Experimental Evaluation of a Gas Sensors Array for the Identification of Complex VOCs Mixtures in the Breath of Patients by Pattern Recognition Techniques

Justin D.M. MARTIN[1]
[1]*University of Liège – Campus Environnement, S.A.M. Laboratory, Avenue de Longwy 185, Arlon, Belgium*
Jdm.martin@uliege.be

1 RESEARCH PROBLEM

Lung cancer is one of the deadliest form of cancer in Europe, being the first and second cause of cancer death respectively for men and women (Ferlay et al., 2018).

This high death toll has to be blamed on the lack of obvious symptoms in the early stages of the illness, during which surgery has to be performed in order to give high chances of recovery to the patient (The National Lung Screening Trial Research Team, 2011). Current diagnostic methods tend to be expensive, cumbersome, slow and requiring rare, qualified personnel. This often means that screening is costly and difficult to organise at a large scale. Asymptomatic subjects and people in remote areas are rarely tested overall, leading to late discovery of the cancer and poor survival chances (Silvestri et al., 2016; Westeel et al., 2007).

There is therefore a need for a diagnostic method that could be used remotely while being simple enough to be used with little prior formation. Ideally, that method would also be cheap enough to be widely deployed, and non-invasive for simplicity.

An interesting technology would be gas sensor arrays. Already widely used in a variety of sectors (food industry, odour characterization, environmental monitoring), they have gained interest as an experimental medical diagnostic method.

This thesis is part of the PATHACOV (*Pathacov project*, 2020) research project, funded by Interreg France-Wallonie-Vlaanderen. This project aims a gas sensor array to serve as a screening apparatus to detect early stage lung cancer.

2 OUTLINE OF OBJECTIVES

This thesis aims at creating and testing a sensor array in order to build a benchmark on which one can compare the discriminative power of different arrays. In order to achieve this, several tasks will be performed simultaneously:

The first is the establishment of a standardized test method of the metrological characteristics of commercial thick film sensors as well as experimental ones, and their qualities within a sensor network.

The second is the integration of experimental sensors into a prototype gas sensor array ("electronic nose") consistent with the framework of the Pathacov project.

The third is the validation of the test method with the prototype electronic nose, which requires the reproducible synthesis of reference gas mixtures. The concentrations of volatile organic compounds (VOC) that characterize the breath of a cancer patient are in very low concentrations (part-per-billion (ppb) to part-per-million (ppm) level), and a specific reliable method has to be developed to reach those concentrations. It is also planned to use real breath from healthy persons and cancer patients as validation of the benchmark's conclusions.

The last task is about the processing and analysis of data and the identification and classification of samples in order to obtain a measurement of the array's discriminatory power. Each sensor has to be characterized and the power of each version of the sensor array evaluated on a common benchmark. This enables iterative improvement of the tested devices.

3 STATE OF THE ART

3.1 Cancer Biomarkers in Breath

Breath has several rather unusual characteristics as a gas sample: temperature close to the body's, increased carbon dioxide content, saturated in humidity, and most interestingly it contains a complex mixture of Volatile Organic Compounds (VOCs) in the ppb to ppm-level concentration. Variations in this mix can be linked with metabolism alterations and pathologies, including cancer.

Therefore, it has been hypothesized that some of these VOCs are lung cancer biomarkers.

The vast majority of supposed VOC cancer biomarkers are also present in a healthy person's breath, but in different concentrations. While afflicted by metabolic disorders or various illnesses, ratios change and sometimes new VOC markers appear. Cells normally produce a certain number of reactive oxygen species (ROS) as by-products of mitochondrial metabolism, inducing oxidative stress that is managed by anti-oxidant mechanisms. The presence of these ROS creates VOCs by reacting with organic material in the cytoplasm (Aksenov et al., 2012).

Cancer cells have an over-active metabolism that create a large quantity of ROS (which can leak in and out of a cell) and a hypoxic environment. Without oxygen, cancer cells continue to create energy by glycolysis, acidifying the environment and draining more glucose than normal cells (Warburg effect). Genetic mutations causing the build-up of Cytochrome p450 and other oxidase enzymes have been linked with tobacco and lung cancer, and as they react with ROS, they tend to change the ratios in emitted VOCs (Filipiak et al., 2016).

Mechanics linking breath's VOCs composition and health status aren't clearly understood. What is known so far is that the "volatilome" (the ensemble of VOCs in breath) isn't entirely due to the cancer cells alone, as in-situ cancer tissues do not emit the same signature as the cells in culture, regardless of the culture medium (Pesesse, 2019). It is probable that the VOCs signature is mainly influenced by the reaction of healthy cells to the presence of the cancer cells in the body (Capuano et al., 2015; Huang et al., 2018). It is unknown if the phenomenon is localised around the afflicted tissues, or if a greater share of body cells reacts to the cancer presence. Interestingly, it has been shown that the cancerous lung can be distinguished from the other (cancer-free) lung by its volatilome on the same patient (Capuano et al., 2015).

The first objective of the state of the art for this project was to find out what kind of VOCs were characteristic of a "healthy patient" or a "cancer patient" breath, and in which concentrations VOCs were observed. The best way to do so was to make a full inventory of gas chromatography mass spectrometry (GC-MS) studies aiming at cancer breath characterization. About 42 articles from 1985 to 2019 were selected and synthesized into a grid grouping biomarkers by frequency of citation as potential biomarkers. The number of publications increased exponentially over the prospected 34 years: more than two-third of the publications appeared after 2010, hence the rather small number of publications on this specific subject.

Studies on the identification of cancer biomarkers are widely divergent, although it is possible to identify certain trends and identify the best candidates. This review was the basis for the selection of VOCs for use in the benchmark's synthetic atmospheres. The sources of contamination for the most cited biomarkers are varied and frequently encountered on a daily basis, adding a certain level of incertitude to the conclusions, as patient's breath can be contaminated by exposure to those sources.

On the matter of lung cancer detection, it is important to note that there is no consensus on which VOCs are linked to cancer presence, since there haven't been two independent studies with the same results, as observed by Jia (Jia et al., 2019) and during this literature review. The vast majority of studies include less than 100 people, as illustrated below (Figure 1). The biggest study included 484 people (lung cancer patients and controls included), and the smallest as few as 8 people. The mean number of subjects was 131, and the median 89.

A lot of studies concentrate on the hypothesis that cancer cells are the origin of the volatilomic difference of cancer patients (Wang et al., 2014), including in-vitro cancer cell cultures studies (Chen et al., 2007; Jia et al., 2019; Thriumani et al., 2018). Since most of the published results seem to conclude that this hypothesis isn't to be followed, it was decided not to consider results from in-vitro-only studies.

The most frequently cited compounds are as follows:

- With 10 occurrences each (amongst the 42 selected articles), the most cited compounds are 2-butanone (or methyl-ethyl-ketone), isoprene and 1-propanol.
- Hexanal has 9 occurrences.
- Ethylbenzene and acetone were cited 8 times each.
- Pentane and 2-Propanol were cited 7 times each.
- Benzene, hexane and decane were cited 6 times each.
- Toluene, propanal, nonanal, styrene, heptanal and ethanol were cited 5 times each.

11 compounds were cited 4 times, 9 compounds were cited 3 times, and more than 181 compounds were cited 2 times or less. This last category was not studied any further, as the relevance of each compound was likely to be very low.

Nevertheless, the review proved itself useful to diminish the number of potentially interesting VOCs to choose from. A variety of putative biomarkers was chosen with an educated guess from the list, with the intention to use them during lab testing. Main choosing criterions were: short half-life in the body, not closely linked to cigarette usage, found as relevant for studies cumulating a large number of test subjects, not found to be exclusively exogenous, not highly correlated with physical activity.

From the previously cited compounds, 4 have been picked as interesting. The compounds are: 2-butanone, decane, 2-pentanone, dodecane.

To complete our testing toolbox, several other compounds were acquired. The purpose of these supplementary compounds is to observe their influence on the tested device's ability to correctly identify "sick" breath from "healthy" breath. Among them, compounds frequently found in breath, likely confounders (smoking-related compounds for example), or possible biomarkers that are also likely to be exogenous. The compounds are: Pentane, 1-Propanol, Methanol, Ethanol, 2-Pentanone, Acetone, Heptanal, Hexanal, Benzaldehyde.

3.2 Gas Sensor Array

A gas sensor array (GSA) aims at qualifying (and sometimes quantifying) mixes of gases. Since it's been historically used to detect odours and the general working principle is inspired from the olfactory system, the term electronic nose (e-nose) is often used to name that kind of sensor arrays, even if what the "nose" detects isn't necessarily odorant compounds (Romain et al., 2002). For example, carbon monoxide and methane are odourless molecules, but can be detected by metal oxide gas sensors.

The particularity of an e-nose is that its working principle relies on the non-specificity of its sensors: such a system will not identify what compounds are in a gas sample (like a GC-MS would), but will look at the general imprint the whole mixture makes on each sensor. A GSA can recognize mixes of several hundreds of compounds at a time and tell them apart.

On a single measurement, it is highly probable that several sensors will react to a single compound, and that several compounds will react at the same time on a single sensor. Since each sensor has a different sensitivity, the response of the whole array will be characteristic to the mixture's composition. The "imprint" of the mix will therefore be a vector composed of the response of each sensor in the array.

That vector will enable to identify the mix by comparison with a reference set of previously measured reference mixes, and therefore qualify it – e.g. healthy or sick, coffee or tea, pleasant or unpleasant – depending on what we aim to do (Gardner & Bartlett, 1999).

GSA are composed of at least the same 4 types of elements (Gardner & Bartlett, 1999):

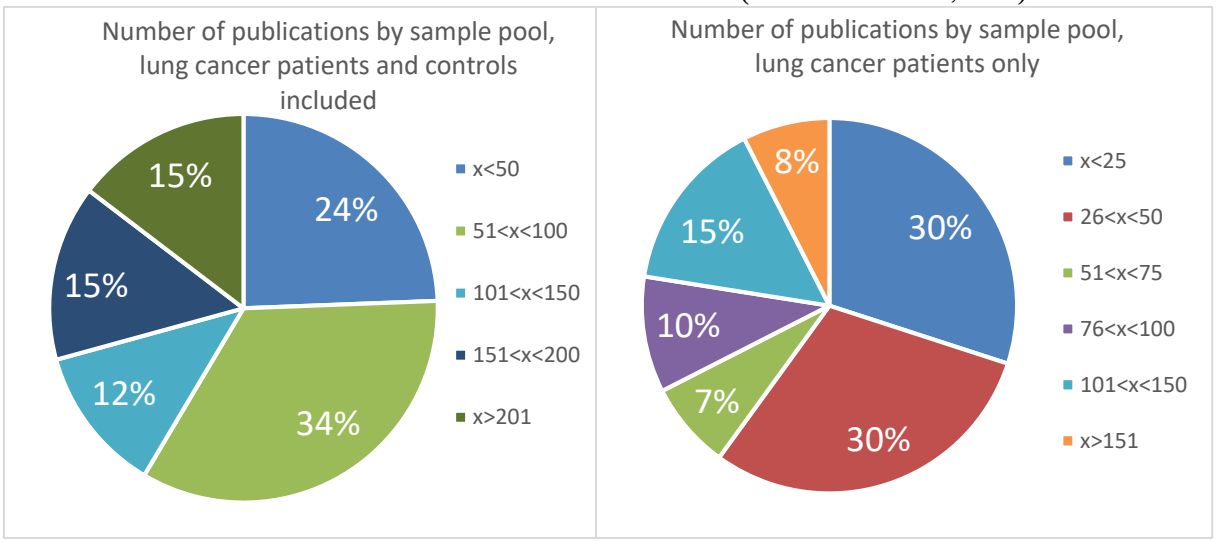

Figure 1a : Number of publications by sample pool, including controls. The majority of them have less than 100 people. Figure 1b : Number of publications by sample pool, lung cancer patients only. We notice than more than two third of the publications had access to only 75 patients or less. The low statistical significance of the results is always cautiously noted by the authors, calling for larger sized experiments.

• A sample treatment system, storing and putting the samples in contact with the array, and in some cases having a form of pre-concentration or pre-treatment of the sample. It can be as simple as a tube with a mouthpiece, in the case of inline analysis, or as complex as a sorption/thermal desorption system.

• The sensor array itself, usually composed of 4 to 32 sensors housed in one (or several) chamber(s). Usually, they are of a single type (e.g. metal oxide semiconductor sensors (MOS)) but the model and/or the working temperature varies, giving the sensors different sensitivity ranges.

• A signal treatment system, which converts the analogic output of the sensors to a numeric output interpretable by the processing unit.

• A processing unit. Basically a computer that will pilot the other units, collect and register the data from the sensors, process them and display the results to the user accordingly to its user-created protocol.

Most reported methods managed a good discrimination of lung cancer with sensor arrays, some even managed to isolate specific mutations of cancer using the breathprint (Shlomi et al., 2017). However, not unlike GCMS studies on cancer volatilome, the samples of population for cancer discrimination by GSA in literature are rather small (usually, 20-100 patients, half of them being controls, with some exceptions having up to 300 patients) and further researches are needed.

3.3 Breath Sampling and VOCs

In breath, VOCs are exogenous (environmental contamination) or endogenous (produced by the metabolism) in origin. For disease detection, endogenous compounds are more likely to be interesting. To understand how to sample breath correctly, it is important to study the factors influencing its composition.

Exposure to VOCs in ambient air can result in durable contamination of the breath. The rate of removal of VOCs from the body depend on initial concentration of VOCs in the air, the duration of exposure, the solubility in blood and lipid tissues, and physiology. Some studies have shown that the inspired VOCs are partially retained by the body, depending on their affinity for fatty tissues and blood, and can be exhaled later on (Jia et al., 2019).

During sampling, different parameters have been reported to influence VOCs concentration in literature.

Sampled part of breath: the first part of the breath is often “dead-space” air that is different in composition to alveolar air – end-of-breath air is in close contact with blood and lung tissues, and often regarded as more interesting to sample (Doran et al., 2017). Using capnography as a method of fractioning breath samples is possible, and envisioned in this PhD thesis.

Ventilation rhythm: breath holding tends to rise the concentration of some VOC species, and hyperventilation seems to have the opposite effect. This is likely related to the time needed for alveolar air to reach an equilibrium with the blood (Boshier et al., 2011; Herbig et al., 2008). It is possible to ask the patient to blow at a given rhythm to moderate this effect.

Heartbeat rate (HR) and blood pressure: since the vascular system is closely linked to the lungs, the behaviour of the heart is an important factor, as a greater blood flow can expose a greater amount of dissolved VOC to the air-blood interface. Isoprene, for example, has been shown to be linked with heartbeat rate (Karl et al., 2001). In a similar manner, airway resistance has been shown to increase isoprene levels (Sukul et al., 2017). It is therefore important to ensure HR is stable during sampling. Stress and effort can raise HR, an acclimation time should be respected before sampling to give the patient some time to reach as stable HR.

Contamination from sampling materials is a concern. Ideally, sampling tools should be made in inert materials such as Teflon, stainless steel or glass. Contamination of the apparatus by bacteria and saliva should be avoided (Doran et al., 2017). Using bacterial filters and inert materials is an easily implementable solution.

Age and gender of the patient: Some compounds like ammonia increase in concentration with age, and several VOCs seem to be gender specific, although data are conflicting (Horváth et al., 2017; Jia et al., 2019). McWilliams et al. has shown that gender played a minor role in the discrimination capacity of their gas sensor array, female patients being less likely to be correctly classified by the breath-sensing device (McWilliams et al., 2015). In most studies, the studied groups of people are matched in age and gender repartition to ensure no false conclusions are drawn from the study. However, as lung cancer has a lower incidence with women, studies often report having fewer women patients included.

Diet: one should note that diet might have a lasting influence on breathprint, whether or not fasting is involved before sampling (Doran et al., 2017; Horváth et al., 2017; Jia et al., 2019). As this aspect is unavoidable in practice, the best approach would be to take note of the patient’s diet and observe the variation on the subject on a long period of time.

A 4h nothing-by-mouth policy can also be implemented to moderate the effects. The device should, ideally, not be affected in its classification performances if its design is robust enough.

Smoking: cigarette smoke contains several hundreds of VOCs, and seems to be affecting a number of metabolic pathways, such as those linked with oxidative stress, that influence the breathprint further. Smokers, former smokers and non-smokers can be accurately identified with sensor arrays (Horváth et al., 2017). Effects of smoking can be moderated with a 12h no-smoking policy. Smoking as a risk-increaser could also be used in the decision algorithm of the device, smoking markers becoming an additional information instead of an interference.

Other diseases: a large number of medical conditions have a known effect on the breathprint (asthma, liver diseases, chronic obstructive pulmonary disease (COPD) to name a few) (Jia et al., 2019). McWilliams et al has shown that a gas sensor array could discriminate COPD and lung cancer, without the COPD status of the patient playing a role in lung cancer discrimination (McWilliams et al., 2015). Comorbidity can only be compensated by including a similar share of affected people in all studied groups.

Medication: drugs have been investigated as a confounding factor. Any recent use of drugs with anti-inflammatory effect or causing dilatation of the airways should be recorded before sampling (Horvath et al., 2009).

Histology: Different type of cancer, and different affected zones in the lung, seem to give different breathprint. E-noses managed to discriminate between stages in the past with good accuracy, but on small sample populations (Barash et al., 2012; Broza & Haick, 2013; Peled et al., 2012). It is unfortunately impossible to confirm presently if this level of discrimination is attainable on larger sample populations. Such precision is not mandatory for a screening device as envisioned by the project.

Human genotype and/or habitat of subject: an international study has looked into international variations in breath VOCs, which are important to consider while building a device that has to be usable around the world. Their study showed significant differences and classification accuracy between Latvian and Chinese patients (Amal et al., 2013). It is however difficult to determine what is the direct cause of the observed difference from the reported data. Diet is a responsible cause as likely as genotype or other environmental factors. Sampling people of varying genotypes living in a similar environment could give some insight on this matter.

Time (diurnal cycle, seasons…) has an influence of its own, and also because it influences the sampling conditions (humidity in winter and summer differs, time of day changes the time from last meal…). Most studies avoid this influence by sampling in a stable, controlled environment at a defined pace. However, mobile sampling devices should also be assessed for robustness. There have been several studies regarding VOCs measurement in real time, and we know for example that acetone levels in breath slowly increases during the night (de Lacy Costello et al., 2014). Sampling at the same time of the day should moderate the effects of diurnal variability. Prolonged study of the control group should give some insight on the general time-related variability.

Mental status: general happiness, depression or mental illnesses can cause metabolic variations. There have been several studies exploring the discrimination of schizophrenia patients from controls with breath sampling, showing that there seems to be higher levels of ethane linked with oxidative stress in people with the disorder (Ross et al., 2011).

If some of the previous effects can be avoided or moderated with careful sampling, others are simply unavoidable. Fortunately, GSA's are able to overlook some interfering compounds thanks to careful array conception and multivariate analysis. Main confounding factors can be identified during the patient's interview and possibly compensated during data treatment.

Evaluating the sensitivity and the specificity of a screening test is a well-treated subject in literature, with a number of statistical tools designed to help researchers calculate the adequate sample size. The size of "sick" and "control" groups can be estimated based on several values: the prevalence of the disease (fraction of the general population that can be affected), the confidence level, the desired precision of the estimation (the maximum difference between the estimation and the true value) and the estimated sensitivity and specificity (from previous studies or clinical expertise). For a screening tool, sensitivity has to be high, but a lower degree of specificity can be tolerated (Bujang & Adnan, 2016; Hajian-Tilaki, 2014).

4 METHODOLOGY

4.1 Sensor Array

In order to establish a benchmark, a prototype sensor array is needed. The device needs to be usable for breath technically sensing to stay relevant.

The prototype should therefore have the following characteristics:

- Quick sensor response, to keep the analysis time as short as possible.
- CO_2 sensor included. It can be used for capnography and easy signal treatment. This also enables the automatic selection of the wanted part of the exhalation (see 3.4 Breath Sampling and VOCs), as CO_2 levels tend to rise toward the end of the breath.
- Small volume sensor chamber, as this avoids dilution of samples, provides quicker signal stabilisation, and keeps the design compact and lightweight.
- Temperature/humidity measurement and sensor chamber heating for temperature regulation: as breath is water-saturated at about 37°C, rising the temperature of the system and monitoring these parameters guarantees the absence of condensation. Water in the system could remove some chemical species from the gas phase, which would alter responses from sensors. Also, MOS sensors are known to be sensitive to humidity.
- Contamination avoidance: all materials in contact with the samples should be non-emissive and resistant to chemical alterations. Stainless steel or PTFE are preferred. Devices that cannot be made in those materials will be placed down flow from the sensor chamber.
- Flow monitoring: a flow control device (for instance a rotameter) and pump are placed downflow to ensure constant flow from the sample's storage (FEP sampling bag). Offline analysis of collected samples was found to be easier to obtain a stable signal from, and was chosen as the preferred way of operation.

By comparing the characteristics supplied by the manufacturers, a varied range of sensors was acquired for evaluation. During the first tests, the Figaro Engineering TGS® 2603, Umwelt Sensor Technik® G3530, G1430, G2530, G8530 and Winsen® MP901 sensors were assembled on the prototype. To measure temperature and humidity, a Bosh BME680 sensor was placed in the sensor chamber. This last sensor also includes a thin film MOS sensor.

For the carbon dioxide sensor, high sensitivity and short response time were preferred for capnography. It is also necessary to have a sensor operating in the range 0-6% CO_2 content. Therefore, the Sprint-IR infrared (GGS®) sensor was selected.

The electronic part is built around a Teensy® 3.5 (PJRC) card (Figure 2). Data is sent to a computer which records and displays in real time the graphs of the parameters (conductance of the sensors, oxygen, carbon dioxide, temperature, and humidity values).

Figure 3 : Prototype of breath sensing gas sensor array used for benchmark building, nicknamed SAMBRE_1

4.2 Gaseous samples

4.2.1 Synthesis

To test the discrimination power of an experimental device, it is necessary to create samples to be analyzed by it. Those samples should be as close as possible to real breath in composition while being sound logistically and reproducibility-wise.

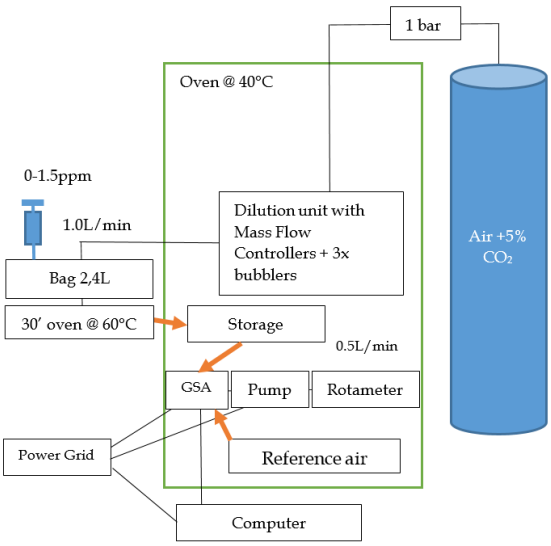

Figure 2: General setup for gas samples synthesis and GSA testing

One of the challenge in making breath-like mixes for volatolomics is the concentration range, which requires accurate dilution of compounds down to the part per billion (ppb) level. A usual approach, chosen for this project, consists in the insertion of a few microliters of the liquid compounds in a gas sampling bag filled with a known volume of air (Figure 3), the gas volume of the vaporized compounds diluted in the air allows to reach ppm-level concentrations. Volatilization of VOCs is ensured by heating the bag for 30 minutes at 60°C. The obtained mix is then fed into a system using Mass Flow Controllers (MFC) diluting the mix further down to the ppb level. In this project, the final sample is stored for a short time in another gas sampling bag before usage. Storage is made at elevated temperature (40°C) to avoid any condensation.

Humidification of the gas samples is made by routing a fraction of the dilution air with MFCs to bubblers. As shown on figure 3, most of the dilution setup is kept at elevated temperature.

Canisters of CO_2 enriched air are used as dilution air and pre-made VOC mix canisters (Westfalen®) are used to provide background VOC interfering compounds (ppm level) to the dilution unit.

4.2.2 Gas Chromatography

In order to confirm the composition of the samples given to the gas sensor array, and evaluate the reproducibility of the dilution method, a reference method is needed: thermal desorption gas chromatography mass spectrometry (TD-GC-MS) has been taken as the method of choice, as it is the most commonly used method for volatilome analysis in literature.

A Trace GC oven and DSQII mass spectrometer (Thermo Fisher Scientific®) is coupled with a TD-100xr (Markes®) thermal desorber to analyze Tenax® TA packed cartridges. Bags are connected to the cartridges and emptied with a GilAir Plus Personal (Sensidyne®) sampling pump. To ensure good separation, a VOC-specialized capillary column is used (Rxi®-624Sil MS, 0.25mm ID 1.4um 60m, Restek®).

The samples are analyzed using the following GC oven program: hold 50°C for 5min, 10°C/min to 180°C, 6°C/min to 205°C, hold 205°C for 5min.

Quantitative analysis is made after calibration on four compounds (pentanone, heptanal, decane, toluene). Calibration is made using standard-spiked cartridges on three different concentrations with three replications each. The concentration range for the calibration is based on the likely amounts of a single VOC biomarker found in a 2L breath sample according to literature (10ppb to 1ppm). Using a linear regression model from the peak areas for each compound, following samples can be quantified and in-bag concentration can be calculated. Compounds without a calibration line are measured in toluene equivalent. For every gas bag sample processed by the GSA, a duplicate is created to be adsorbed on a cartridge and analyzed in parallel.

4.3 Sensor Characterization

This step corresponds to the metrological analysis of the sensors both individually and within a network. Since some of the tested sensors will be prototypal in nature, characteristics such as response time, limit of detection, recovery time and sensitivities to different compounds are evaluated. The effect of variables – i.e. absolute humidity, VOC content (propanol, pentanone, heptanal, decane were chosen for this part), and O_2/CO_2 ratio – is studied for each sensor.

With a large number of parameters and possible second order effects between variables, it is necessary to reduce the number of experiments to make the analysis technically affordable. One of the best methods is to use an experimental design that will find the experiments providing the maximum amount of information in as few steps as possible.

The interpretation of the results of the experimental design (with Design of Experiment (DoE) modelling and Principal Component Analysis (PCA)) made it possible to assess the sensitivity of the sensors to the parameters. The interaction between parameters and the reproducibility of the experiment are also assessed.

DoE experience plan are synthesized and analysed using R code (FrF2 and DoE.base packages are used).

Correlation between sensors is evaluated. This is an important aspect of the analysis as a sensor highly correlated to another sensor bring no new information of the sample. This often suggests that the sensor is redundant and could be swapped for another model, improving the general performance of the array.

4.4 Gas Sensor Array Benchmark

4.4.1 Benchmarking

To evaluate how well a Gas Sensor Array can distinguish between two mixtures, the most straightforward way is to expose the array to two populations of the mixtures, and observe how well it

is possible to class the individuals using multivariate analysis.

The synthetic mixtures should have common points which may vary slightly to simulate a real breath population (varying absolute humidity and CO_2 concentration, different interfering compounds, varying VOC concentrations). The goal is to identify the extent to which the network is able to separate two given populations of mixtures (resolution) despite some heterogeneities. The mixtures are always created in a reproducible way, and all benchmarked devices measure statistically similar samples.

The dataset obtained from the GSA's is split into two: one is used to train a classification model, the second to test it (4.4.4 Multivariate Analysis).

4.4.2 Human Breath Sampling

Actual breath populations will also be used in the datasets, bringing the prototype as close as possible to actual field use. Sick breath populations will be provided by the university hospitals partners of the project, while healthy breath populations will be sampled among volunteers in the Arlon campus teams.

Sampling is operated using a prototype device that stores the breath in a bag, kept at elevated temperature to avoid condensation effects. The device is currently being developed by one of the project's teams and will be detailed in future articles.

Sample bags are quickly processed by electronic noses and GC-MS in parallel, in a way similar to the synthetic atmospheres analysis detailed before.

The experimental setting is the main source of concern when it comes to exogenous compounds. If the ideal sampling room is most likely void of any contamination, this approach is incompatible with the door-to-door diagnosis envisioned for the final device. The device should be able to differentiate categories of breath with an estimated degree of confidence regardless of the common interferences. Nevertheless, interferences from the environment should be investigated. The influence of the room on the classification should be studied by making repeated measurements on several subjects in varying environments. Repeated measurements of the breath of the same group of persons on several months should give some insight on the effects of variation of activities and habits on the measurements. Regarding confounders such as genotype, hormonal status or mental health, they will be considered as out of the scope of the thesis for the time being. Test subjects will be chosen within a homogeneous population for these aspects. However, a proper study of these aspects is advisable in the near future.

A questionnaire covering aspects influencing the composition of exhaled air has also been designed: medical history, lifestyle and potential sources of contamination encountered in the past few days are treated. This data is used to understand the influence of confounding factors. Aside from the useful information, filling the questionnaire also gives some resting time to the patient. It is useful to let the patient breathe the air of the sampling room for several minutes to avoid exogenous contaminations, and reach a stable heartbeat rate. As it is impossible to control every confounding factor, we thrive to control what can be reliably controlled. Other factors are taken note of in order to observe their influence on the data.

4.4.3 Pre-treatment and Data Visualization

The sensors signal dataset is too plentiful to be used directly, and therefore needs to be pretreated. RStudio was used to create a program in R language, which process the data in a usable form.

The first part of the code extracts the useful features. It can identify where the signal of interest starts, ends, and then computes various parameters (height of the signal's plateau, area under the curve, start and end slopes) taking into account the baseline signal and smoothing the noise as needed using a moving average. Normalisation or Standardisation of the array's signals may also be applied to improve classification.

This pre-processed data is recorded and sent to another code which is responsible for Principal Component Analyses, data structure visualisation and multivariate analysis (Linear Discriminant Analysis for instance).

Several functionalities can also be implemented at this stage, such as baseline drift compensation to compensate the progressive conductance shift of aging sensors. Humidity compensation can also be considered, using data from a recent humidity calibration of the sensors.

4.4.4 Multivariate Analysis

Considering the creation of a benchmark, pre-treated dataset will be processed using multivariate analysis.

Various methods will be tested to create a classification algorithm. Amongst the candidates are Neural Networks, k-Nearest-Neighbours, Linear Discriminant Analysis, Partial Least Squares, Random Forest and AdaBoost. These methods have

been used previously in literature for GSA classification. It is however necessary to assess which one is the most interesting for our task.

The dataset is split into two parts: one for the training of the classification model, and another for the validation of the algorithm's classification performances. An external dataset is also used to validate the model, which evaluates how robust the classification is. Based on classification performances of a chosen method and characteristics of sensors, different GSAs will be compared. The classification error on the external dataset would be used as a metric of GSA performance.

Since the benchmarking procedure is always the same for tested GSAs, it is possible to compare them and optimize them iteratively to create the best prototype possible. This will ensure final hospital field testing better chances of success.

5 EXPECTED OUTCOME

This PhD Thesis has several expected outcomes. Experiments on ppb-level dynamic gas dilutions and gas sensor array optimisation for VOC discrimination will constitute a base of knowledge for other projects studying VOCs at low concentrations. The benchmarking approach can also be used for other projects using GSAs. The produced data would have interesting features: comparison of lab-made "synthetic breath" and real patient breath, confusion factors and potential contaminations being taken into account, availability of large project-wise data.

As a whole, this PhD Thesis is contributing to the creation of a portable screening device against lung cancer. Such a device could later be re-purposed to detect other health conditions, opening the way to new diagnostic methods.

6 STAGE OF THE RESEARCH

The thesis began on May 15, 2019 for a period of 4 years.

According to the state of the art research, the first reproducibility tests of atmospheres dilution for sub-ppm concentrations have been made. A reproducible method has been established with less than 20% variation between days, for the worst case.

A working prototype of GSA fitting the requirements has been assembled (SAMBRE_1), and tested using synthetic atmospheres. Commercial sensors have been characterized and the first sets of synthetic atmospheres have been processed for classification. Code on pre-processing, DoE and PCA are operational and have been tested on several datasets from SAMBRE_1.

The next steps will start in the coming months: real breath sample pool collection and analysis, in-depth analysis of DoE outputs, creation of the code for automatic multivariate analysis of datasets, first tests with prototype sensors and new versions of SAMBRE, comparison of classification performances.

ACKNOWLEDGEMENTS

We thank Prof. Dr. Anne-Claude ROMAIN for supervising this thesis. We thank Dr. Nathalie REDON and Dr. Pierre-Hughues STEFANUTO for their work as thesis comitee members.

We thank Laurent COLLARD and Noémie MOLITOR for their help in the realisation of the experiments. We thank Claudia FALZONE and Marie SCHEUREN for the proofreading of this article.

We thank the partners of the PATHACOV project for the helpful discussions.

This work was supported by Interreg France-Wallonie-Vlaanderen.

REFERENCES

Aksenov, A. A., Gojova, A., Zhao, W., Morgan, J. T., Sankaran, S., Sandrock, C. E., & Davis, C. E. (2012). Characterization of Volatile Organic Compounds in Human Leukocyte Antigen Heterologous Expression Systems : A Cell's "Chemical Odor Fingerprint". *ChemBioChem*, *13*(7), 1053-1059. https://doi.org/10.1002/cbic.201200011

Amal, H., Leja, M., Broza, Y. Y., Tisch, U., Funka, K., Liepniece-Karele, I., Skapars, R., Xu, Z., Liu, H., & Haick, H. (2013). Geographical variation in the exhaled volatile organic compounds. *Journal of Breath Research*, *7*(4), 047102. https://doi.org/10.1088/1752-7155/7/4/047102

Barash, O., Peled, N., Tisch, U., Bunn, P. A., Hirsch, F. R., & Haick, H. (2012). Classification of lung cancer histology by gold nanoparticle sensors. *Nanomedicine: Nanotechnology, Biology and Medicine*, *8*(5), 580-589. https://doi.org/10.1016/j.nano.2011.10.001

Boshier, P. R., Priest, O. H., Hanna, G. B., & Marczin, N. (2011). Influence of respiratory variables on the on-line detection of exhaled trace gases by PTR-MS. *Thorax*, *66*(10), 919-920. https://doi.org/10.1136/thx.2011.161208

Broza, Y. Y., & Haick, H. (2013, mai 9). *Nanomaterial-based sensors for detection of disease by volatile organic compounds*. Https://Doi.Org/10.2217/Nnm.13.64. https://doi.org/10.2217/nnm.13.64

Bujang, M. A., & Adnan, T. H. (2016). Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. *Journal of Clinical and Diagnostic Research : JCDR*, *10*(10), YE01-YE06. https://doi.org/10.7860/JCDR/2016/18129.8744

Capuano, R., Santonico, M., Pennazza, G., Ghezzi, S., Martinelli, E., Roscioni, C., Lucantoni, G., Galluccio, G., Paolesse, R., Di Natale, C., & D'Amico, A. (2015). The lung cancer breath signature : A comparative analysis of exhaled breath and air sampled from inside the lungs. *Scientific Reports*, *5*, 16491. https://doi.org/10.1038/srep16491

Chen, X., Xu, F., Wang, Y., Pan, Y., Lu, D., Wang, P., Ying, K., Chen, E., & Zhang, W. (2007). A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. *Cancer*, *110*(4), 835-844. https://doi.org/10.1002/cncr.22844

de Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., Osborne, D., & Ratcliffe, N. M. (2014). A review of the volatiles from the healthy human body. *Journal of Breath Research*, *8*(1), 014001. https://doi.org/10.1088/1752-7155/8/1/014001

Doran, S. L. F., Romano, A., & Hanna, G. B. (2017). Optimisation of sampling parameters for standardised exhaled breath sampling. *Journal of Breath Research*, *12*(1), 016007. https://doi.org/10.1088/1752-7163/aa8a46

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., & Bray, F. (2018). Cancer incidence and mortality patterns in Europe : Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, *103*, 356-387. https://doi.org/10.1016/j.ejca.2018.07.005

Filipiak, W., Mochalski, P., Filipiak, A., Ager, C., Cumeras, R., Davis, C. E., Agapiou, A., Unterkofler, K., & Troppmair, J. (2016). A Compendium of Volatile Organic Compounds (VOCs) Released By Human Cell Lines. *Current Medicinal Chemistry*, *23*(20), 2112-2131. https://doi.org/10.2174/0929867323666160510122913

Gardner, J. W., & Bartlett, P. N. (1999). *Electronic noses, principles and applications*.

Hajian-Tilaki, K. (2014). Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of Biomedical Informatics*, *48*, 193-204. https://doi.org/10.1016/j.jbi.2014.02.013

Herbig, J., Titzmann, T., Beauchamp, J., Kohl, I., & Hansel, A. (2008). Buffered end-tidal (BET) sampling—A novel method for real-time breath-gas analysis. *Journal of Breath Research*, *2*(3), 037008. https://doi.org/10.1088/1752-7155/2/3/037008

Horváth, I., Barnes, P. J., Loukides, S., Sterk, P. J., Högman, M., Olin, A.-C., Amann, A., Antus, B., Baraldi, E., Bikov, A., Boots, A. W., Bos, L. D., Brinkman, P., Bucca, C., Carpagnano, G. E., Corradi, M., Cristescu, S., de Jongste, J. C., Dinh-Xuan, A.-T., … Vink, T. J. (2017). A European Respiratory Society technical standard : Exhaled biomarkers in lung disease. *European Respiratory Journal*, *49*(4), 1600965. https://doi.org/10.1183/13993003.00965-2016

Horvath, I., Lazar, Z., Gyulai, N., Kollai, M., & Losonczy, G. (2009). Exhaled biomarkers in lung cancer. *European Respiratory Journal*, *34*(1), 261-275. https://doi.org/10.1183/09031936.00142508

Huang, C.-H., Zeng, C., Wang, Y.-C., Peng, H.-Y., Lin, C.-S., Chang, C.-J., & Yang, H.-Y. (2018). A Study of Diagnostic Accuracy Using a Chemical Sensor Array and a Machine Learning Technique to Detect Lung Cancer. *Sensors (Basel, Switzerland)*, *18*(9). https://doi.org/10.3390/s18092845

Jia, Z., Patra, A., Kutty, V., & Venkatesan, T. (2019). Critical Review of Volatile Organic Compound Analysis in Breath and In Vitro Cell Culture for Detection of Lung Cancer. *Metabolites*, *9*(3), 52. https://doi.org/10.3390/metabo9030052

Karl, T., Prazeller, P., Mayr, D., Jordan, A., Rieder, J., Fall, R., & Lindinger, W. (2001). Human breath isoprene and its relation to blood cholesterol levels : New measurements and modeling. *Journal of Applied Physiology*, *91*(2), 762-770. https://doi.org/10.1152/jappl.2001.91.2.762

McWilliams, A., Beigi, P., Srinidhi, A., Lam, S., & MacAulay, C. E. (2015). Sex and Smoking Status Effects on the Early Detection of Early Lung Cancer in High-Risk Smokers Using an Electronic Nose. *IEEE Transactions on Biomedical Engineering*, *62*(8), 2044-2054. https://doi.org/10.1109/TBME.2015.2409092

PATHACOV - Diagnostic de pathologies humaines par analyse de COV dans l'air expiré. (2020). [Project website]. PATHACOV. https://pathacov-project.com/

Peled, N., Hakim, M., Bunn, P. A., Miller, Y. E., Kennedy, T. C., Mattei, J., Mitchell, J. D., Hirsch, F. R., & Haick, H. (2012). Non-invasive Breath Analysis of Pulmonary Nodules. *Journal of Thoracic Oncology*, *7*(10), 1528-1533. https://doi.org/10.1097/JTO.0b013e3182637d5f

Pesesse, R. (2019). *Contribution of comprenhensive two-dimensional gas chromatography to untargeted volatilomics of lung cancer*. Université de Liège.

Romain, A.-C., Nicolas, J., & Andre, P. (2002). Three years experiment with the same tin oxide sensor arrays for the identification of malodorous sources in the environment. *Sensors and Actuators. B, Chemical*, *84*. https://doi.org/10.1016/S0925-4005(02)00036-9

Ross, B. M., Maxwell, R., & Glen, I. (2011). Increased breath ethane levels in medicated patients with schizophrenia and bipolar disorder are unrelated to erythrocyte omega-3 fatty acid abundance. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *35*(2), 446-453. https://doi.org/10.1016/j.pnpbp.2010.11.032

Shlomi, D., Abud, M., Liran, O., Bar, J., Gai-Mor, N., Ilouze, M., Onn, A., Ben-Nun, A., Haick, H., & Peled, N. (2017). Detection of Lung Cancer and EGFR Mutation by Electronic Nose System. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, *12*(10), 1544-1551. https://doi.org/10.1016/j.jtho.2017.06.073

Silvestri, G. A., Pastis, N. J., Tanner, N. T., & Jett, J. R. (2016). Clinical Aspects of Lung Cancer. In *Murray and Nadel's Textbook of Respiratory Medicine* (p. 940-964.e22). Elsevier. https://doi.org/10.1016/B978-1-4557-3383-5.00053-1

Sukul, P., Schubert, J. K., Kamysek, S., Trefz, P., & Miekisch, W. (2017). Applied upper-airway resistance instantly affects breath components : A unique insight into pulmonary medicine. *Journal of Breath Research*, *11*(4), 047108. https://doi.org/10.1088/1752-7163/aa8d86

The National Lung Screening Trial Research Team. (2011). Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *New England Journal of Medicine*, *365*(5), 395-409. https://doi.org/10.1056/NEJMoa1102873

Thriumani, R., Zakaria, A., Hashim, Y. Z. H.-Y., Jeffree, A. I., Helmy, K. M., Kamarudin, L. M., Omar, M. I., Shakaff, A. Y. M., Adom, A. H., & Persaud, K. C. (2018). A study on volatile organic compounds emitted by in-vitro lung cancer cultured cells using gas sensor array and SPME-GCMS. *BMC Cancer*, *18*(1), 362. https://doi.org/10.1186/s12885-018-4235-7

Wang, C., Dong, R., Wang, X., Lian, A., Chi, C., Ke, C., Guo, L., Liu, S., Zhao, W., Xu, G., & Li, E. (2014). Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation. *Scientific Reports*, *4*, 7312. https://doi.org/10.1038/srep07312

Westeel, V., Pitard, A., Martin, M., Thaon, I., Depierre, A., Dalphin, J.-C., & Arveux, P. (2007). Negative Impact of Rurality on Lung Cancer Survival in a Population-based Study. *Journal of Thoracic Oncology*, *2*(7), 613-618. https://doi.org/10.1097/JTO.0b013e318074bb96