margin, otherwise misclassification of some points is allowed. C controls the influence of ξ in the minimisation criterion Ψ. An high tolerance to misclassification corresponds to a large value of C, whereas if C is small the influence of ξ will be increased, leading to data points being misclassified with overall shorter margins. This description above focused on the binary cases in which this algorithm was designed for and it is applied to. In order to apply it to a multi-class problem like this one, a combination of several binary SVM classifiers is employed. From the various strategies that exist to combine those, this thesis uses the one-vs-one scheme. A separate classifier is created for each different pair of class labels.

2.2.5.2 Backpropagation Neural Network

A feedforward neural network with backpropagation is the most widespread and more generally used artificial network architecture. An artificial neural network is an algorithm that is loosely modelled after the neural structure of the mammalian cerebral cortex although on a small scale. They are organized in layers which in turn have a series of interconnected nodes also called neurons. A pattern is presented to the network in its input layer, which communicates to one or more hidden layers

where a system of weighted connections performs the pattern processing. Every node does a summation of their respective weights and it passes through a non-linear function known as a transfer function that is commonly shared with their counterparts in the same layer. The structure of a neural network is shown in Figure 2.13

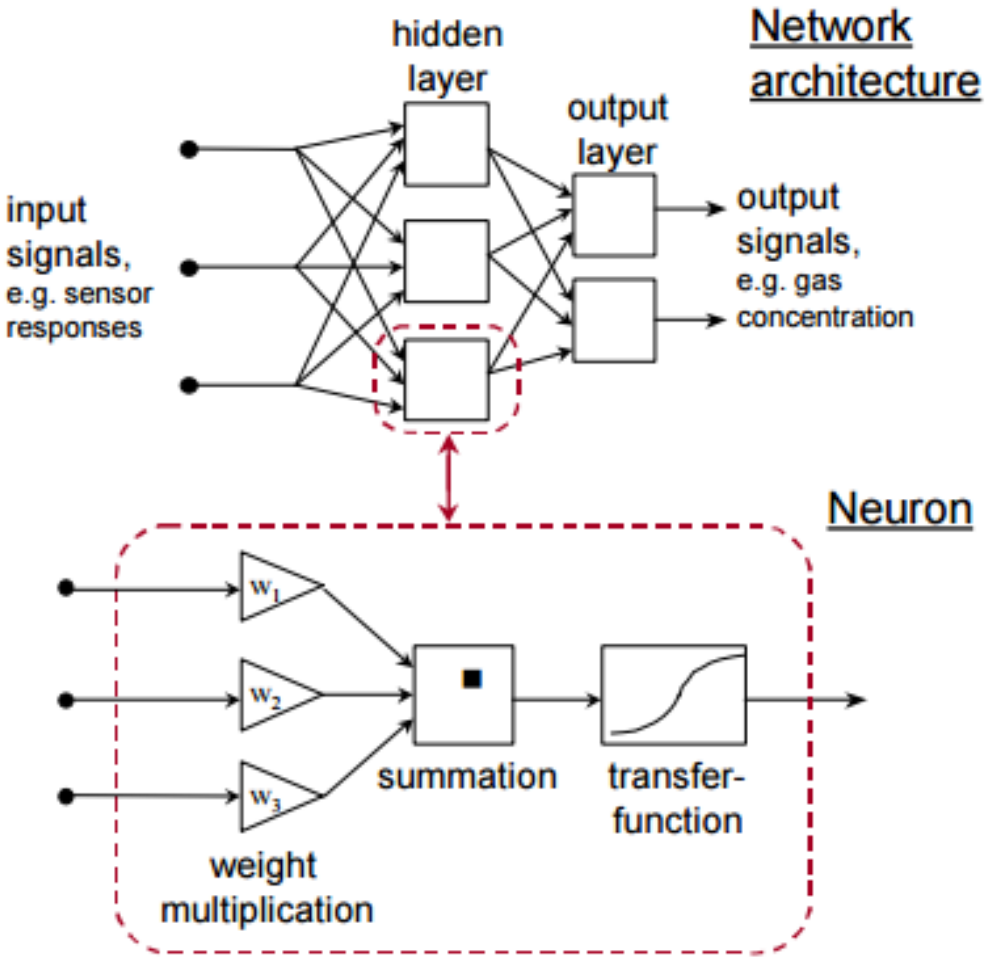


Figure 2.13: Schematics of the structure of a neural network and its neuron structure (Adapted from: [71])

Afterwards, the hidden layers share their outcomes to an output layer that is responsible to present the output obtained. The loss function of this process provides the error rate which in turn is used to backpropagate over the neural network so it can minimize the learning error and thus learn better and deeper the training cases it has been fed with. This property gives the network feedback about its current learning procedure and it is the main advantage of this type of architecture. Although there are several methodologies to perform backpropagation, the most common is

the generalized version of the Delta rule that is capable of dealing with networks with hidden layers [72]. Gradient descent is the optimization algorithm often used to search the net weights that minimize its learning error [73].

The backpropagation takes place after the feedforward computation of the artificial neural network. The process starts in the output layer, where the first set of derivatives is calculated. Then it goes throughout all the hidden layers in order to compute the error that it is trying to minimize [74].

Gradient descent is the most used optimization algorithm in the backpropagation step of a neural network and has as objective the objective the minimization of the loss function considered and this, the learning error.

Generally a standard error function of a neural network is presented in equation 2.24, where p is the size of the training set, o_i is the output pattern and t_i is the desired target pattern.

$$E = \frac{1}{2}\sum_{i=1}^{p} ||o_i - t_i||^2 \quad for\ i = 1,..,p \tag{2.24}$$

The gradient alongside every dimension of the error is calculated via equation 2.25. Derivatives of the error give the algorithm the direction in which the minimization should follow.

$$\nabla E = (\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, ..., \frac{\partial E}{\partial w_l}) \tag{2.25}$$

Then, the algorithm searches the neighbourhood of a weight w in which the error function can decrease the fastest if one goes from w in the direction of the negative gradient of E at w, $\nabla E(w)$, as mentioned in equation 2.26.

$$\Delta w_i = -\gamma \frac{\partial E}{\partial w_i}\ for\ i = 1, ..., l \tag{2.26}$$

Where l is the number of weights and γ is the learning constant and dictates the quality of the gradient descent. With different values the step length of each iteration differs and so if the step is too large the algorithm might discover a local minimum

labelling it as the global minimum because the search step was too large. If the parameter is too low, the step might be so small that the duration of this search for the global minimum would take a considerate amount of time and processing power. This process is repeated until the error is not decreasing anymore.

2.2.6 Evaluation

In order to validate our results in the subsequent classification procedures we need to essentially make sure the results are not influenced too much with overfitting, which would decline our models capacity to generalize when new data is provided. Cross-validation is a technique used to divide the data available for training in small, balanced sets. Classifiers are trained with each different set of the original data and, at the end, classification results are obtained using the mean of all those results. In this case, the dataset was shuffled and 75% of the samples were randomly selected as training data and the other 25% as the testing data. K-fold cross validation was not applied but it certainly is a good strategy for future work. Accuracy will be the main metric of evaluation of classifiers applied in this thesis.

Chapter 3

Implementation

In this chapter, details about implementation of the experimental setup and sampling methodology employed are shown in Sections 3.1 and 3.2 respectively. How the data acquisition took place is described in Section 3.3 and data processing applied scripts are focused on the last Section, 3.4.

3.1 Experimental Setup

Taking into account the materials described in the previous chapter of this thesis, the experimental setup assembled is presented in Figure 3.1. In figure 3.2 the various steps and procedures of the setup are shown and more details about them are described in the following section.

3.2 Sampling Methodology

The methodology used to extract samples from the experimental setup was dynamic headspace sampling in which a motor coupled with a diaphragm is used as an air pump. The concept of this methodology consists in air being effectively drawn from an odour bottle towards the target sensing chamber via a thin tube. This motion promotes a sweep of volatile compounds released by the sample into the air stream that subsequently flows through the chamber. Scenario conditions in which the samples were collected is described in Table 3.1.

In order to obtain a guaranteed reaction from the sensors, one bottle is softly attached to the acquisition chamber with a tube. On the other side of the chamber,

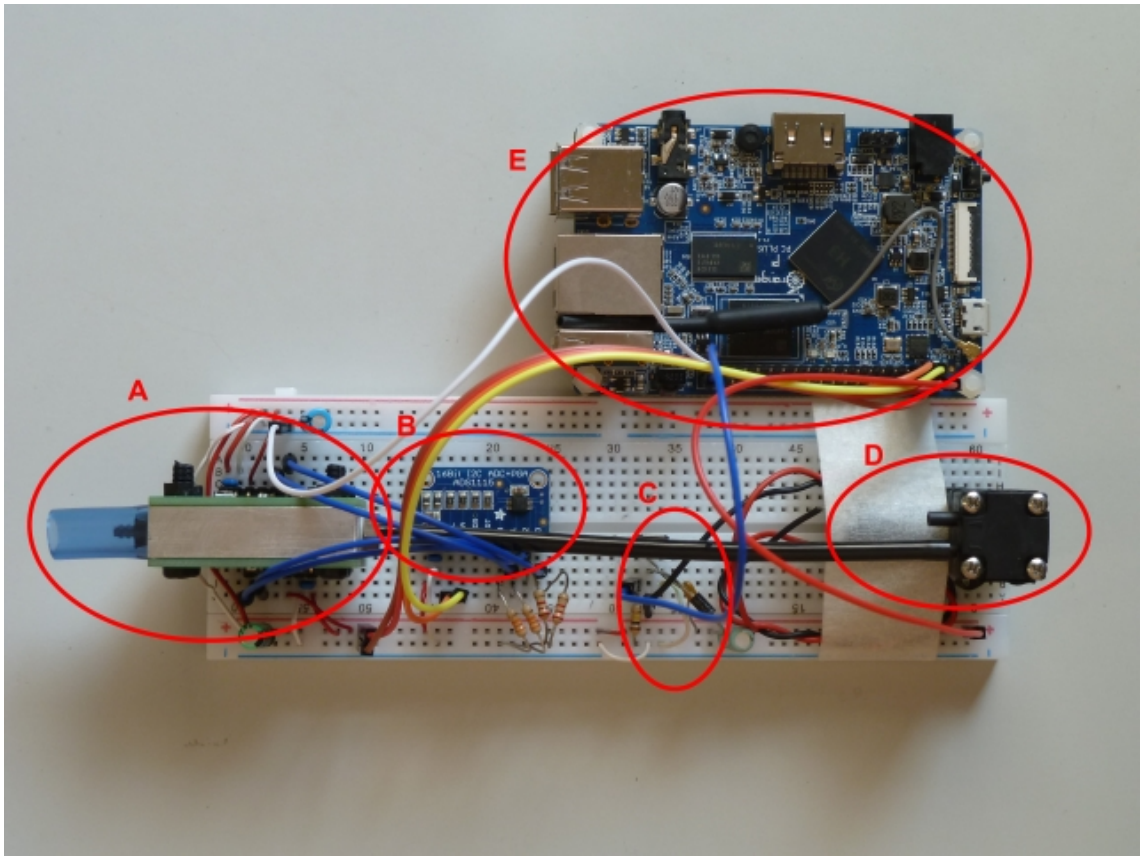


Figure 3.1: Experimental setup implemented: A - Electronic nose (sensor boards coupled with gas chamber); B - ADS1115 analog-to-digital converter; C - Air pump handling circuit; D - Air pump; E - Orange Pi PC Plus.

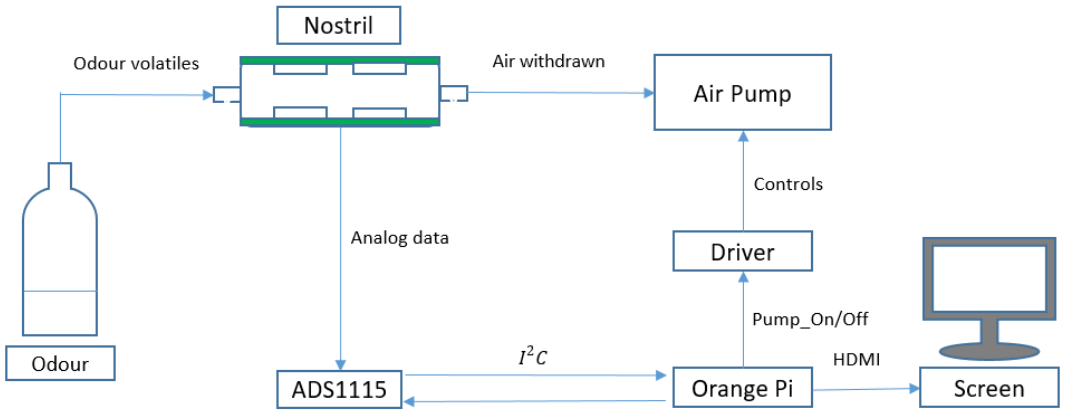


Figure 3.2: Experimental setup pipeline.

an air pump together with a soft and steady bottle squeeze ensure that the passage of airflow is facilitated as required. At the same time, the internal resistance values of the sensor array that are attached to the chamber are constantly being monitored

Table 3.1: Scenario and material specifications and conditions.

Parameters	Specifications
Chamber size	40 x 20 x 10 mm
Chamber gas volume	2585 mm^3
Sampling rate	9 Hz
Sampling time	420 s
Purging time	30 s
Working voltage	$\sim$ 5 V
ADC resolution	16-bit (ADS1115)
Temperature	19 $\sim$ 31 C°
Pressure	Ambient pressure

and registered with the ADS1115 analog-to-digital converter with 16 bit resolution in order to obtain a precise response value of each sensor for that specific odour. A pipeline that shows the various steps of the sampling procedure is shown in 3.2.

After the acquisition, there is a sensor desorption time in which the chamber loses its previous odour presence with the ambience odours replacing it and reaching a stable baseline. Each sample comprises 420 seconds, in which the first 20 seconds are dedicated to obtain the current baseline response; the pump works for 10 seconds extracting air out of the odour bottle and the last 390 seconds are dedicated to give time to the sensors approach their baseline values. An extra purging time of 30 s was applied between consecutive sample measurements. This procedure occurred for 8 different odours with 25 samples for each one of them. The Orange Pi is responsible for the electrical supply of the sensors chamber, the ADC and the air pump. The analog-to-digital converter is connected to an Orange Pi board and sends 16 bits data to the card-size computer. This board was configured for the purpose of being a controller board for the experiment, a data storage for the various samples collected and, eventually, a data processing platform in which the pattern recognition algorithms would be applied and a decision would be made about the odour facilitated. During sampling time, values arriving in the Orange Pi are printed in a monitor connect to it via HDMI port facilitated by the sampling acquisition script below presented in Code Snippet 3.1. This way it was possible to disregard sampling anomalies and impactful errors that would damage the next steps of this work.

Code Snippet 3.1: Some Py code

```
import time
```

```
import Adafruit_ADS1x15
from pyA20.gpio import gpio
from pyA20.gpio import port
%
t_start = time.time()+20;
t_end = time.time()+30;
t_finish = time.time()+420;)
gpio.init()
gpio.setcfg(port.PG6, gpio.OUTPUT)
# Create an ADS1115 ADC (16-bit) instance.
# Bus number used is 0
adc = Adafruit_ADS1x15.ADS1115(address=0x48 , busnum= 0)
GAIN = 2/3
print('Reading ADS1x15 values , press Ctrl-C to quit ...')
# Print nice channel column headers.
print('| {0:>6} | {1:>6} | {2:>6} | {3:>6} | '.format(*range
    (4)))
print('-' * 37)
odourname = raw_input("Odour: ")
filename = time.strftime(odourname+'---' +"%Y-%m-%d_%Hh%Mm%
    Ss", time.gmtime())+'_log.txt'
f = open(filename , 'w')
data = []
while True:
        if time.time()>=t_start or time.time()<t_start+1:
        gpio.output(port.PG6, HIGH);
        # Read all the ADC channel values in a list.
        values = [0]*4
        for i in range(4):
        # Read the specified ADC channel using the
            previously set gain value.
                values[i] = adc.read_adc(i, gain=GAIN,
                    data_rate = 128)

        s=str(values[0])+'\t'+str(values[1])+'\t'+str(values
            [2])+'\t'+str(values[3])+'\t'+time.strftime("%Hh%
            Mm%Ss", time.gmtime())+'\n'
```

```
            print(s)
            if(len(data)==1000):
                    for d in data:
                            f.write(d)
                    f.write(s)
                    data=[]
                    else:
            data.append(s)
            if time.time()>=t_end and time.time()<t_end+1
            gpio.output(port.PG6, LOW);
            if time.time()>=t_finish:
            break;
            time.sleep(0.1)
f.close()
```

3.3 Data Acquisition

When the data collecting script is running, sensor voltage differences pass from sensors to the circuit board that, in turn sends to the breadboard where adequate load resistors are allocated. Then, new transformed differences are sent to the analog-to-digital converter, where the data is transformed to a digital state. This device gather the sensor differences in one single stream of data that is communicated through a specific GPIO pin - PA12/SDA, that stands for serial data - of the Orange Pi utilising the I^2C (inter-integrated circuit) protocol, where this data stream is controlled by a clock pulse sent also by the ADC to the serial clock pin PA11/SCK. The disposition of the GPIO pins used is shown in Table 3.2.

This pulse dictates when the receiver reads the data from the data stream sent by the ADC. Considering the data transmitter originates both the data and the reading timing pulses, the Orange Pi only reads the data when the ADC determines, enabling a guaranteed synchronization and a lossless data transmission. Afterwards, this data is stored in the single-board computer labelled by odour name prompted by the user.

Table 3.2: GPIO bus pin information from Orange Pi PC Plus.

2x20 Header			
1	*3.3V*	2	*5V*
3	PA12 (TWI0_SDA/DI_RX/PA_EINT12)	4	*5V*
5	PA11 (TWI0_SCK/DI_TX/PA_EINT11)	6	*GND*
7	PA6 (SIM_PWREN/PWM1/PA_EINT6)	8	PA13 (SPI1_CS/UART3_TX/PA_EINT13)
9	*GND*	10	PA14 (SPI1_CLK/UART3_RX/PA_EINT14)
11	PA1 (UART2_RX/JTAG_CK/PA_EINT1)	12	PD14
13	PA0 (UART2_TX/JTAG_MS/PA_EINT0)	14	*GND*
15	PA3 (UART2_CTS/JTAG_DI/PA_EINT3)	16	PC4
17	*3.3V*	18	PC7
19	PC0 (SPI0_MOSI)	20	*GND*
21	PC1 (SPI0_MISO)	22	PA2 (UART2_RTS/JTAG_DO/PA_EINT2)
23	PC2 (SPI0_CLK)	24	PC3 (SPI0_CS)
25	*GND*	26	PA21 (PCM0_DIN/SIM_VPPPP/PA_EINT21)
27	PA19 (PCM0_CLK/TWI1_SDA/PA_EINT19)	28	PA18 (PCM0_SYNC/TWI1_SCK/PA_EINT18)
29	PA7 (SIM_CLK/PA_EINT7)	30	*GND*
31	PA8 (SIM_DATA/PA_EINT8)	32	PG8 (UART1_RTS/PG_EINT8)
33	PA9 (SIM_RST/PA_EINT9)	34	*GND*
35	PA10 (SIM_DET/PA_EINT10)	36	PG9 (UART1_CTS/PG_EINT9)
37	PA20 (PCM0_DOUT/SIM_VPPEN/PA_EINT20)	38	PG6 (UART1_TX/PG_EINT6)
39	*GND*	40	PG7 (UART1_RX/PG_EINT7)

3.4 Methods For Classification

In order to apply the various data processing steps defined on the previous chapter to the data collected, a series of scripts were developed in Python. One that contains the features extraction techniques implemented and a second script that as 8 clones for each different odour, in which the data is visually represented and normalized if desired, samples are assembled together in an odour specific matrix, features are extracted and added to a new file labelled accordingly. There is also one script that is responsible for the feature ranking algorithm implemented and one for dimensionality reduction application to the newly extracted datasets. It is also prepared to build 2D or 3D visualizations of subspace vectors that result from either PCA or LDA and it may also store projected data on one of the referred subspaces in order to feed the classifiers afterwards. At last, in the sixth one, classifiers are built and introduced to the datasets and their performance is evaluated.

The series of scripts developed during this thesis are presented in Appendix D.

3.5 Visualization

Although the visualization objective of this thesis was not implemented, a projected view of a potential application is shown in Figure 3.3.

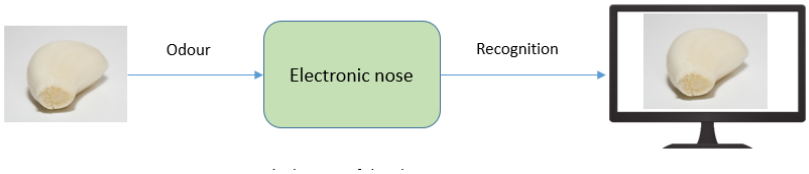


Figure 3.3: Scheme of a intended electronic nose application. A garlic clove is used as an example.

The envisioned objective was a deployment of the electronic nose assembled with a fully trained recognition model that would be doing consequent timed air scans 24/7 and would be capable of detecting the odour set considered and communicating that information in a monitor, either via a HDMI cable or via a SSH wireless connection, allowing the localization of recognition display to be flexible.

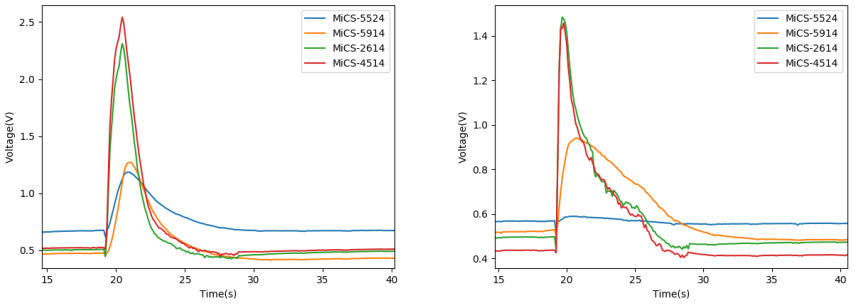
Chapter 4

Test and Results

In this section the results obtained applying the previous described methodology are shown. Visual data exploration of the samples collected is presented in 4.1, feature ranking results with normalized 0 to 1 data are shown in 4.2, dimensionality reduction techniques employed on the dataset and their results are displayed in 4.3 and classifiers are evaluated in 4.4.

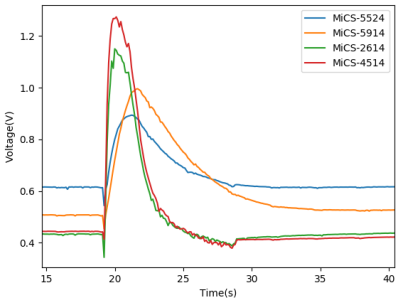
4.1 Data Exploration

As mentioned in the previous chapter, 25 samples were collected from each of the 8 odours amounting to a 200 samples dataset. A sample of sensor responses for each one of the odours in the experiment is shown in Figure 4.1. From a visual analysis of the graphs, it is somewhat clear that the responses of the four sensors for each odour are quite distinct which, in turn, is encouraging to the application of the following data processing steps. There is clearly a similarity of sensor responses between lemon and bay leaf samples. MiCS-5524 response might be the main difference between those two odours, considering that its conductivity increases at a higher rate when exposed to bay leaf even if this odour promotes lower response in magnitude in comparison with the lemon sample. In resemblance, gasoline and naphthalene provoke similar response traits in the sensor array and it can be argued that MiCS-5524 makes the difference, although the gasoline response shows higher magnitude and that factor can be reason for that difference. Ethanol and garlic responses both share similar MiCS-2614, MiCS-4514 and even MiCS-5914 responses however, MiCS-5524 does seem to differentiate them. Both ammonia and vinegar appear to have some similarities but their response profiles are still moderately distinct.

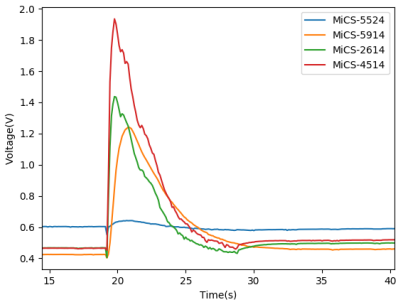


(a) Ethanol

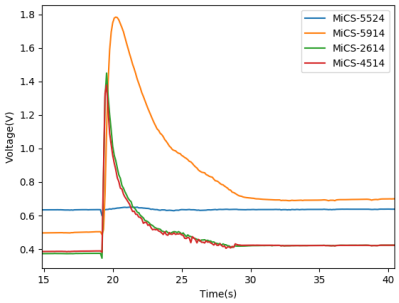
(b) Garlic



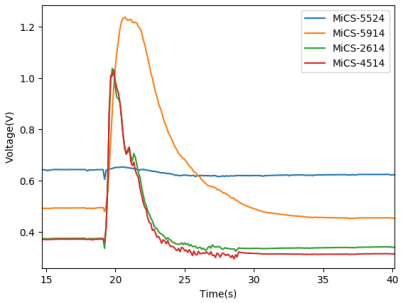
(c) Ammonia



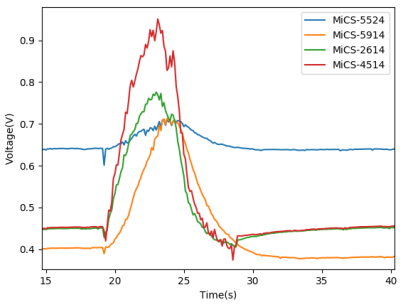
(d) Gasoline



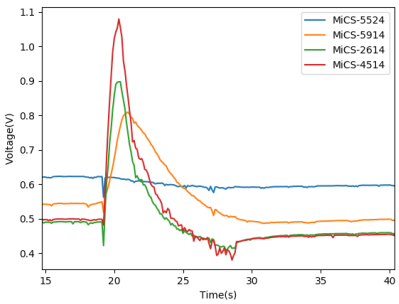
(e) Lemon



(f) Bay leaf



(g) Vinegar



(h) Naphthalene

Figure 4.1: Examples of sensor responses for each of the odours: a) Ethanol; b) Garlic; c) Ammonia; d) Gasoline; e) Lemon; f) Bay leaf; g) Vinegar and h) Naphthalene.

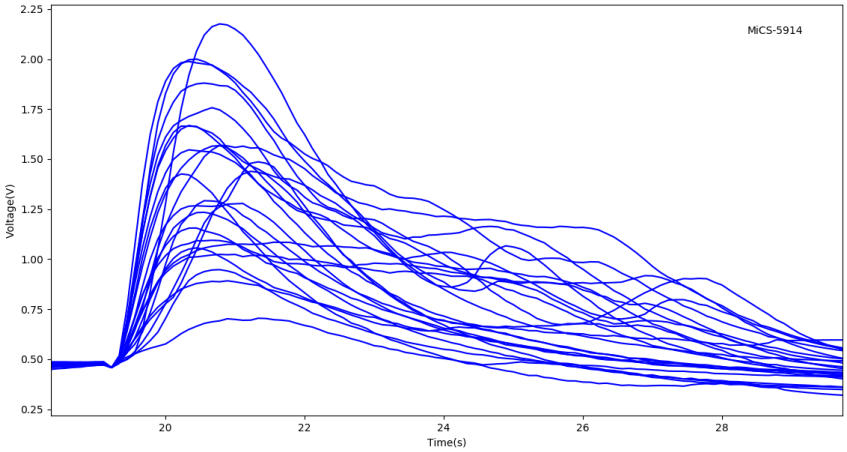


Figure 4.2: Response curves of MiCS-5914 to ethanol exposure (25 responses).

So far, one sample of each odour was presented but that information does not really provide any information about the variability inside the group of samples from the same odour. In a attempt to obtain some insight about this issue, all 25 responses of MiCS-5914 in exposure to ethanol are normalized around a time stamp where the sensor starts to react, is displayed in Figure 4.2. From a initial analysis standpoint, it is safe to say that the magnitude of the responses oscillates reasonably across the sample spectre which in turn points out to sampling method intrinsic issues, namely the absence of control in the force applied during the initial bottle squeeze. When to stop it and the rate in which the bottle is uncompressed might unravel other types of variability in the collection of responses shown like late peak responses and soft secondary peaks during desorption phase. Although this might be perceived as systematic error, variability of the dataset is crucial to create a robust dataset capable of containing reasonable discriminative power and, as a result, help preventing classification overfitting.

4.2 Feature Ranking

In the following subsection, the results of the methodology chosen for feature ranking is described. It is worth noting that the dataset was already divided in two groups, as mentioned in subsection 2.2.6: a training group encapsulating 75% of the data and a second group, a testing set with the remaining 25%.

4.2.1 Recursive Feature Elimination

A ranking of the different features was created applying this algorithm and the results for the 10 most discriminatory features are shown in Table 4.1. The reason for selecting the first ten and not a smaller number is linked to the fact that the main objective for this algorithm application was to try to understand to which extent each sensor contributes to the differentiation of the odours considered.

Table 4.1: Ten most discriminative features for 0 to 1 normalized data according to recursive feature elimination.

Feature Ranking	Feature Abbreviation	Corresponding sensor
1	m3	MiCS-2614
2	dm2	MiCS-5914
3	um4	MiCS-4514
4	udm1	MiCS-5524
5	udm3	MiCS-2614
6	wv410	MiCS-4514
7	wv211	MiCS-5914
8	ddm2	MiCS-5914
9	dvm4	MiCS-4514
10	wv29	MiCS-5914

All four different sensors are represented in this ranking with more than one feature each which implies that all sensors have a substantial contribution in the discrimination of the odours they were exposed to. From the number of features in the first ten, sensor MiCS-5914 seems to be the most dominant one with four features out of ten. A more individual sensor focused research with information like this ranking taken into consideration may improve the overall results of similar experiments. Feature removal with basis on this ranking was not performed on the dataset although it might be a reasonable enhancement in classifications tasks that come ahead.

4.3 Feature Transformation

Dimensionality reduction techniques are applied in the following subsection and their results presented.

4.3.1 Dimensionality reduction

Normalization of the data vastly improves PCA results in most cases. Since PCA is a variance maximizing methodology, having all the features in the same scale increases the reliability of the projection of the data onto directions that maximize the variance. From the explained variances of the first five principal components presented in Table 4.2 it can be inferred that the dimensionality reduction algorithm was able to condense the majority of the dataset information in the first five components. The most discriminative component is the first from logarithmic scaled data with 89.7% of the variance explained followed by 74.5 % in min-max normalization. It is also worth mentioning that the three data statuses contain more then 95.0% of cumulative variance in the first five components.

Table 4.2: First five principal components explained variances for each of the feature normalization status of the data.

Data normalization	Component	Explained Variance	Cumulative Variance
Raw	PC 1	0.526	0.526
	PC 2	0.236	0.762
	PC 3	0.098	0.860
	PC 4	0.071	0.931
	PC 5	0.034	0.965
Scaled 0 to 1	PC 1	0.745	0.745
	PC 2	0.157	0.902
	PC 3	0.042	0.944
	PC 4	0.022	0.966
	PC 5	0.016	0.982
Log scaled	PC 1	0.897	0.897
	PC 2	0.089	0.986
	PC 3	0.007	0.993
	PC 4	0.006	0.999
	PC 5	0.000	1.000

Due to linear discriminant analysis intricate mathematics operations, normalization of the data prior to the analysis does not take any toll in the final outcome of its discriminants are shown in Table 4.3. The cumulative variance of the first five linear discriminants is 99.8% which is similar to the previous results of the PCA counterpart.

Table 4.3: First five principal components explained variances for each of the feature scaling status of the data.

Discriminant	Explained Variance	Cumulative Variance
LD 1	0.573	0.573
LD 2	0.208	0.781
LD 3	0.127	0.908
LD 4	0.064	0.972
LD 5	0.026	0.998

PCA components projected in 2D and 3D in Figure 4.3 show that the classes of odours including in the experiment are somewhat distinctive aside from a specific group that overlaps quite a lot. Ethanol, gasoline and ammonia are perceived as discriminable but there is a clear overlap between garlic, bay leaf, vinegar, lemon and naphthalene to a certain extent.

On the other hand, LDA discriminants projection that can be consulted in Figure 4.4 shows better discriminative performance. Ethanol, gasoline and ammonia have a distinct position in the projection as well as vinegar and naphthalene to a lesser extent. Garlic and bay leaf overlap in several samples and lemon ones are presented really close to this pair of odours.

Concluding, the results seem to indicate that, for the collected dataset, the classes seem to be more separable using a linear discriminant transformation than its principal component counterpart, which is expected considering this algorithm maximizes the distance inter-class. Perhaps a more adequate feature extraction or a more robust number of samples can dictate a different outcome.

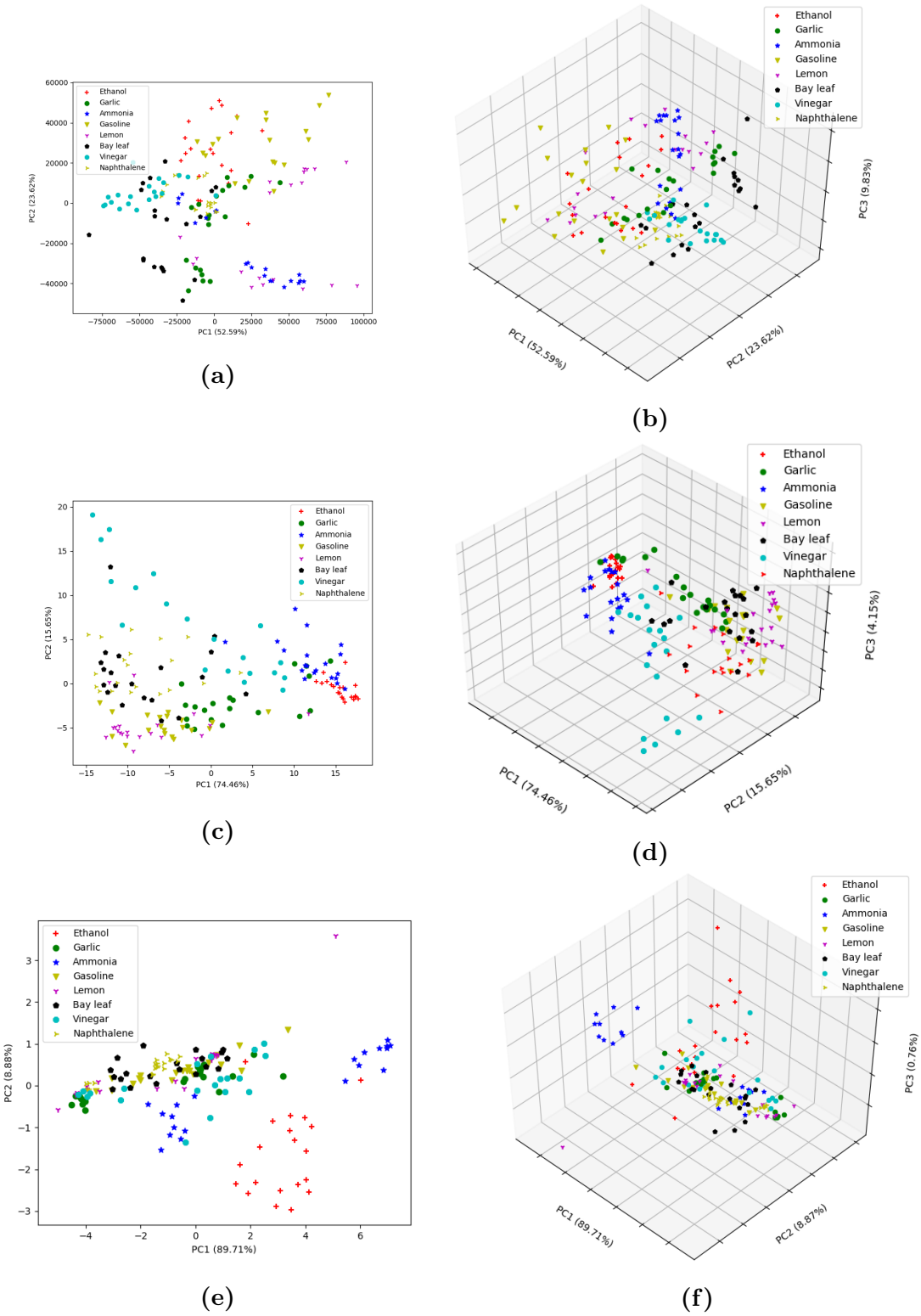
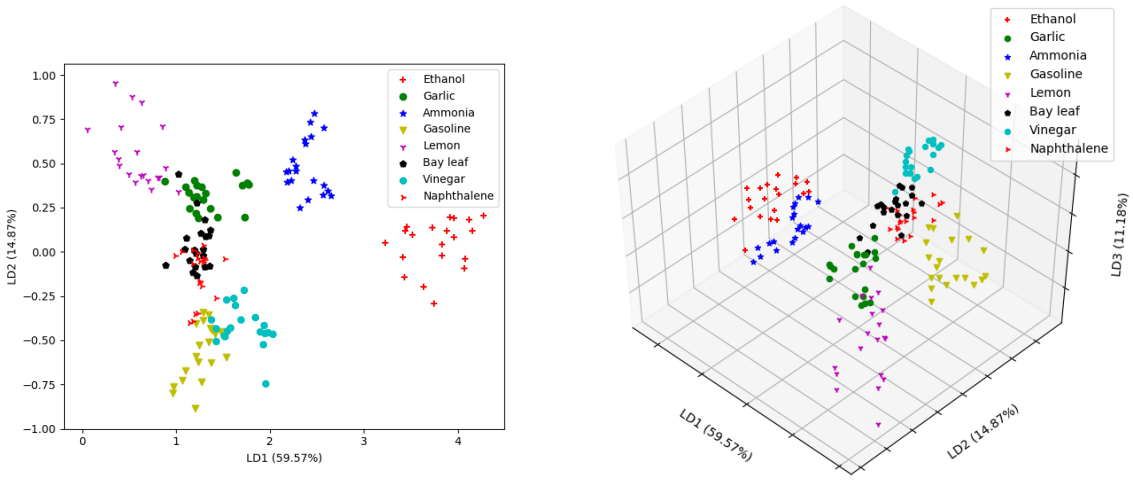


Figure 4.3: 2D and 3D PCA discrimination of the odours dataset collected with different data normalization procedures: raw data a)2D b)3D; 0 to 1 normalized data c)2D d)3D and logarithmic normalized data e)2D f)3D.



(a) (b)

Figure 4.4: 2D and 3D LDA discrimination of the odours dataset collected.

4.4 Classification

Different supervised models were built to predict the classes of the testing data assembled. For each classifier, three different sets of data were used: raw data, first five principal components projected data and first five linear discriminant analysis projected data. Within each set, three different normalization procedures were applied a priori. The results obtained for each model are presented in the following chapters.

4.4.1 Support Vector Machine

After algorithmic tuning trying several combinations of hyper-parameters, the group of hyper-parameters for the support vector machine classifier applied are as follows: linear *kernel*, penalty parameter C is 1 and the *kernel* coefficient gamma is 0.001. Surprisingly, as can be seen in 4.4, the most accurate classification results come from both raw data analysis and raw data projected in linear discriminant subspace with 92.50% of accuracy. Data normalized 0 to 1 projected in principal components subspace achieved 85.00%. Results from logarithmic normalization are overall poor with just 34.90% accuracy with non projected and data projected in the principal component subspace.

Table 4.4: SVM results with different data processing techniques.

Data projection	Data normalization	Accuracy(%)
Raw	Raw	92.50
	Scaled 0 to 1	75.00
	Log scaled	34.90
PCA	Raw	85.00
	Scaled 0 to 1	62.50
	Log scaled	34.90
LDA	Raw	92.50

4.4.2 Backpropagation Neural Network

Regarding hyper-parameter tuning, several combinations were tested and overfitting was taking into consideration to avoid loss of discriminative capability from the model. The parameters chosen are the following: two hidden network layers with 15 neurons on the inner and 7 on the outer layer; *tanh* as activation function; *adam* as the learning methodology; *alpha* as 0.00001 and 10000 as the maximum of iterations.

Similarly to what has been shown in SVM results, Figure 4.5 describes the best results of this classifier that correspond to LDA projected data without normalization with 97.50% of accuracy and 82.50% from both raw data projected with PCA and non projected data with min-max normalization. PCA projected data with logarithmic normalization reaches 80.00% and 67.50% with data scaled from 0 to 1. Data without any type of transformation reaches 40.00% of accuracy with this algorithm.

Table 4.5: BPNN results with different data processing techniques.

Data projection	Data normalization	Accuracy(%)
Raw	Raw	40.00
	Scaled 0 to 1	82.50
	Log scaled	12.50
PCA	Raw	82.50
	Scaled 0 to 1	67.50
	Log scaled	80.00
LDA	Raw	97.50

From the two best accuracy results of both classifiers, Table 4.6 describe how each odour was classified and which misclassification occurred. As can be seen, this

Table 4.6: Classification results of the raw test samples LDA projected with a BPNN as classifier with 97.50% accuracy.

	Samples	Classified as							
		Ethanol	Garlic	Ammonia	Gasoline	Lemon	Bay leaf	Vinegar	Naphthalene
Ethanol	5/5	5/5							
Garlic	5/5		5/5						
Ammonia	5/5			5/5					
Gasoline	5/5				5/5				
Lemon	5/5					5/5			
Bay leaf	5/5						5/5		
Vinegar	5/5							4/5	1/0
Naphthalene	5/5								6/5

misclassification was a vinegar sample that was classified as a naphthalene one. This error happened for both instances but the misclassified sample was different.

The second overall best accuracy is a tie between two different SVM classifier results: data not normalized nor projected and raw data projected in linear discriminants subspace. Both achieved an accuracy of 92.50%. Although both present the same accuracy, the misclassified samples are different but one, a vinegar sample that is classified wrongly as a naphthalene one. This same sample is misclassified as well in the best BPNN result shown in Table 4.6. SVM classifier best classification details are present in Tables 4.7 and 4.8. Aside from the mentioned vinegar sample, in the former table, one lemon sample is misclassified as garlic and one naphthalene one is wrongly classified as bay leaf. On the latter table, a gas sample is misclassified as garlic and a bay leaf sample is considered a naphthalene one wrongly.

Table 4.7: Classification results of the test samples without normalization and projection with a SVM as classifier with 92.50% accuracy.

	Samples	Classified as							
		Ethanol	Garlic	Ammonia	Gasoline	Lemon	Bay leaf	Vinegar	Naphthalene
Ethanol	5/5	5/5							
Garlic	5/5		6/5						
Ammonia	5/5			5/5					
Gasoline	5/5				5/5				
Lemon	5/5		1/0			4/5			
Bay leaf	5/5						6/5		
Vinegar	5/5							4/5	1/0
Naphthalene	5/5						1/0		5/5

Table 4.8: Classification results of the test samples without normalization and projected in linear discriminants subspace with a SVM as classifier with 92.50% accuracy.

	Samples	Classified as							
		Ethanol	Garlic	Ammonia	Gasoline	Lemon	Bay leaf	Vinegar	Naphthalene
Ethanol	5/5	5/5							
Garlic	5/5		6/5						
Ammonia	5/5			5/5					
Gasoline	5/5		1/0		4/5				
Lemon	5/5					5/5			
Bay leaf	5/5						4/5		1/0
Vinegar	5/5							4/5	1/0
Naphthalene	5/5								7/5

Chapter 5

Conclusions

In this final chapter, discussion of the work done in this thesis is presented, with focus on what particularly was a success and, more importantly, what failed so in the future researchers do not make the same mistakes and build up on what was achieved. Discussion about the experiment takes place in section 5.1 and future work and ideas are referenced in section 5.2.

5.1 Discussion

Nowadays there is an urgent need of personal care devices that are useful to the daily life as well as affordable and robust. The experimental setup developed in this thesis is an attempt in achieving a reasonable proof of concept and with further more controlled experiments, it may be possible to aspire to fully imitate a human nose condition and decision wise and perhaps surpass its capabilities. The platform developed to collect, treat and analyse data is modular and can be modified for different experiments with other sensors and technologies. The data sampling procedure in this experiment was not controlled as it could be. Temperature and humidity were not supervised and this type of variations may affect the experiment results considering this condition promotes different behaviours from the sensors exposed. A soft squeeze in order to facilitate the air flow from the odour containers headspace to the gas chamber is not a controlled mechanism which in turn produced variability that a more restrained procedure would not. However, this variability enriches the dataset collected and counterbalances to some extent the low amount of samples obtained. Baseline manipulation of the samples collected was not conducted because comparatively speaking the outcome of that action would not differ much from the

raw status of the data and from the other samples as well. The features extracted in this experiment may be suitable in part for classification considering the reasonable results presented in the previous chapter. A good result was achieved on both classifiers with linear discriminant analysis projected data and that may be an argument in favour of a good feature extraction methodology although logarithmic scaling of the data in its raw state and its projections in PCA subspace achieved uninspiring results except on the neural network with PCA projection data. Is it also worth noting that features like the sample mean and derivatives mean for the whole sample might have had a negative impact in the information extraction for the dataset. One of the constraints of this experiment is the small number of samples collected and analysed. With a quantity so small it is difficult to make a conclusion even about good classification results. Overfitting of the data is a common phenomenon in similar classification tasks and the only way to dampen its influence is to collect more data and apply an adequate cross-validation strategy. The fact that sample length is exceedingly vast, it is safe to say that significant information present on the sensor responses might have been diluted.

5.2 Future Work

In the following subsections, several areas in which this experiment can improve are exposed.

5.2.1 Sensor Drifting

For future works and an eventual deployment or development of a portable electronic nose, sensor drifting should be taken seriously. After long periods of time, sensor projected characteristics start to change due to large amounts of time of exposure to the most varied chemical compounds. This exposure enables the deterioration of sensor coatings and their sensitivity drifts from the projected initially. Among the possibilities to tackle this problem, software workarounds usually are applied to try to soften this phenomenon.

5.2.2 Sensor Load Resistance

In the specific case of the electronic nose developed, changing load resistance values can enable the sensor array to detect different layers of information that may be

valuable in the presence of certain odours.

5.2.3 Odour Chemical Composition and Concentration

Controlling or having a deeper understanding about the chemical compounds present in an odour and their threshold for the human olfaction is a clear future enhancement in this type of analysis. With this chemical information, it should be possible to evaluate different types of certain odour enablers and even make inferences about the living stage of certain organic items like plants, fruits or bakery products. Having this type of information may be crucial to identify and discriminate odour mixtures.

5.2.4 Different Array of Sensors

A different array of sensors with distinct target sensitivities might increase the sensing capability and the overall scope of the device. Not only metal oxide gas sensors but joining different technologies like polymer based sensor will broaden the target odour space.

5.2.5 Larger Amount of Odours

Applying different odours from the most variate sources to this sensing chamber is one of the next steps to improve this technology. With a larger odour space we should be able to be nearer a fully functional portable electronic nose. Once an electronic nose is fully trained to recognize daily life smells, it could then detect unusual odours related to forgotten disposables, burning food, expired refrigerator contents, or even the individual's personal hygiene.

5.2.6 Integration with smart phone

One of the partial objectives of this thesis was the integration of the electronic nose with a mobile application that could be accessible in a smart phone by anyone. In this scenario, an electronic nose would either be placed or carried around and the monitoring of odours could be performed and communicated 24/7.

Bibliography

[1] K Mori. Grouping of odorant receptors: odour maps in the mammalian olfactory bulb, 2003.

[2] Yoshihito Niimura. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Human genomics*, 4(2):107, 2009.

[3] Andreas FP Temmel, Christian Quint, Bettina Schickinger-Fischer, Ludger Klimek, Elisabeth Stoller, and Thomas Hummel. Characteristics of olfactory disorders in relation to major causes of olfactory loss. *Archives of Otolaryngology–Head & Neck Surgery*, 128(6):635–641, 2002.

[4] H Tennen, G Affleck, R Mendola, and TV Getchell. Coping with smell and taste disorders. *Smell and taste in health and disease*, pages 787–802, 1991.

[5] Takaki Miwa, Mitsuru Furukawa, Toshiaki Tsukatani, Richard M Costanzo, Laurence J DiNardo, and Evan R Reiter. Impact of olfactory impairment on quality of life and disability. *Archives of Otolaryngology–Head & Neck Surgery*, 127(5):497–503, 2001.

[6] Steven Nordin, Ebba Hedén Blomqvist, Petter Olsson, Pär Stjärne, Anders Ehnhage, and NAF2S2 Study Group. Effects of smell loss on daily life and adopted coping strategies in patients with nasal polyposis with asthma. *Acta oto-laryngologica*, 131(8):826–832, 2011.

[7] Ann M Ferris and Valerie B Duffy. Effect of olfactory deficits on nutritional status: does age predict persons at risk? *Annals of the New York Academy of Sciences*, 561(1):113–123, 1989.

[8] Annika Brämerson, Steven Nordin, and Mats Bende. Clinical experience with patients with olfactory complaints, and their quality of life. *Acta otolaryngologica*, 127(2):167–174, 2007.

[9] Ebba Hedén Blomqvist, A Bramerson, P Stjarne, and Steven Nordin. Consequences of olfactory loss and adopted coping strategies. *Rhinology*, 42(4):189–194, 2004.

[10] Katja Aschenbrenner, Cornelia Hummel, Katja Teszmer, Franziska Krone, Tadashi Ishimaru, Han-Seok Seo, and Thomas Hummel. The influence of olfactory loss on dietary behaviors. *The Laryngoscope*, 118(1):135–144, 2008.

[11] Bamini Gopinath, Kaarin J Anstey, Carolyn M Sue, Annette Kifley, and Paul Mitchell. Olfactory impairment in older adults is associated with depressive symptoms and poorer quality of life scores. *The American Journal of Geriatric Psychiatry*, 19(9):830–834, 2011.

[12] Ilona Croy, Steven Nordin, and Thomas Hummel. Olfactory disorders and quality of life—an updated review. *Chemical Senses*, 39(3):185–194, 2014.

[13] Adam Kepecs, Naoshige Uchida, and Zachary F Mainen. The sniff as a unit of olfactory processing. *Chemical senses*, 31(2):167–179, 2005.

[14] Matt Wachowiak. All in a sniff: olfaction as a model for active sensing. *Neuron*, 71(6):962–973, 2011.

[15] Thomas Hummel, Basile N Landis, and Karl-Bernd Hüttenbrink. Smell and taste disorders. *GMS current topics in otorhinolaryngology, head and neck surgery*, 10, 2011.

[16] Mechtild M Vennemann, Thomas Hummel, and Klaus Berger. The association between smoking and smell and taste impairment in the general population. *Journal of neurology*, 255(8):1121–1126, 2008.

[17] Woo Hyun Lee, Jee Hye Wee, Dong-Kyu Kim, Chae-Seo Rhee, Chul Hee Lee, Soyeon Ahn, Ju Hyun Lee, Yang-Sun Cho, Kun Hee Lee, Kyung Soo Kim, et al. Prevalence of subjective olfactory dysfunction and its risk factors: Korean national health and nutrition examination survey. *PloS one*, 8(5):e62725, 2013.

[18] Janina Seubert, Erika J Laukka, Debora Rizzuto, Thomas Hummel, Laura Fratiglioni, Lars Bäckman, and Maria Larsson. Prevalence and correlates of olfactory dysfunction in old age: A population-based study. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, page glx054, 2017.

[19] Shristi Rawal, Howard J Hoffman, Kathleen E Bainbridge, Tania B Huedo-Medina, and Valerie B Duffy. Prevalence and risk factors of self-reported smell

and taste alterations: results from the 2011–2012 us national health and nutrition examination survey (nhanes). *Chemical senses*, 41(1):69–76, 2015.

[20] Denis Lafreniere and Norman Mann. Anosmia: loss of smell in the elderly. *Otolaryngologic Clinics of North America*, 42(1):123–131, 2009.

[21] Steven Nordin, Andreas U Monsch, and Claire Murphy. Unawareness of smell loss in normal aging and alzheimer's disease: discrepancy between self-reported and diagnosed smell sensitivity. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(4):P187–P192, 1995.

[22] Claire Murphy, Carla R Schubert, Karen J Cruickshanks, Barbara EK Klein, Ronald Klein, and David M Nondahl. Prevalence of olfactory impairment in older adults. *Jama*, 288(18):2307–2312, 2002.

[23] Chih-Hung Shu, PO Lee, Ming-Ying Lan, and Yi-Lun Lee. Factors affecting the impact of olfactory loss on the quality of life and emotional coping ability. *Rhinology*, 49(3):337–341, 2011.

[24] Steven Nordin, Annika Brämerson, and Mats Bende. Prevalence of self-reported poor odor detection sensitivity: the skövde population-based study. *Acta otolaryngologica*, 124(10):1171–1173, 2004.

[25] Luciano Lobato Gregorio, Fábio Caparroz, Leonardo Mendes Acatauassú Nunes, Luciano Rodrigues Neves, and Eduardo Kosugi Macoto. Olfaction disorders: retrospective study. *Brazilian journal of otorhinolaryngology*, 80(1):11–17, 2014.

[26] T Metin Önerci. *Nasal physiology and pathophysiology of nasal disorders*. Springer, 2013.

[27] D Nunes, E Janz, and A Naumann. An atypical etiology of anosmia. *J Otolaryngol ENT Res*, 6(1):00146, 2017.

[28] Steven Nordin and Annika Brämerson. Complaints of olfactory disorders: epidemiology, assessment and clinical implications. *Current opinion in allergy and clinical immunology*, 8(1):10–15, 2008.

[29] Yong-ming Zou, Da Lu, Li-ping Liu, Hui-hong Zhang, and Yu-ying Zhou. Olfactory dysfunction in alzheimer's disease. *Neuropsychiatric disease and treatment*, 12:869, 2016.

[30] Yuriko Hori, Osamu Matsuda, and Sachiko Ichikawa. Olfactory function in elderly people and patients with alzheimer's disease. *Psychogeriatrics*, 15(3):179–185, 2015.

[31] Jens Reden, Antje Mueller, Christian Mueller, Iordanis Konstantinidis, Johannes Frasnelli, Basile N Landis, and Thomas Hummel. Recovery of olfactory function following closed head injury or infections of the upper respiratory tract. *Archives of Otolaryngology–Head & Neck Surgery*, 132(3):265–269, 2006.

[32] AM Seiden, HJ Duncan, and DV Smith. Office management of taste and smell disorders. *Otolaryngologic clinics of North America*, 25(4):817–835, 1992.

[33] Ilona Croy, Simona Negoias, Lenka Novakova, Basile N Landis, and Thomas Hummel. Learning about the functions of the olfactory system from people without a sense of smell. *PLoS One*, 7(3):e33365, 2012.

[34] V Bojanowski, T Hummel, and I Croy. Isolated congenital anosmia–clinical and daily life aspects of a life without a sense of smell. *Laryngo-rhino-otologie*, 92(1):30–33, 2013.

[35] J Frasnelli, BN Landis, S Heilmann, B Hauswald, KB Hüttenbrink, JS Lacroix, DA Leopold, and T Hummel. Clinical presentation of qualitative olfactory dysfunction. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 261(7):411–415, 2004.

[36] B N Landis, C G Konnerth, and T Hummel. A study on the frequency of olfactory dysfunction. *The Laryngoscope*, 114(10):1764–1769, 2004.

[37] Steven Nordin, Annika Brämerson, Eva Millqvist, and Mats Bende. Prevalence of parosmia: the skövde population-based studies. *Rhinology*, 45(1):50–53, 2007.

[38] Daniel A Deems, Richard L Doty, R Gregg Settle, Victoria Moore-Gillon, Paul Shaman, Andrew F Mester, Charles P Kimmelman, Vernon J Brightman, and James B Snow. Smell and taste disorders, a study of 750 patients from the university of pennsylvania smell and taste center. *Archives of otolaryngology–head & neck surgery*, 117(5):519–528, 1991.

[39] Steven Nordin, Claire Murphy, Terence M Davidson, Carlo Quiñonez, Alfredo A Jalowayski, and Dennard W Ellison. Prevalence and assessment of qualitative olfactory dysfunction in different age groups. *The Laryngoscope*, 106(6):739–744, 1996.

[40] P Faulcon, F Portier, B Biacabe, and P Bonfils. Anosmie secondaire à une rhinite aiguë: sémiologie et évolution à propos d'une série de 118 patients. In *Annales d'oto-laryngologie et de chirurgie cervico-faciale*, volume 116, pages 351–357. Elsevier Masson, 1999.

[41] C Quint, AF Temmel, B Schickinger, S Pabinger, P Ramberger, and T Hummel. Patterns of non-conductive olfactory disorders in eastern austria: a study of 120 patients from the department of otorhinolaryngology at the university of vienna. *Wiener klinische Wochenschrift*, 113(1-2):52–57, 2001.

[42] Anthony PF Turner and Naresh Magan. Electronic noses and disease diagnostics. *Nature reviews. Microbiology*, 2(2):161, 2004.

[43] Victor E Henrich and Paul Anthony Cox. *The surface science of metal oxides*. Cambridge university press, 1996.

[44] Chengxiang Wang, Longwei Yin, Luyuan Zhang, Dong Xiang, and Rui Gao. Metal oxide gas sensors: sensitivity and influencing factors. *Sensors*, 10(3):2088–2106, 2010.

[45] Ghenadii Korotcenkov. Metal oxides for solid-state gas sensors: What determines our choice? *Materials Science and Engineering: B*, 139(1):1–23, 2007.

[46] Mos type sensors operating principle. `http://www.figaro.co.jp/en/technicalinfo/principle/mos-type.html`. Accessed: 2017-07-27.

[47] K Arshak, E Moore, GM Lyons, J Harris, and S Clifford. A review of gas sensors employed in electronic nose applications. *Sensor review*, 24(2):181–198, 2004.

[48] J Watson. The stannic oxide gas sensor. *Sensor Review*, 14(1):20–23, 1994.

[49] N Barsan, D Koziej, and U Weimar. Metal oxide-based gas sensor research: How to? *Sensors and Actuators B: Chemical*, 121(1):18–35, 2007.

[50] Common types of chemical that cause health risks. `http://old.iupac.org/publications/cd/essential_toxicology/IUPACTOX7.pdf`. Accessed: 2017-08-10.

[51] Amalia Berna. Metal oxide sensors for electronic noses and their application to food analysis. *Sensors*, 10(4):3882–3910, 2010.

[52] Amy Loutfi, Silvia Coradeschi, Ganesh Kumar Mani, Prabakaran Shankar, and John Bosco Balaguru Rayappan. Electronic noses for food quality: A review. *Journal of Food Engineering*, 144:103–111, 2015.

[53] Corrado Di Natale, Alessandro Mantini, Antonella Macagnano, Daniela Antuzzi, Roberto Paolesse, and Arnaldo D'Amico. Electronic nose analysis of urine samples containing blood. *Physiological Measurement*, 20(4):377, 1999.

[54] Alexandros K Pavlou, Naresh Magan, Cliodna McNulty, Jeff Meecham Jones, Dorothy Sharp, Jonathon Brown, and Anthony PF Turner. Use of an electronic nose system for diagnoses of urinary tract infections. *Biosensors and Bioelectronics*, 17(10):893–899, 2002.

[55] Wang Ping, Tan Yi, Xie Haibao, and Shen Farong. A novel method for diabetes diagnosis based on electronic nose. *Biosensors and Bioelectronics*, 12(9):1031–1036, 1997.

[56] Ehab I Mohamed, R Linder, G Perriello, N Di Daniele, SJ Pöppl, and A De Lorenzo. Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis. *Diabetes, nutrition & metabolism*, 15(4):215–221, 2002.

[57] Silvano Dragonieri, Robert Schot, Bart JA Mertens, Saskia Le Cessie, Stefanie A Gauw, Antonio Spanevello, Onofrio Resta, Nico P Willard, Teunis J Vink, Klaus F Rabe, et al. An electronic nose in the discrimination of patients with asthma and controls. *Journal of Allergy and Clinical Immunology*, 120(4):856–862, 2007.

[58] Rossella Blatt, Andrea Bonarini, Elisa Calabro, Matteo Della Torre, Matteo Matteucci, and Ugo Pastorino. Lung cancer identification by an electronic nose based on an array of mos sensors. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 1423–1428. IEEE, 2007.

[59] Corrado Di Natale, Antonella Macagnano, Eugenio Martinelli, Roberto Paolesse, Giuseppe D'Arcangelo, Claudio Roscioni, Alessandro Finazzi-Agrò, and Arnaldo D'Amico. Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosensors and Bioelectronics*, 18(10):1209–1218, 2003.

[60] Roberto F Machado, Daniel Laskowski, Olivia Deffenderfer, Timothy Burch, Shuo Zheng, Peter J Mazzone, Tarek Mekhail, Constance Jennings, James K Stoller, Jacqueline Pyle, et al. Detection of lung cancer by sensor array analyses

of exhaled breath. *American journal of respiratory and critical care medicine*, 171(11):1286–1291, 2005.

[61] M Bernabei, G Pennazza, M Santonico, C Corsi, C Roscioni, R Paolesse, C Di Natale, and A D'Amico. A preliminary study on the possibility to diagnose urinary tract cancers by an electronic nose. *Sensors and Actuators B: Chemical*, 131(1):1–4, 2008.

[62] Antti Roine, Erik Veskimäe, Antti Tuokko, Pekka Kumpulainen, Juha Koskimäki, Tuomo A Keinänen, Merja R Häkkinen, Jouko Vepsäläinen, Timo Paavonen, Jukka Lekkala, et al. Detection of prostate cancer by an electronic nose: a proof of principle study. *The Journal of urology*, 192(1):230–235, 2014.

[63] Tim G Meij, Ilhame Ben Larbi, Marc P Schee, Yvette E Lentferink, Tamara Paff, Jochim S Terhaar sive Droste, Chris J Mulder, Adriaan A Bodegraven, and Nanne K Boer. Electronic nose can discriminate colorectal carcinoma and advanced adenomas by fecal volatile biomarker analysis: proof of principle study. *International journal of cancer*, 134(5):1132–1138, 2014.

[64] Jessica E Fitzgerald, Eric TH Bui, Naomi M Simon, and Hicham Fenniri. Artificial nose technology: Status and prospects in diagnostics. *Trends in biotechnology*, 35(1):33–42, 2017.

[65] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[66] Xiuzhen Guo, Chao Peng, Songlin Zhang, Jia Yan, Shukai Duan, Lidan Wang, Pengfei Jia, and Fengchun Tian. A novel feature extraction approach using window function capturing and qpso-svm for enhancing electronic nose performance. *Sensors*, 15(7):15198–15217, 2015.

[67] Jia Yan, Fengchun Tian, Qinghua He, Yue Shen, Shan Xu, Jingwei Feng, and Kadri Chaibou. Feature extraction from sensor data for detection of wound pathogen based on electronic nose. *Sensors and Materials*, 24(2):57–73, 2012.

[68] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[69] Lindsay I Smith et al. A tutorial on principal components analysis. *Cornell University, USA*, 51(52):65, 2002.

[70] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18, 1998.

[71] F Davide, M Holmberg, and I Lundström. Virtual olfactory interfaces: electronic noses and olfactory, 2001.

[72] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[73] Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[74] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Appendix A

Sensors in Electronic Noses

Table A refers to various type of sensors and their detection principles incorporated in different electronic noses works.

Sensor type	Sensitive material	Detection Principle
Metal oxide semiconducting (MOS)	Doped semiconducting metal oxides (Sn_2, WO_3, GaO)	Resistance change
Electrochemical	Solid or liquid electrolytes	Current or voltage change
Conducting polymer	Modified conducting polymers	Resistance change
Acoustic: Quartz crystal microbalance (QMB); surface & bulk acoustic wave (SAW,BAW)	organic/inorganic film layers	Mass change (frequency shift)
Calorimetric; catalytic bead (CB)	Pellistor	Temperature or heat change (chemical reaction)
Catalytic field-effect (MOSFET)	Catalytic metals	Electric field change
Colorimetric	Organic dyes	Colour changes, absorbency
Fluorescence	Fluorescence-sensitive detector	Fluorescence-light emissions

Infra-red	IR-sensitive detector	Infra-red radiation absorption
Electro-optical	Photodiode, light sensitive	Light modulation, optical changes

Appendix B

Printed Circuits Boards

In this appendix PCB circuit designs are presented.

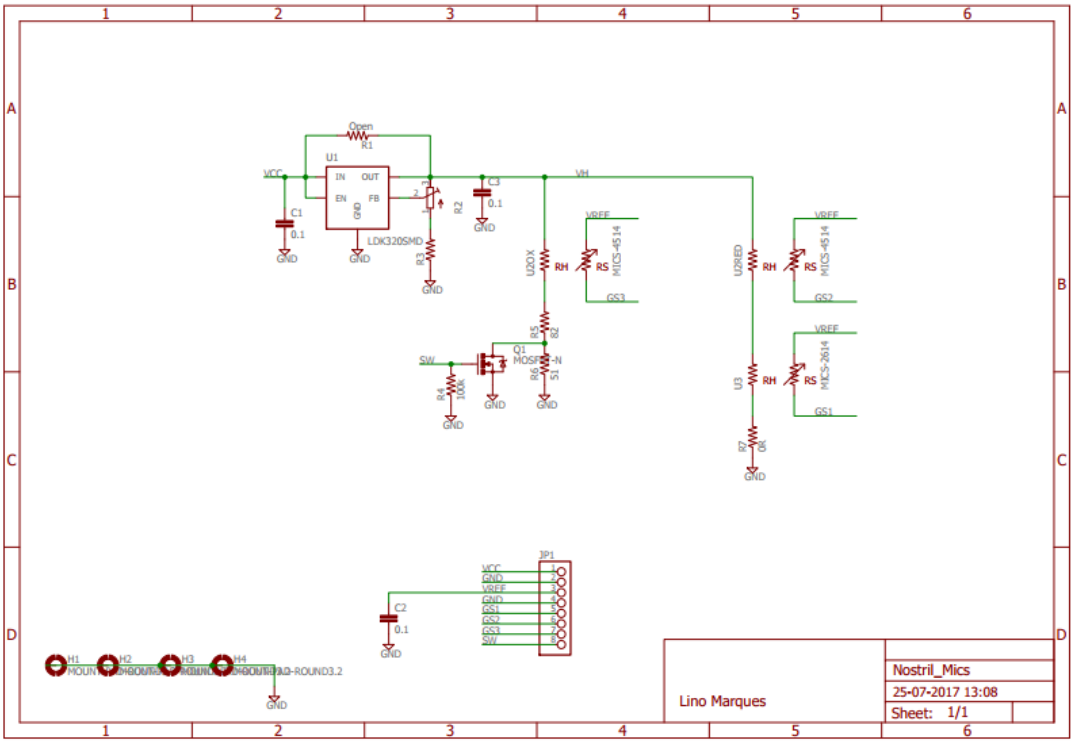
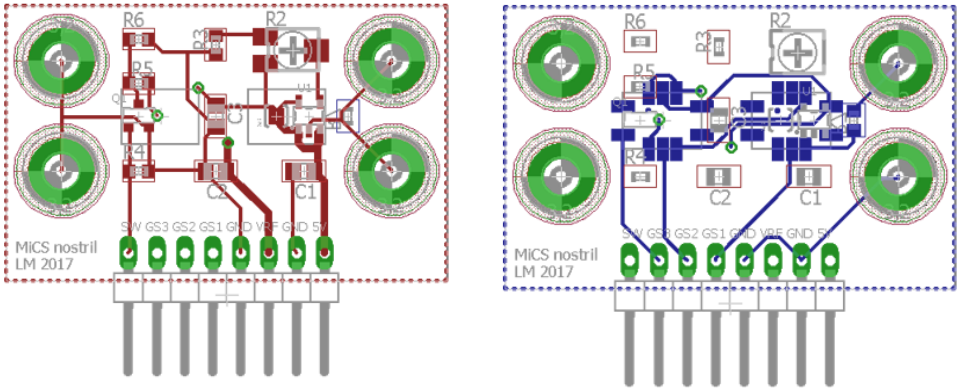


Figure B.1: Circuits of the printed circuit boards aimed for the sensors.



(a) **(b)**

Figure B.2: Printed board circuit designed to supply the sensors. a) Board for MiCS-5524 and MiCS-5914 and b) Board for MiCS-4514 and MiCS-2614.

Appendix C

Gas Chamber

This appendix is dedicated to the 3d chamber designs.

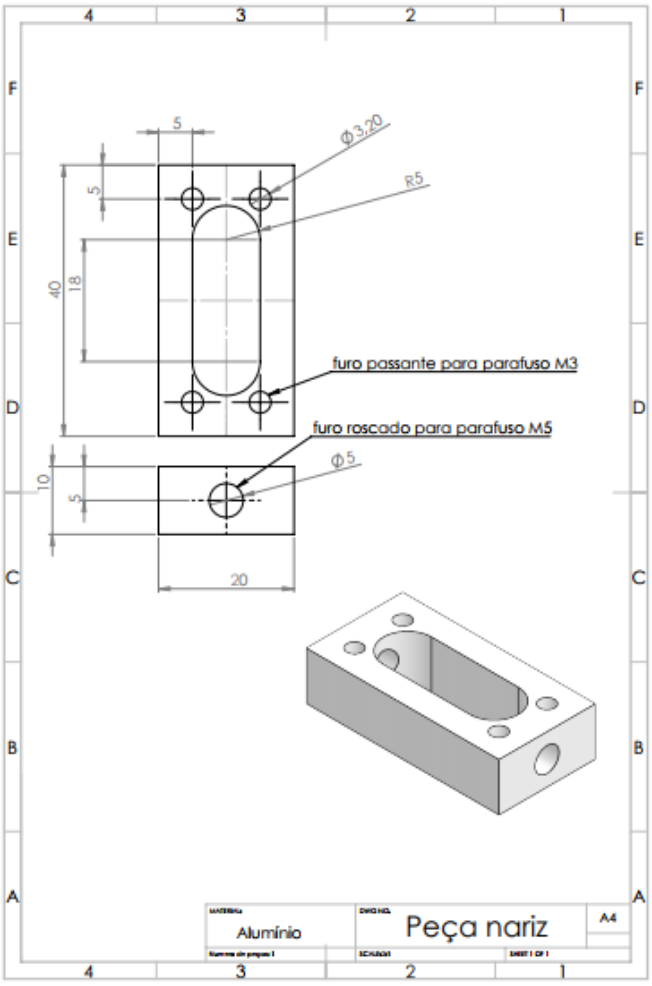


Figure C.1: 3D aluminium chamber designs in SolidWorks®.

Appendix D

Python Scripts

In this appendix, the scripts developed with data processing as aim during the development of this thesis are presented.

D.1 Feature Extraction

The group of functions that implement the selected feature extraction techniques are described in this script.

```
import numpy as np
import matplotlib as mpl
import pywt
from pywt import wavedec
from scipy.fftpack import fft, rfft
import re

def mean (data):
# Mean of the data prompted when calling this function
cenas = np.array(data);
mean_data = np.mean(cenas);
return mean_data

def slope_mean(data):
# Mean of the data points contained in the slope when the
    sensor response
```

```
# kicks in.
maxim= np.amax(data)
maxindex = np.where(data==maxim);
sample = str(maxindex);
maxind = re.findall(r'\[([^]]*)\]',sample);
maxinde = map(int, maxind[0].split(','))
maxi = maxinde[0];
data_slope = data[165:maxi];
mean_slope = np.mean(data_slope);
return mean_slope

def downslope_mean(data):
# Mean of the data points contained in the downwards slope,
   from the maximum point
# till the 300 s point
maxim= np.amax(data)
maxindex = np.where(data==maxim)
sample = str(maxindex);
maxind = re.findall(r'\[([^]]*)\]',sample);
maxinde = map(int, maxind[0].split(','))
maxi = maxinde[0];
pointer=maxi+300
data_slope = data[maxi:pointer];
mean_slope = np.mean(data_slope);
return mean_slope

def upderiv_mean(data):
# Mean of the derivatives data points contained in the
   upwards slope, from the maximum point
# till the 300 s point
maxim= np.amax(data)
maxindex = np.where(data==maxim)
sample = str(maxindex);
maxind = re.findall(r'\[([^]]*)\]',sample);
maxinde = map(int, maxind[0].split(','))
maxi = maxinde[0];
data_slope = data[165:maxi];
```

```
dx=1;
dataderiv=np.array(data_slope);
dy = np.diff(dataderiv)/dx;
mean_deriv = np.mean(dy);
return mean_deriv

def downderiv_mean(data):
# Mean of the derivatives data points contained in the
    downwards slope, from the maximum point
# till the 300 s point
maxim= np.amax(data)
maxindex = np.where(data==maxim)
sample = str(maxindex);
maxind = re.findall(r'\[([^]]*)\]',sample);
maxinde = map(int, maxind[0].split(','))
maxi = maxinde[0];
pointer=maxi+300
data_slope = data[maxi:pointer];
dx=1;
dataderiv=np.array(data_slope);
dy = np.diff(dataderiv)/dx;
mean_deriv = np.mean(dy);
return mean_deriv

def derivatives_mean(data):
# Mean of the derivatives of each data point
data = np.array(data);
dx=1;
dy = np.diff(data)/dx;
mean_deriv = np.mean(dy);
return mean_deriv

def wvt_transf(data):
# cA - approximation coefficients
# cD - details coefficients
# Decomposition of level 8 is the higher recommended
    considered the length of the data samples
```

```
# cA is the array of approximation coefficients
    corresponding to low frequencies
# cDs are arrays of details coefficients corresponding to
    high frequencies
# db6
cA8,cD8, cD7, cD6, cD5, cD4, cD3, cD2, cD1 = wavedec(data, '
    db6', level=8)
# Most descriminative coefficients: 7,8,9,10,11
mdcA = [cA8[6], cA8[7],cA8[8], cA8[9], cA8[10]];
return mdcA
```

D.2 Individual Odour Dataset

This appendix is dedicated to programming code developed in Python where individual odour data can be normalized and visualized. Feature extraction also takes place in this script.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import csv
from preprocess import *

## Load the data and Pandas column designation
rawdata1 = pd.read_table('../alc---2017-09-06_11h11m15s_log.
    txt', sep='\t', header=None);rawdata1.columns =
    [1,2,3,4,5];
          ...
          ...
          ...
rawdata25 = pd.read_table('../alc---2017-09-10_06h52m50s_log
    .txt', sep='\t', header=None); rawdata25.columns =
    [1,2,3,4,5];

## Data junction in a single matrix
datalist = [dataraw1, dataraw2, ..., dataraw24, dataraw25]
```

```
datajunction = datalist.concat(datalist);

## Drop the fifth collumn (timestamps)
datajunction = datajunction.drop(5,1)

## Plot sensor readings. In this case the sample considerd
    is the 18th
x=np.arange(3612);
ce =datajunction[0+(3612*17):3612*(18)]
y1=ce[1];y2=ce[2];y3=ce[3];y4=ce[4];
y1=(y1*4.096/32200);y2=(y2*4.096/32200);y3=(y3*4.096/32200);
    y4=(y4*4.096/32200);
x=x/float(9)
plt.figure()
plt.plot(x,y1,label='MiCS-5524',)
plt.plot(x,y2,label='MiCS-5914',)
plt.plot(x,y3,label='MiCS-2614',)
plt.plot(x,y4,label='MiCS-4514',)
plt.xlabel('Time(s)')
plt.ylabel('Voltage(V)')
plt.legend()
plt.show()

procdata1=[]; procdata2=[]; ... ; procdata25=[];
proclist=[procdata1, procdaaa2 , ... , procdata25];

label = 1;
j=0;

for i in datalist:
        #0 to 1 Normalization for each sensor for each
            sample
        ## datalist[j][1] = (datalist[j][1] - datalist[j
            ][1].min())/(datalist[j][1].max()-datalist[j][1].
            min())
```

```
## datalist[j][2] = (datalist[j][2] - datalist[j
    ][2].min())/(datalist[j][2].max()-datalist[j][2].
    min())
## datalist[j][3] = (datalist[j][3] - datalist[j
    ][3].min())/(datalist[j][3].max()-datalist[j][3].
    min())
## datalist[j][4] = (datalist[j][4] - datalist[j
    ][4].min())/(datalist[j][4].max()-datalist[j][4].
    min())

# Log normalization
## datalist[j][1] =datalist[j][1].apply(np.log);
## datalist[j][2] =datalist[j][1].apply(np.log);
## datalist[j][3] =datalist[j][1].apply(np.log);
## datalist[j][4] =datalist[j][1].apply(np.log);

# Wavelet coefficients list
w1=wvt_transf(datalist[j][1])
w2=wvt_transf(datalist[j][2])
w3=wvt_transf(datalist[j][3])
w4=wvt_transf(datalist[j][4])

proclist[j].extend((mean(datalist[j][1]),mean(
    datalist[j][2]),mean(datalist[j][3]),mean(
    datalist[j][4]),
slope_mean(datalist[j][1]),slope_mean(datalist[j
    ][2]),slope_mean(datalist[j][3]),slope_mean(
    datalist[j][4]),
downslope_mean(datalist[j][1]),downslope_mean(
    datalist[j][2]),downslope_mean(datalist[j][3]),
    downslope_mean(datalist[j][4]),
derivatives_mean(datalist[j][1]),derivatives_mean(
    datalist[j][2]),derivatives_mean(datalist[j][3]),
    derivatives_mean(datalist[j][4]),
upderiv_mean(datalist[j][1]),upderiv_mean(datalist[j
    ][2]),upderiv_mean(datalist[j][3]),upderiv_mean(
    datalist[j][4]),
```

```
        downderiv_mean(datalist[j][1]),downderiv_mean(
            datalist[j][2]),downderiv_mean(datalist[j][3]),
            downderiv_mean(datalist[j][4]),
        w1[0],w1[1],w1[2],w1[3],w1[4], w2[0],w2[1],w2[2],w2
            [3],w2[4], w3[0],w3[1],w3[2],w3[3],w3[4], w4[0],
            w4[1],w4[2],w4[3],w4[4],
        label))

# At last, data is written in a .csv file
with open("features_log.csv", "w") as f:
writer = csv.writer(f)
writer.writerows(proclist)
```

D.3 Training-Test Data Division

This script has as objective to separate randomly the initial dataset into training and testing set, respecting the 75-25% proportion.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

featuresFile = pd.read_table('features_log.csv', sep=',',
    header=None);

data = featuresFile.ix[:,0:43]
labels = featuresFile[44];
data.columns =
    [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,
22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43];


# Random selection of 20 samples of each odour to serve as
    the training set
# and 5 samples of each serve as testing set (75%-25%).
k=0;
```

```
training_set = pd.DataFrame();
testing_set=pd.DataFrame();

for p in range(8):
        data_odour = data.ix[k:k+24];
        #Samples for the training set
        data_75 = data_odour.sample(n=20)
        #Samples for the test set
        data_25 = data_odour.loc[~data_odour.index.isin(
           data_75.index)]
        training_set=training_set.append(df_75);
        testing_set = testing_set.append(df_rest);
        k=k+25;
training_set.to_csv('featuresFiles_train.csv', sep=',');
testing_set.to_csv('featuresFiles_test.csv', sep=',')
```

D.4 Recursive Feature Elimination

The following script implements the feature ranking algorithm applied in this thesis.

```
import numpy as np
import pandas as pd
from pandas import read_csv
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

training_set = pd.read_table('featuresFiles_train.csv', sep=
   ',', header=None);
X_values = training_set.values[1:161,1:43]
Y_values = training_set.values[1:161,44]
model = LogisticRegression()
rfe = RFE(model, 1)
fit = rfe.fit(X, Y)
print("Feature Ranking : %s") % fit.ranking_
```

D.5 Dimensionality Reduction

In respect to next script, data is projected and plotted in both dimensionality techniques sub-spaces. It is also stored in external files.

```
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.discriminant_analysis import
    LinearDiscriminantAnalysis
import pylab as pl
from mpl_toolkits.mplot3d import Axes3D

training_set = pd.read_table('featuresFiles_train.csv', sep=
    ',', header=None);
testing_set = pd.read_table('featuresFiles_test.csv', sep=',
    ', header=None);

training_set.columns =
    [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,
22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,
40,41,42,43,44,45];
testing_set.columns =
    [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,
22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,
40,41,42,43,44,45];

train_ready = training_set.ix[1:161,1:44];
classes = training_set.ix[1:161,45];
test_ready = testing_set.ix[1:41,1:44];
classes1= testing_set.ix[1:41,45];

## PCA
pca = PCA(n_components=5).fit(train_ready)
pca_train = pca.transform(train_ready);
pca_test = pca.transform(test_ready)
print(pca.explained_variance_ratio_)
```

```
explained_variance = pca.explained_variance_ratio_
## LDA

#lda = LinearDiscriminantAnalysis(n_components=5, solver='
    eigen',shrinkage='auto').fit(train_ready,classes)
#lda_train = lda.transform(train_ready);
#lda_test = lda.transform(test_ready)
#print(lda.explained_variance_ratio_)
#explained_variance = lda.explained_variance_ratio_

d1 = float("{0:.2f}".format(explained_variance[0]))
d2 = float("{0:.2f}".format(explained_variance[1]))
d3 = float("{0:.2f}".format(explained_variance[2]))

## Plot PCA/LDA 2D. PCA in this case
fig = plt.figure(1, figsize=(4, 3))
for i in range(0, pca_train.shape[0]):
        if classes.iloc[i] == 1:
        c1 = pl.scatter(pca_train[i,0],pca_train[i,1],c='r',
           marker='+')
        elif classes.iloc[i] == 2:
        c2 = pl.scatter(pca_train[i,0],pca_train[i,1],c='g',
           marker='o')
        elif classes.iloc[i] == 3:
        c3 = pl.scatter(pca_train[i,0],pca_train[i,1],c='b',
           marker='*')
        elif classes.iloc[i] == 4:
        c4 = pl.scatter(pca_train[i,0],pca_train[i,1],c='y',
           marker='v')
        elif classes.iloc[i] == 5:
        c5 = pl.scatter(pca_train[i,0],pca_train[i,1],c='m',
           marker='1')
        elif classes.iloc[i] == 6:
        c6 = pl.scatter(pca_train[i,0],pca_train[i,1],c='k',
           marker='p')
        elif classes.iloc[i] == 7:
```

```
        c7 = pl.scatter(pca_train[i,0],pca_train[i,1],c='c',
            marker='8')
        elif classes.iloc[i] == 8:
        c8 = pl.scatter(pca_train[i,0],pca_train[i,1],c='r',
            marker='4')

pl.xlabel('PC1 ('+str(d1)+'%)')
pl.ylabel('PC2 ('+str(d2)+'%)')
pl.legend([c1, c2, c3, c4, c5, c6, c7,c8], ['Ethanol', '
    Garlic', 'Ammonia', 'Gasoline', 'Lemon', 'Bay leaf', '
    Vinegar','Naphthalene'])
pl.show()

##Plot PCA/LDA 3D. PCA in this case
fig=pl.figure(2)
ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
for i in range(0, pca_train.shape[0]):
        if classes.iloc[i] == 1:
        c1 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2], c='r',marker='+')
        elif classes.iloc[i] == 2:
        c2 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='g',marker='o')
        elif classes.iloc[i] == 3:
        c3 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='b',marker='*')
        elif classes.iloc[i] == 4:
        c4 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='y',marker='v')
        elif classes.iloc[i] == 5:
        c5 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='m',marker='1')
        elif classes.iloc[i] == 6:
        c6 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='k',marker='p')
        elif classes.iloc[i] == 7:
```

```
        c7 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='c',marker='8')
        elif classes.iloc[i] == 8:
        c8 = ax.scatter(pca_train[i,0],pca_train[i,1],
            pca_train[i,2],c='r',marker='4')

ax.set_xlabel('PC1 ('+ str(d1)+'%)');ax.w_xaxis.
    set_ticklabels([])
ax.set_ylabel('PC2 ('+ str(d2)+'%)');ax.w_yaxis.
    set_ticklabels([])
ax.set_zlabel('PC3 ('+ str(d3)+'%)');ax.w_zaxis.
    set_ticklabels([])
pl.legend([c1, c2, c3, c4, c5, c6, c7,c8], ['Ethanol', '
    Garlic', 'Ammonia', 'Gasoline', 'Lemon', 'Bay leaf', '
    Vinegar','Naphthalene'])
pl.show()
```

D.6 Classification

At last, both classifiers are constructed and applied to the groups of data considered.

```
import numpy as np
import pandas as pd
from sklearn import svm
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score

# Data loading
train = pd.read_table('featuresFiles_train.csv', sep=',',
    header=None);
test = pd.read_table('featuresFiles_test.csv', sep=',',
    header=None);

train_n = train.ix[1:161,1:44];
trainlabels=train.ix[1:161,45];
```

```
test_n = test.ix[1:41,1:44];
testlabels=test.ix[1:41,45];

## Support Vector Machine
clg = svm.SVC(kernel = 'linear');
svc = svm.SVC();
clg.fit(train_n,trainlabels);
acc = accuracy_score(clg.predict(test_n),testlabels);
print('SVM accuracy: ',acc)
print('Labels predicted: ',clg.predict(test_n))

## Neural Network
cln = MLPClassifier(activation ='tanh',solver='adam', alpha
   =1e-5,
hidden_layer_sizes=(15, 7), random_state=1, max_iter =10000)
cln.fit(train_n,trainlabels);
print('NN accuracy: ',accuracy_score(testlabels, y_pred))
print('Labels predicted: ',cln.predict(test_n))
```