UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

TESIS DOCTORAL

Non-invasive diagnosis of human diseases by combining breath analysis and neural network modeling

Diagnóstico no invasivo de patologías humanas combinando análisis de aliento y modelización con redes neuronales

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Juan Carlos Cancilla Buenache

Director

José Santiago Torrecilla Velasco

Madrid, 2017

TESIS DOCTORAL

Non-Invasive Diagnosis of Human Diseases by Combining Breath Analysis and Neural Network Modeling

Diagnóstico No Invasivo de Patologías Humanas
Combinando Análisis de Aliento y
Modelización con Redes Neuronales

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

Autor: Juan C. Cancilla Buenache
Director: Prof. Dr. José S. Torrecilla Velasco

Madrid, 2016

Acknowledgements

Firstly, and before getting into expressing my gratitude to my advisor, research group and collaborators, friends, and family, I would like to show great appreciation to the 1171 volunteers that provided samples for this research to be possible. Even though many of them were suffering from serious diseases at the time, they were willing to help us successfully carry out this work and, hopefully, lead to a better future for generations to come. Thank you for your invaluable generosity.

My most sincere gratitude goes to my PhD advisor Prof. José Santiago Torrecilla, who has become much more than my scientific mentor during the past few years. Not only have I learned how to research and develop new ideas thanks to him, but he has made me see that being a good and caring human being is compatible within the scientific research world, which at times seems blinded by resumes and competitiveness. I truly appreciate how he always works non-stop for his research team before himself and I will pursue and encourage this mentality for the rest of my career, wherever I may go. This has made me see him not only as a great leader but, most importantly, as a great person and friend.

I would also like to thank the rest of my research team: AlgoReach. I have learned everything I know in terms of research during the time I have spent in the group, as well as achieved my only and quite satisfying working experience up to date. I would like to specifically thank Mrs. Gemma Matute, the behind-the-scenes boss of the group. She is the glue that has kept us together with her words of knowledge, and has always helped us think outside the box. She is someone who seems to know the right words under any circumstance, willing to always offer a helping hand. Furthermore, I also thank my colleagues and friends Kacper, Regina, Enrique, and Miguel, the present and future of AlgoReach, for having found a home in this small and unselfish group, which, despite its size, is capable of achieving so much due to the hard-work and solidarity found within.

Next, I thank everyone else that has been directly involved some way or another in the research that has been carried out here. I would particularly like to thank Dr. Hossam Haick, whom I had the pleasure to meet in person, for the opportunity of collaborating with him and his group during the course of a European Project (LCAOS), and for inspiring many of the ideas and concepts of this research. I would also like to thank Dr. Nisreen Shehada, who works with Dr. Haick and recently presented her thesis as well, for the many answered messages regarding this research as well as her hospitality when I visited her home country of Israel.

Last but not least, I would like take advantage of this opportunity to acknowledge my family and friends, for their vital spiritual (and sometimes material) support while writing this thesis. Thanks to my mother Mercedes, who has taught me that love, respect, and communication are indispensable for success, and that perseverance and determination can get you anywhere you desire. To my father John, who was my first friend, and has been my role model since I can remember, because of his kindness, intelligence, and, to my eyes, unprecedented wisdom. To my father's wife Ana, for always having respected me and treated me like a son. To my grandparents Mari, Hilario, Josephine, and Myron, for their tireless love and support. To all my friends who have made me who I am: Arthur, Albertito, Marco, Cris, Justine, and Chad. And, finally, to the person that has been by my side every day since I started writing this thesis, on the good days and on the better ones, my girl and best friend, Albertina. I love you all. Thank you.

Index

Summary

Non-Invasive Diagnosis of Human Diseases by Combining Breath Analysis and Neural Network Modeling

It is currently known that there is a direct relation between the moment a disease is detected or diagnosed and the consequences it will have on the patient, as an early detection is generally linked to a more favorable outcome. This concept is the basis of the present research, due to the fact that its main goal is the development of mathematical tools based on computational artificial intelligence to safely and non-invasively attain the detection of multiple diseases. To reach these devices, this research has focused on the breath analysis of patients with diverse diseases, using several analytical methodologies to extract the information contained in these samples, and multiple feature selection algorithms and neural networks for data analysis.

In the past, it has been shown that there is a correlation between the molecular composition of breath and the clinical status of a human being, proving the existence of volatile biomarkers that can aid in disease detection depending on their presence or amount. During this research, two main types of analytical approaches have been employed to study the gaseous samples, and these were cross-reactive sensor arrays (based on organically functionalized silicon nanowire field-effect transistors (SiNW FETs) or gold nanoparticles (GNPs)) and proton transfer reaction-mass spectrometry (PTR-MS). The cross-reactive sensors analyze the bulk of the breath samples, offering global, fingerprint-like information, whereas PTR-MS quantifies the volatile molecules present in the samples.

All of the analytical equipment employed leads to the generation of large amounts of data per sample, forcing the need of a meticulous mathematical analysis to adequately interpret the results. In this work, two fundamental types of mathematical tools were utilized. In first place, a set of five filter-based feature selection algorithms (χ^2 (chi^2) score, Fisher's discriminant ratio, Kruskal-Wallis test, Relief-F algorithm, and information gain test) were employed to reduce the amount of independent in the large databases to the ones which contain the greatest discriminative power for a further modeling task. On the other hand, and in relation to mathematical modeling, artificial neural networks (ANNs), algorithms that are categorized as computational artificial intelligence, have been employed. These non-linear tools have been used to locate the relations between the independent variables of a system and the dependent ones to fulfill estimations or classifications. The type of ANN that has been used in this thesis coincides with the one that is more commonly employed in research, which is the supervised multilayer perceptron (MLP), due to its proven ability to create reliable models for many different applications.

Having presented the scope of this research, as well as the tools employed, it is time to briefly look into the four main experimental sections that have been carried out and their results. The first one is a preliminary study in which a series of artificial gaseous mixtures were prepared with low concentrations of 11 known polar and non-polar volatile organic compounds (VOCs), chemically similar to molecules found in human breath. Pure samples of each VOC were prepared, as well as binary and ternary mixtures, and they were analyzed using an array of functionalized cross-reactive SiNW FET sensors. These sensors provided signals which were mathematically treated and interpreted using MLPs in an attempt to identify and quantify the

VOCs in the samples. The results of this study were promising, as all the compounds were perfectly identified, and the mean prediction error during their quantification was below 1.5%. These results proved the existence of a relation between the signals provided by the sensors and the chemical composition of the gaseous samples, validating the combination of the selected chemical and mathematical tools, as well as opening the door for the analysis of biological breath samples using this approach.

In the second experiment of this research, a study was carried out using the same type of sensors as the previous ones, but this time, real human breath samples were analyzed. In this case, samples from patients of different diseases were gathered, and a set of binary classifiers based on MLPs were designed to distinguish samples from each group. Samples were obtained from lung cancer, gastric cancer, chronic obstructive pulmonary disease, and asthma patients, as well as others from healthy controls, and they were processed with SiNW FET sensors. Once the data was available, it was treated and employed to design models which would distinguish the diseases from each other, as well as from the healthy control group. The results ranged from 80% to 100% correct classification rate (100% was attained during a series of cross-validations when discriminating lung cancer patients from gastric cancer ones, and gastric cancer patients from chronic obstructive pulmonary disease or asthma ones), using information from a single sensor of the array. With these results, it can be confidently stated that the composition of breath is in fact related to the clinical status of a person, as only using information extracted from these gaseous samples it is possible to accurately classify the people according to the diseases. Additionally, the combination of these sensors and ANNs for breath analysis is confirmed for diagnostic purposes.

Next, during the third experiment, PTR-MS was employed to analyze the breath samples of a group of lung cancer patients and compare them to others from healthy controls, yet high-risk individuals, with the goal set to classify the samples using MLPs. This quantitative approach enables the location of potential volatile biomarkers that can aid in the detection of lung cancer (in this case, feature selection algorithms were employed to locate those compounds that possess the greatest discriminative power to distinguish the samples from both groups). The breath samples were collected during an oral glucose tolerance test, gathering two samples per volunteer, one before and one after the glucose uptake. This permitted two different studies where the samples can be analyzed regardless of glucose consumption, or considering it. In the first case, irrespective of glucose uptake, the classifying models would distinguish the breath samples without taking into account if the person had consumed glucose or not. These models offered around a 94% correct classification rate. On the other hand, if the influence of glucose is considered, compounds which are potentially affected by the Warburg effect could be identified, as a change in their concentration would not only be determined by the presence of cancer, but by the altered glucose metabolism of the patient as well. In this case, the classification rate was approximately 90%. Therefore, during this experiment, two main goals have been achieved, as an accurate and non-invasive tool has been developed to detect lung cancer, and the location of potential volatile lung cancer biomarkers has been enabled (such as acetic acid, ethylbenzene, 1,2-dichlorobenzene, and glutamic acid).

Finally, in the fourth and last experiment, a set of 34 functionalized cross-reactive GNP sensors were used to analyze the breath samples from patients of seven different diseases, as well as healthy controls for each group which were matched in terms of age, gender, and smoking history. The diseases analyzed were chronic kidney disease, head and neck cancer, inflammatory bowel disease, multiple sclerosis, Parkinson’s disease, preeclampsia, and pulmonary arterial

hypertension. The objective of this study was the design of binary classifiers that would distinguish samples from patients from others which belonged to their matched healthy controls, in order to reach disease detecting tools. In this case, the mathematical treatment phase also allowed the identification of the sensors that led to the best possible classifications, enabling the development of more cost-effective and specialized tools for specific biomedical sectors. The statistical performance of the models in terms of correct classification rate ranged from about 80% for the detection of multiple sclerosis to over 90% for head and neck cancer and pulmonary arterial hypertension during a cross-validation procedure. On the other hand, all the models were above 80% classification rate during an internal validation which used an independent set of samples to test them. These results reveal and further confirm the relation between the clinical status of a person and the composition of his or her breath, and, furthermore, reflect the huge potential behind the combination of cross-reactive sensors and neural networks, as these algorithms have been able to locate specific patterns in the signals originated by the breaths of patients of such a broad span of diseases.

To sum up, with these experiments and the results they have provided, it has been possible to demonstrate the undoubtable correlation between the clinical status of a human being and the composition of their exhaled breath. The door has been opened for the design of countless tools to reach an early, safe, non-invasive, accurate, and reliable detection of multiple diseases, which could have a priceless repercussion over many biomedical sectors worldwide, giving a relevant support to current techniques and, most importantly, saving lives and improving their quality.

Resumen en Castellano

Diagnóstico No Invasivo de Patologías Humanas Combinando Análisis de Aliento y Modelización con Redes Neuronales

Actualmente es sabido que existe una relación directa entre el momento en el cual se detecta o diagnostica una enfermedad y las consecuencias que tendrá sobre el paciente, ya que una detección temprana va generalmente ligada a un desarrollo más favorable. Este concepto es el cimiento de la presente investigación, cuyo objetivo fundamental es el desarrollo de herramientas basadas en inteligencia artificial computacional que consigan, mediante medios seguros y no invasivos, la detección de diversas enfermedades. Para alcanzar dichos sistemas, los estudios han sido enfocados en el análisis de muestras de aliento de pacientes de diversas enfermedades, empleando varias técnicas para extraer información, y diversos algoritmos de selección de variables y redes neuronales para el procesamiento matemático.

En el pasado, se ha comprobado que hay una correlación entre la composición molecular del aliento y el estado clínico de una persona, evidenciando la existencia de biomarcadores volátiles que pueden ayudar a detectar enfermedades, ya sea por su presencia o por su cantidad. Durante el transcurso de esta investigación, se han empleado esencialmente dos tipos de técnicas analíticas para estudiar las muestras gaseosas, y estas son conjuntos de sensores de reactividad cruzada (basados en transistores de efecto de campo con nanocables de silicio (SiNW FETs) o en nanopartículas de oro (GNPs), ambos funcionalizados con cadenas orgánicas) y equipos de reacción de transferencia de protones con espectrometría de masas (PTR-MS). Los sensores de reactividad cruzada analizan el aliento en su conjunto, extrayéndose información de la muestra global, mientras que usando PTR-MS, se cuantifican las moléculas volátiles presentes en las muestras analizadas.

Todas las técnicas empleadas desembocan en la generación de grandes cantidades de datos por muestra, por lo que un análisis matemático exhaustivo es necesario para poder sacar el máximo rendimiento de los estudios. En este trabajo, se emplearon principalmente dos tipos de herramientas matemáticas. Las primeras son un grupo de cinco algoritmos de selección de variables, concretamente, filtros de variables (cálculos basados en estadística de χ^2 (chi^2), ratio discriminante de Fisher, análisis de Kruskal-Wallis, algoritmo relief-F y test de ganancia de información), que se han empleado en las bases de datos con grandes cantidades de variables independientes para localizar aquellas con mayor importancia o poder discriminativo para una tarea de modelización matemática posterior. Por otro lado, en cuando a dicha modelización, se ha empleado un tipo de algoritmo que se cataloga dentro del área de la inteligencia artificial computacional: las redes neuronales artificiales (ANNs). Estas herramientas matemáticas de naturaleza no lineal se han utilizado para localizar las relaciones existentes entre las variables independientes de un sistema y las variables dependientes o parámetros a estimar o clasificar. Se ha empleado el tipo de ANN supervisada más extensamente usado en investigación, que son los perceptrones multicapa (MLPs), debido a su habilidad contrastada para originar modelos fiables para numerosas aplicaciones.

Habiendo presentado la temática y las herramientas empleadas en la presente investigación, ahora se explicarán los cuatro bloques experimentales que se han desarrollado y

los resultados obtenidos en los mismos. El primero se trata de un estudio preliminar en el que se prepararon una serie de muestras gaseosas artificiales con concentraciones bajas y conocidas de 11 compuestos orgánicos volátiles (VOCs) polares y no polares, de naturaleza similar a compuestos que se encuentran en el aliento humano. Se prepararon muestras puras además de mezclas binarias y ternarias, y se analizaron empleando conjuntos de sensores de reactividad cruzada basados en SiNW FETs funcionalizados. Estos sensores originaban señales que fueron tratadas matemáticamente e interpretadas a través de MLPs para intentar identificar y cuantificar los VOCs de las muestras. Los resultados de este estudio fueron satisfactorios, ya que todos los compuestos volátiles fueron perfectamente identificados, y el error de la cuantificación en términos de error de predicción medio siempre fue inferior al 1,5%. Estos resultados demuestran la existencia de una relación entre las señales originadas por los sensores y la composición química de las muestras gaseosas, dando validez a la combinación de las herramientas químicas y matemáticas seleccionadas, además de representar resultados prometedores a la hora de enfrentarse a muestras biológicas de aliento.

En segundo lugar, se llevó a cabo un estudio empleando sensores de la misma naturaleza que los del primer experimento, pero para el análisis de muestras de aliento reales. En este caso, se recogieron muestras de una serie de pacientes de diferentes enfermedades, y se diseñaron clasificadores binarios basados en MLPs para diferenciar muestras de cada grupo. Se obtuvieron muestras de pacientes que padecían cáncer de pulmón, cáncer gástrico, enfermedad pulmonar obstructiva crónica o asma, además de una serie de muestras de individuos sanos o controles, y se analizaron con los sensores de SiNW FET. Una vez obtenidos los datos, se procesaron y se emplearon para diseñar modelos que distinguiesen las enfermedades entre sí y de los controles, y los resultados oscilaron entre un 80% de acierto en los peores casos hasta el 100% (el 100% se consiguió durante una serie de validaciones cruzadas al distinguir casos de cáncer de pulmón de otros de cáncer gástrico, y al diferenciar casos de cáncer gástrico de otros de enfermedad pulmonar obstructiva crónica o asma), llegándose a usar información de un solo sensor del conjunto. Con estos resultados se pone de manifiesto la relación entre el estado clínico de un paciente y la composición de su aliento, ya que solamente usando datos obtenidos del aliento se pueden clasificar las muestras de diferentes enfermedades de forma precisa. Asimismo, se confirma la valía de la combinación de los sensores empleados con las redes neuronales para el análisis de muestras de aliento reales con fines diagnósticos.

En el tercer experimento se empleó PTR-MS para procesar las muestras de aliento de un grupo de pacientes de cáncer de pulmón y compararlas a otras pertenecientes a controles sanos, pero con elevado riesgo a desarrollar dicha enfermedad, con el objetivo de clasificar las muestras empleando MLPs. Al emplear esta técnica analítica cuantitativa, se permite la localización de potenciales biomarcadores volátiles que puedan ayudar en la detección de cáncer de pulmón (en este caso, se emplearon los algoritmos de selección de variables para localizar aquellos compuestos que poseen mayor poder para diferenciar las muestras de los dos grupos). Las muestras de aliento se recogieron durante el transcurso de un test oral de tolerancia a la glucosa, obteniéndose muestras de cada voluntario antes y después de la ingesta de glucosa. Esto permitió un estudio dual, donde se tratan las muestras de forma individual sin importar la ingesta de glucosa o considerando la misma. En el primer caso, no considerando la glucosa, se obtendrían modelos que usan datos de compuestos que diferencian pacientes de controles independientemente de si han ingerido glucosa o no. En estos modelos se obtuvieron precisiones del 94% aproximadamente en cuanto a la clasificación de muestras. Por otro lado, si se considera la influencia de la glucosa, se pueden potencialmente localizar compuestos que se ven afectados por el efecto Warburg, ya

que el cambio en su concentración en aliento viene determinada no solo por la presencia del cáncer, sino también por el metabolismo de la glucosa alterado. En este caso, el rendimiento estadístico de los modelos ronda el 90% de acierto. Por lo tanto, este experimento cumple un doble propósito, ya que se ha desarrollado una herramienta precisa para la detección de cáncer de pulmón de forma no invasiva, y se permite la localización de potenciales biomarcadores de dicha enfermedad (como el ácido acético, etilbenceno, 1,2-diclorobenceno o ácido glutámico).

Finalmente, en el cuarto y último experimento se empleó un conjunto de 34 sensores de reactividad cruzada basados en GNPs funcionalizadas para analizar muestras de aliento provenientes de pacientes de siete enfermedades diferentes, además de muestras de controles para cada enfermedad que se hicieron coincidir en términos de edad, género e historial fumador. Las enfermedades analizadas fueron enfermedad renal crónica, cáncer de cabeza y cuello, enfermedad inflamatoria intestinal, esclerosis múltiple, enfermedad de Parkinson, preeclampsia e hipertensión arterial pulmonar. El objetivo del estudio fue el diseño de clasificadores binarios que diferenciasen las muestras de enfermedades concretas de su grupo correspondiente de controles para alcanzar herramientas de detección o diagnóstico de enfermedades. En este caso, el tratamiento matemático de los datos también permitió la identificación de aquellos sensores que mejor permitían llevar a cabo las clasificaciones de los pacientes y sus respectivos controles, pudiendo reducir costes en el desarrollo potencial de herramientas para sectores específicos de la biomedicina. Los resultados de los modelos en cuanto a porcentajes de acierto oscilaban entre el 80% aproximado para esclerosis múltiple y por encima del 90% para la detección de cáncer de cabeza y cuello e hipertensión arterial pulmonar, durante una validación cruzada. Por otro lado, con una validación interna, empleando muestras separadas de la base de datos, los modelos siempre superaban el 80% de aciertos, lo cual pone de manifiesto la relación entre el estado clínico y la composición del aliento, además de reflejar la inmensa potencialidad de la combinación de sensores de reactividad cruzada y redes neuronales, ya que se han podido localizar con estos algoritmos patrones propios de enfermedades tan variadas en las señales proporcionadas por los sensores.

En resumen, con esta serie de experimentos y sus resultados, se ha podido demostrar que hay una relación innegable entre el estado clínico de un ser humano y la composición del aire que exhala. Se ha abierto la puerta al desarrollo de cuantiosas herramientas para alcanzar una detección precoz, segura, no invasiva, precisa y fiable de diversas enfermedades, que podrían tener una repercusión incalculable sobre numerosos sectores de la biomedicina a nivel mundial, sirviendo de apoyo a las técnicas actuales y ayudando a salvar vidas y mejorar la calidad de las mismas.

Abbreviations

AD: Alzheimer's disease
ANN: artificial neural network
APTES: 3-aminopropyl-triethoxysilane
AS: asthma
CD: Crohn's disease
CKD: chronic kidney disease
COPD: chronic obstructive pulmonary disease
DFA: discriminant factor analysis
ED: Euclidean distance
FET: field-effect transistor
FS: feature selection
GC: gastric cancer
GC-MS: gas chromatography-mass spectrometry
GFR: glomerular filtration rate
GNP: gold nanoparticle
HNC: head and neck cancer
HNN: hidden neuron number
HPV: human papillomavirus
IBD: inflammatory bowel disease
I_{ds}: source-drain current
LC: lung cancer
Lc: Marquardt adjustment parameter
Lcd: decrease factor for Marquardt adjustment parameter
Lci: increase factor for Marquardt adjustment parameter
LNBD: laboratory for nanomaterial based devices
m/z: mass charge ratio
μ_h: hole mobility
MLP: multilayer perceptron
MOSFET: metal-oxide-semiconductor field-effect transistor
MPE: mean prediction error
MS: multiple sclerosis
OGT: oral glucose tolerance
p_a: partial pressure
PAH: pulmonary arterial hypertension
PD: Parkinson's disease
PE: preeclampsia
p_o: vapor pressure
PTR-MS: proton transfer reaction-mass spectrometry
SEM: standard error of the mean
SIFT-MS: selected ion flow tube-mass spectrometry
SiNW: silicon nanowire
SOM: self-organizing maps
SS: subthreshold swing
SVM: support vector machine
THF: tetrahydrofuran
trainBR: Bayesian regularization
trainLM: Levenberq-Marquardt back-propagation
UC: ulcerative colitis
V_g: gate voltage
VOC: volatile organic compound
V_{th}: voltage threshold

Non-Invasive Diagnosis of Human Diseases by Combining Breath Analysis and Neural Network Modeling

1) Introduction

It is well known that the early detection or diagnosis of many and diverse diseases is highly correlated with successful treatments which lead to more favorable patient prognosis (Befeler and Di Bisceglie, 2002; O'Sullivan and Freedman, 2009). The research behind the present thesis is completely and utterly influenced by this concept, as it is focused on attaining novel, non-invasive diagnosing tools, which combine breath tests and intelligent non-linear computational modeling. In many cases, diseases and syndromes develop and pass unnoticed and are asymptomatic for determined time periods that are likely to be crucial for the patient's outcome. The goal that has been set during this study is to reach fast, safe, and reliable systems that can potentially help lower the mortality rate of terrible diseases such as cancer by consistently detecting the diseases in the earliest stages possible. Discovering diseases early also facilitates the following and necessary treatments or procedures, greatly increasing the quality-of-life of the patients. Therefore, not only would survival rates increase, but also the wellbeing of cured people or patients under treatment.

In the following subsections of the introduction, the background of the most relevant elements of this research will be revealed and sensibly linked together to fully understand the thought process behind its development. In the first phase, the history of breath analysis and breath tests will be uncovered, and how they can be used to determine and relate volatile biomarker profiles with different diseases and even disease subtypes. It will be backed up by the general developments and progress within the biomarker field, which will guide the definition of the wide set of diseases included in this research. Next, the analytical methodologies employed for the preliminary and breath tests will be shown as well as their past and present applications. Finally, the data analysis process followed will be thoroughly displayed and explained as well as how it has enabled the design of useful mathematical tools and models for many fields within the biomedical sector for disease detection or diagnosis.

1.1) Breath Tests – Exhaled Information

Tens of centuries back, at the time of Ancient Greece, particularly, during the Classical period, the father of western medicine, Hippocrates of Kos (c.460-c.370 BC), was already starting to consider that human breath potentially contained evidence concerning a person's medical status and physiology (Grammaticos and Diamantis, 2008; Risby and Solga, 2006). For instance, he understood that determined breath smells such as a sweet scent or a urine odor could be representative of diabetes or kidney failure, respectively (Phillips, 1992). Many years later, during the late 18th Century, Antoine Lavoisier demonstrated that carbon dioxide could be found in exhaled breath and he interpreted that the production of this metabolic byproduct was due to a process of slow combustion that takes place in the human body (Amann *et al.*, 2014). In this section, we will look into the background of breath and breath analysis as well as its current applications and involvement in the biomedical field and, specifically, how its study enables the diagnosis of diseases.

1.1.1) Breath – A Volatile Tattletale

Breath can be defined as a gaseous matrix that mainly contains nitrogen, oxygen, carbon dioxide, water vapor, inert gases, and very small, perhaps seemingly insignificant, amounts of volatile organic compounds (VOCs) (Cao and Duan, 2007) (**Figure 1**). Nevertheless, these "insignificant" VOCs are the key components that provide underlying clues behind a human being's medical status. A few decades ago, in 1971, the Nobel Prize winning Linus Pauling and his research team quantified around 250 compounds in breath using gas-liquid partition chromatography (Pauling *et al.*, 1971; Teranish *et al.*, 1972). Most of these detected molecules were the mentioned VOCs, and they are the reason why it is possible to link the story that breath analysis tells us with a person's physiology (**Figure 1**).

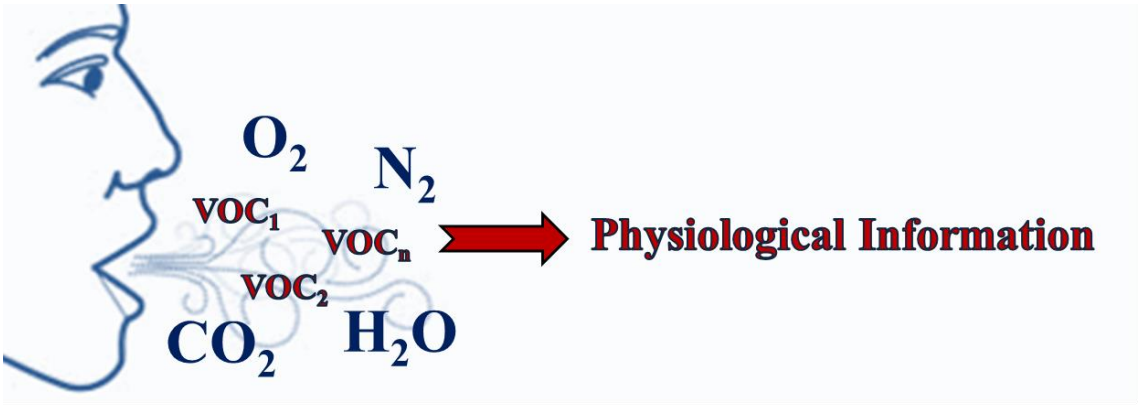

Figure 1. Main molecular components of exhaled breath. The small amounts of VOCs present in breath provide physiological information from the person offering potential diagnostic evidence.

It is currently known that there are over 1000 trace VOCs present in human breath, with concentrations varying from about 100 ppm all the way to the ppt range (Cao and Duan, 2007; Risby and Solga, 2006). They vary qualitatively and quantitatively from person to person and only a limited set of them (e.g., isoprene, acetone, ethane, pentane, ethanol, methanol, and other alcohols; *vide infra*) are common to everyone and are the result of essential metabolic processes (Mukhopadhyay, 2004; Sánchez and Sacks, 2003). VOCs reach the exhaled breath deep within the lungs, from the insides of the alveoli. "Alveolar breath" is the air that has suffered the gaseous exchange process with the blood, which is the reason why these VOCs end up in exhaled air and act as a true reflection of what is happening inside our bodies (Mukhopadhyay, 2004).

The VOCs in breath are subdivided into molecules that possess an endogenous or an exogenous origin. This subdivision is relevant due to the fact that the information provided solely by the endogenous VOCs, many of which are common to all people, is what actually provides a clinical perspective and diagnosing hints. On the other hand, the exogenous VOCs simply interfere with these results, setting clear hurdles for the analyses of breath tests, as many confounding factors appear. Unlike endogenous VOCs, the sources of exogenous ones are inhaled air, food ingestion, tobacco use, or even compounds like anesthetics that remain in the body up to six weeks after their use (Risby and Solga, 2006). Furthermore, compounds present in ambient air may interfere with the breath test results, thus necessary precautions and corrections must be considered (Risby and Solga, 2006). For these reasons, revealing and adequately interpreting the biomedical tale that is being presented by the endogenous VOCs in breath is not at all an easy or straightforward task.

Therefore, the real clinical information in breath is offered by the presence and the concentration of endogenous VOCs, as they mirror metabolic alterations or malfunctions that take place within the body (Hubbard *et al.*, 2009). The biochemical origin of these compounds, which can be used for diagnosing purposes, lies in normal or irregular metabolic processes, such as inflammation or oxidation damage (Smolinska *et al.*, 2014). The main groups of endogenous VOCs, and examples of each one of them, can be seen in **Table 1** (Buszewski *et al.*, 2007; Miekisch *et al.*, 2004). In order to understand their physiological meaning and evaluate their relevance in disease diagnosis and detection, it is important to locate the metabolic pathways in which their generation or presence is involved (**Table 1**) (Buszewski *et al.*, 2007; Miekisch *et al.*, 2004).

Table 1. Main endogenous VOCs in exhaled breath. Examples of each main group are shown, as well as their metabolic pathway of origin (Buszewski et al., 2007; Miekisch *et al.*, 2004). The numbers in bold are used to correlate the molecules with the specific metabolic pathway and clinical importance.

Group of compounds	Examples	Metabolic pathway	Clinical relevance
Saturated hydrocarbons	**1**) Ethane [a,b] **2**) Pentane [a,b]	**1** & **2**) Lipid peroxidation [a,b]	**1** & **2**) Control of oxidative damage [b]
Unsaturated hydrocarbons	**1**) Isoprene [a,b]	**1**) Cholesterol biosynthesis [a,b]	**1**) Cholesterol-related disorders [a] **1**) Control of oxidative damage [b]
Oxygen-containing compounds	**1**) Acetone [a,b] **2**) 2-Propanol [b] **3**) Acetaldehyde [a,b] **4**) Ethanol [b] **5**) Methanol [b]	**1**) Decarboxylation of acetoacetate and acetyl-CoA [a,b] **2**) Reduction of acetone [b] **3**) Oxidation of endogenous ethanol [b] **4** & **5**) Intestinal bacterial flora [b]	**1**) Diabetes mellitus [a,b] **1**) Nutritional issues [a] **1**) Ketonemia [a]
Sulphur-containing compounds	**1**) Ethyl mercaptane [a,b] **2**) Dimethylsulfide [a,b] **3**) Dimethyldisulfide [a,b]	**1**, **2**, & **3**) Metabolism of methionine; transamination pathway [a,b]	**1**, **2**, & **3**) Impairment of liver function [a,b]
Nitrogen-containing compounds	**1**) Ammonia [a,b] **2**) Dimethylamine [a,b] **3**) Trimethylamine [a,b]	**1**) Conversion to urea [a,b]	**1**, **2**, & **3**) Uremia and kidney and liver impairment [a,b]

[a] Buszewski et al., 2007.
[b] Miekisch *et al.*, 2004.

The information collected in **Table 1** clearly demonstrates the biomedical potential behind breath analysis. With only a few molecules it is already possible to assess a relevant span of diseases or general altered processes, which allow narrowing down the type of clinical irregularities. If we were able to develop a methodology that was capable of suitably analyzing the entire bulk of exhaled air (over 1000 trace VOCs), the potential applications would be countless. In the next subsection, the current applications of breath tests will be shown, ranging from well-known examples such as the alcohol measuring breathalyzer, to more specific devices to fulfill biomedical needs.

1.1.2) Breath Tests – Disease Diagnosis & Other Applications

Breath gas analysis is currently carried out for a wide variety of purposes. The implementation of specific applications that rely on breath analysis began when, as mentioned previously, Antoine Lavoisier discovered amounts of carbon dioxide in human exhaled air in the late 1700s (Amann *et al.*, 2014). This led to the birth of the first breath test-based application: capnography (Amann *et al.*, 2014). Capnometry measurements, which are used to monitor the amount of carbon dioxide in breath, are used by clinicians to gather information from the systemic metabolism of the patient, as well as data from both circulatory and respiratory systems (Sanders, 1989).

Since then, many other applications have developed to become regularly employed tests. For instance, the common breathalyzer used to determine the amount of alcohol in blood through indirect ethanol measurements in breath. It has been around since being invented by Robert Frank Borkenstein (Martin, 2002) back in 1958 (Borkenstein, 1958). Its most known application is for law enforcement during traffic controls, to locate inebriated drivers.

On the other hand, more focused on the theme of the present work, a wide variety of breath tests exist which are medically oriented and serve as disease detectors and evaluators (Kim *et al.*, 2012). There are multiple compounds in breath that have shown to be linked to the existence of different significant groups of diseases. For example, breath tests are employed to diagnose lung diseases (e.g., asthma, chronic obstructive pulmonary disease (COPD), cystic fibrosis, bronchiectasis, interstitial lung disease, obstructive sleep apnea, and pneumonia), metabolic disorders (e.g., diabetes), and gastroenteric diseases (e.g., lactase deficiency, starch malabsorption, *Helicobacter pylori* infection, lactose and fructose intolerance, bacterial overgrowth, bile salt wastage, pancreatic insufficiency, liver dysfunction, and abnormal small-bowel transit), as well as to evaluate the state of oxidative stress in the body (Kim *et al.*, 2012).

Many of the above diagnosing breath tests are based on the measurement of specific endogenous VOCs. For instance, asthma and COPD, as well as other inflammatory diseases, can be related to a characteristic increase in the amount of exhaled saturated hydrocarbons, such as pentane and ethane (Miekisch *et al.*, 2004; Paredi *et al.*, 2000-a; Paredi *et al.*, 2000-b). Additionally, particular amounts of nitric oxide (not a VOC) have shown determined correlations with the state and severity of these diseases (Clini *et al.*, 1998; Pijnenburg and De Jongste, 2008). Other examples are listed in **Table 1**, where determined VOCs appear linked to their clinical relevance and/or disease. On the other hand, a different well-known test is the urea breath test to diagnose and monitor *H. pylori* infection. This test does not measure an endogenous VOC as it is based on the fact that these Gram-negative bacteria can transform urea into ammonia and carbon dioxide through the enzymatic activity of urease. Patients ingest urea labeled with an uncommon carbon isotope, and around 10 to 30 minutes later, the amount of labeled carbon dioxide in exhaled breath is determined to confirm or monitor this infection (Chey and Wong, 2007; Malfertheiner *et al.*, 2002).

As can be reasoned from the above paragraphs, metabolic changes that are induced from different diseases are to some extent reflected in the molecular profile in breath (Miekisch *et al.*, 2004; Smolinska *et al.*, 2014). This is the key behind the search of breath-based methodologies to diagnose diseases. The goal is clear: to link the information contained in breath to particular diseases through non-invasive procedures to attain safe, reliable, and real-time diagnosing tools (Peng *et al.*, 2008). Therefore, the premise here is that a specific average molecular breath profile

will exist for determined diseases, which will allow differentiating it from other pathological developments and from healthy people, and, ideally, even different stages of the same disease. These molecules that allow diagnosing diseases are known as biomarkers (Spanel and Smith, 2011), which will be analyzed in the following section.

1.2) Biomarkers – A Safe Highway to Disease Diagnosis

A proper way to begin understanding the clinical relevance of biomarkers and the role they play in disease detection and early diagnosis is by comprehending their definition. The following was provided by a working group from the National Institutes of Health Director's Initiative on Biomarkers and Surrogate Endpoints (Biomarkers Definitions Working Group, 2001):

> "***Biological marker (biomarker)****: A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*".

In other words, biomarkers are molecules that, depending on their presence or amount in a determined body fluid, can indicate the existence of a disease or even predict its development (Wang *et al.*, 2006) and, for instance, assess the extent of the efficiency of a specific treatment or surgical procedure that a patient is undergoing or has undergone, respectively (Biomarkers Definitions Working Group, 2001). Therefore, biomarkers are compounds which hide behind their presence and amount significant clinical information that can potentially save lives. For this reason, it is crucial to design reliable methods that allow the analysis of biological matrices to find and extract this kind of information, with the goal set to catch diseases as early as possible and greatly improve their outcome.

The concept of biological markers appeared in the early 1980s, where research revealed a series of tumor markers in human biological body fluids (Johnson *et al.*, 1984; Paone *et al.*, 1980; Winters, 1983). Since then, a variety of applications related to disease diagnosis and health status monitoring have emerged through biomarker analysis. For example, it is possible to locate people suffering a particular pathological process, reflect or stage the extent of diseases, provide an approximate prognosis, or predict and monitor the clinical development after a medical intervention (Biomarkers Definitions Working Group, 2001). If studied and validated adequately, these biological markers represent a fast and safe diagnosing alterative, possibly limiting the amount of required biopsies. This correct biomarker validation is necessary from the moment of discovery all the way through the required preclinical and clinical trials (Anonymous, 2010). In order for a molecule to become a biomarker, it must meet three requirements: (**a**) must exist in peripheral body tissue and/or fluid such as blood or serum (Kosaka *et al.*, 2010; Maurya *et al.*, 2007), urine (Pisitkun *et al.*, 2006), saliva (Pfaffe *et al.*, 2011), or exhaled breath (Shirasu and Touhara, 2011); (**b**) must be affordably and robustly detectable and quantifiable; (**c**) its presence must be specifically linked to damage in a particular tissue, preferably in a quantifiable fashion (Anonymous, 2010).

Biomarkers, as mentioned before, can be present in a variety of easily accessible body fluids, enabling safe disease diagnosis. Additionally, the chemical nature of these molecules is not restricted to a determined kind of molecule, as a wide variety of them have been described, ranging from proteins (Xiao *et al.*, 2005) to small microRNA molecules (Kosaka *et al.*, 2010) or

even cell-free DNA (Jung *et al.*, 2010). A series of clear examples of biomarkers and their clinical relevance can be seen in **Table 2**, manifesting the broad span of possibilities behind these studies.

Table 2. Biomarker examples from diverse research studies and groups linked to their biochemical nature, biological matrix analyzed, and clinical importance.

Biomarker/s	Biochemical Nature	Body Fluid Analyzed	Clinical Relevance	Reference
Acetone	VOC	Breath	Correlated with diabetes	Di Francesco *et al.*, 2005
α-Fetoprotein	Protein	Serum	Detection of liver cancer	Beneduce *et al.*, 2004
Aquaporin-2	Integral membrane protein	Urine	Related to impaired water excretion	Ishikawa and Schrier, 2003
Cell-free DNA	DNA	Serum	Myocardial infarction marker	Chang *et al.*, 2003
C-reactive protein, Myoglobin & Myeloperoxidase	Proteins	Saliva	Early detection of acute myocardial infarction	Floriano *et al.*, 2009
Cystatin C	Protein	Serum	Glomerular filtration rate marker	Dharnidharka *et al.*, 2002
Isoprene	VOC	Breath	Marker for cholesterol metabolism disorders	Buszewski *et al.*, 2007
miR-92	microRNA	Plasma	Colorectal cancer marker	Ng *et al.*, 2009
miR-210	microRNA	Plasma	Detection of pancreatic cancer	Ho *et al.*, 2010
Carbonic anhydrase VI (protein) & set of 8 mRNAs*	mRNA & protein	Saliva	Detection of breast cancer	Zhang *et al.*, 2010

*The eight mRNAs are: CSTA, TPT1, IGF2BP1, GRM1, GRIK1, H6PD, MDM4, and S100A8.

As can be noticed, the possibilities are immense for safe and non-invasive biomarker-based diagnosis or disease monitoring. Options are broad in terms of both biological fluids and biochemical nature of the markers, greatly widening the potentiality of these kinds of clinical studies. Nevertheless, it has been reported that diagnosis based on single biomarkers suffer from low sensitivity (true positive rate) and low specificity (true negative rate) values, giving a clear advantage to studies based on the analysis of biomarker panels (Kozak *et al.*, 2003).

In the present thesis, the analysis will be centered around exhaled breath and volatonomics (Shehada *et al.*, 2015; Vishinkin and Haick, 2015), based on the well-known concept that the average molecular profiles of exhaled volatile compounds (volatile biomarker panel) vary between patients suffering different diseases and from healthy people (Amman *et al.*, 2014; Buszewski *et al.*, 2007). As mentioned several times already, the volatile profile of exhaled breath acts as a reflection of the clinical status of a human being. These volatile compounds that

can be related to different diseases are great examples of biomarkers, with the clear advantage of being present in such an accessible and safe biological matrix. When perfected, breath analysis for disease detection, early diagnosis, and evaluation will offer an invaluable source of non-invasive and real-time biomedical tools (Peng *et al.*, 2008). Therefore, attaining reliable breath-based diagnosing systems would clearly be devices that many doctors would gladly include into their disease-combating resources.

In the current research breath samples have been treated as a whole, not so much focusing on the search for specific biomarkers for specific diseases, but as a complex matrix with hundreds or even thousands of volatile compounds that can offer a general profile or pattern to distinguish and identify diseases (Cao and Duan, 2007; Risby and Solga, 2006). This approach, rather than seeking for particular evidence from determined molecules, attempts to extract relevant information from the big picture, from the complete exhaled sample, to then link it with a particular disease, thus taking advantage of what a biomarker panel provides, instead of non-sensitive and non-specific single biomarker methods. In other words, we are seeking for the entire story that breath is willing and able to tell us, not partial bits and pieces. To understand this approach, it can be compared with the mechanism of our own olfactory system, which enables us to identify smells (more specifically, from 4 to 10 thousand different ones) as they are integrated by our brain and treated as particular patterns, without necessarily knowing the individual molecules that lead to that smell (Boots *et al.*, 2012).

The present work has been centered on the detection of multiple and very dissimilar diseases through the analysis of exhaled breath or volatolome. This set of diseases will be looked into in the following section of this introduction.

1.3) Diseases Analyzed – Exhale & Detect

The breaths of a broad series of patients with different diseases have been analyzed to try and discover disease-specific patterns that can potentially allow their diagnosis and detection. This study is based on the fact that as all diseases affect and alter the "normal" metabolism, it will also specifically change the composition (presence and/or concentration) of volatile biomarkers in a patient's breath. The main study has been focused around lung cancer (LC), as early detection greatly improves its otherwise terrible prognosis and recovery rate, and is currently one of the main health concerns and deadly diseases (Tisch *et al.*, 2012). In addition, other diseases have also been looked into in order to attain fast and non-invasive diagnosing devices. These are asthma (AS), chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), gastric cancer (GC), head and neck cancer (HNC), inflammatory bowel disease (IBD), multiple sclerosis (MS), Parkinson's disease (PD), preeclampsia (PE), and pulmonary arterial hypertension (PAH). The wide variety of diseases assessed through breath tests are gathered in **Figure 2**.

As can be deduced, successfully attaining reliable breath-based tools that can identify such diverse diseases at early stages safely, quickly, and non-invasively would imply an explosion of new devices for a large variety of health fields, potentially facilitating a trustworthy disease-screening method. If proven suitable, these systems, combined with current techniques, would enable a completely different yet complementary alternative for disease diagnosis, clearly lowering detection and treatment time as well as increasing patient safety and satisfaction. Furthermore, early detection of diseases would allow reducing the global treatment costs, which,

for example, for LC is quite high, as its economic burden represents tens of thousands of euros per patient in Europe (McGuire *et al.*, 2015).

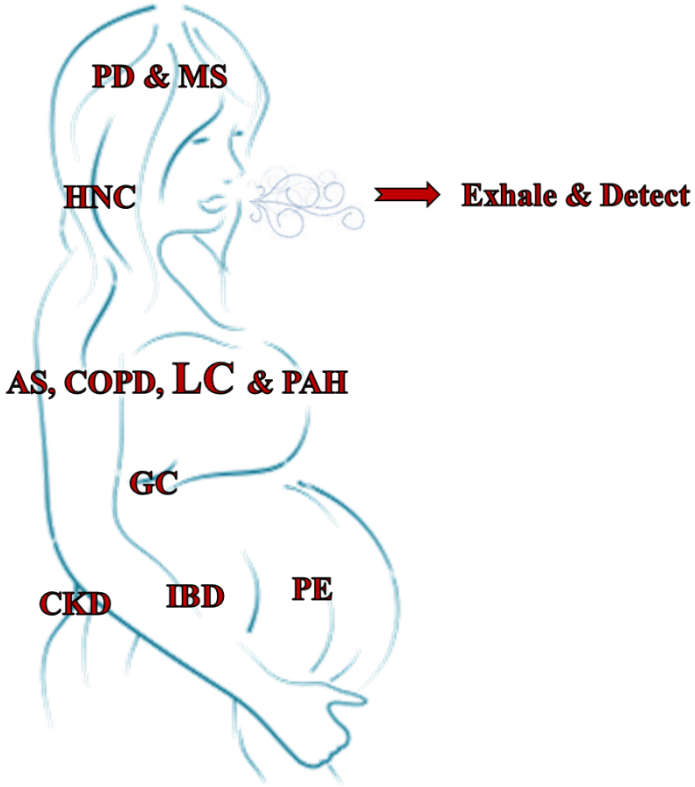

Figure 2. Diseases evaluated through breath tests in the present research and their approximate and main location in the human body (asthma (AS), chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), gastric cancer (GC), head and neck cancer (HNC), inflammatory bowel disease (IBD), lung cancer (LC), multiple sclerosis (MS), Parkinson's disease (PD), preeclampsia (PE), and pulmonary arterial hypertension (PAH)).

Each one of the diseases analyzed in this work will be looked into next. First, a section regarding cancer and LC will be presented, followed by brief descriptions of the other mentioned ten diseases.

1.3.1) Cancer & Lung Cancer – Exhaling the Worst of the Worst

Due to the horrible outcome that cancer originates in many cases, together with the social impact and psychological stress it produces, this disease is one of the most feared, leading to grief in both patients and their family members and friends (Robb *et al.*, 2014; Wess, 2007). Nonetheless, developments for early diagnosis have aided in greatly decreasing the mortality rate of cancer, proving that catching this disease as early as possible is one of the best options to fight its effects and improve its prognosis, as the effectiveness of treatments is much greater during the early stages of cancer (Lau and Lai, 2008; Soerjomataram *et al.*, 2008).

Cancer is a disease that involves dynamic changes in the genome, in other words, which requires genetic alterations or mutations to develop (Bishop and Weinberg, 1996). Despite being a genetic disease, this fact must not be confused with the origin or the causes behind the

appearance of cancer, as 90 to 95% of cancer cases are induced by environmental factors and lifestyle, clearly demonstrating that following adequate habits can prevent many tumor favoring mutations and, in the end, tumorigenic processes (Anand *et al.*, 2008). As a matter of fact, from all cancer-related deaths, around one third of them are related with unhealthy dieting and over one quarter are directly linked to smoking tobacco (Anand *et al.*, 2008). Nevertheless, there is still a relevant group of cancers that are genetic or hereditary (Kinzler and Vogelstein, 1996). For instance, studies have revealed the mutated BRCA-1 and/or BRCA-2 genes clearly predispose women to develop breast cancer (Parmigiani *et al.*, 1998), or people presenting mutations in MSH-2, MSH-6, PMS-1, PMS-2, and/or MLH-1 have greater chances of suffering colorectal cancer (Farrington *et al.*, 1998). In **Figure 3**, a detailed representation of the causes that are involved in the development of cancer can be found.

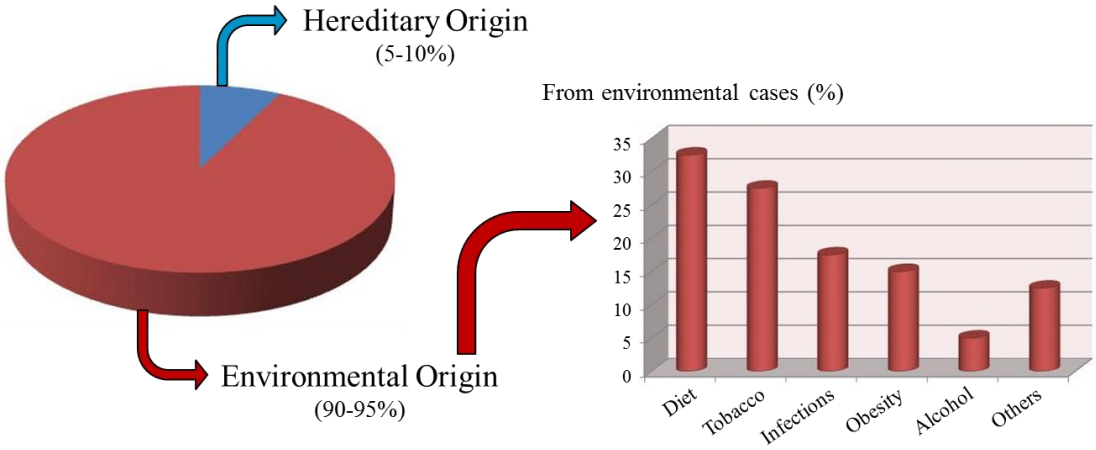

Figure 3. Schematic view of the causes behind the development of cancer and a deeper analysis of the main lifestyle-related factors involved (Anand *et al.*, 2008).

As can be reasoned from this information, the fight against cancer begins in each and every human being, by avoiding common negative habits such as smoking, eating too much fast-food, or consuming excessive amounts of alcohol, and by leaning towards a healthy diet and lifestyle (Block *et al.*, 1992). To sum up, prevention comes first.

Cancer is, to say the least, a complex disease that requires the malfunction of numerous biological phenomena that lead to the excessive proliferation of cells and increase their survival rate. These abnormal activities that favor the appearance and development of tumors are the "hallmarks of cancer", which are shown here as brought to us by Hanahan and Weinberg in two very well-known scientific publications (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011):

1. **Sustained proliferative signaling**; deregulation of growth-promoting signals enables exaggerated proliferation.

2. **Evasion of growth suppressors**; bypassing signals that negatively regulate cell proliferation such as actions from tumor suppression genes.

3. **Cell death evasion**; apoptosis is attenuated, especially in tumors with a high-grade of malignancy.

4. **Enabled replicative immortality**; cells proliferate with no signs of reaching senescence.

5. **Induced angiogenesis**; the creation of new blood vessels are fomented to address the needs of oxygen and nutrients of the tumors. It enables cancerous cells to reach the blood and begin spreading throughout the body.

6. **Activation of tissue invasion and metastasis**; tumor cells start to leave behind the initial site to form new colonies in other near or distant locations. This is when the cancer becomes malignant, greatly worsening prognosis.

7. **Genomic instability and mutations favored**; determined mutant genotypes confer selective advantages to certain cells, allowing them to outgrow other non-cancerous subclones and form a tumor.

8. **Tumor-promoting inflammation**; tumor sites are populated with cells from the immune system leading to typical inflammation processes.

9. **Cellular energetics deregulated**; non-efficient glycolysis is favored in cancerous cells, even in aerobic conditions, to produce energy quickly for fast proliferation. This is known as the Warburg effect.

10. **Immune system avoided**; solid tumors that prevail manage to restrict the actions of the immune system.

In order to clearly break down this list (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011), all of the above basically favor uncontrolled cell proliferation as well as their survival to form, first, localized tumors, and, with time, secondary cancerous sites thanks to phenomena like angiogenesis and metastasis. These ten hallmarks represent an organizing principle that attempts to rationalize the cosmic underlying complexity behind cancer (Hanahan and Weinberg, 2011).

Apart from the incredible organization of the necessary biological events required for tumors to grow and expand in the body, it is also worth mentioning that there are over 200 different types and subtypes of cancer, according to body location or specific affected tissue or cell type in the diseased organ (Cancer Research UK website, 2016-a). Each subtype presents its own statistics in terms of incidence, mortality, and prognosis (Jemal *et al.*, 2011), greatly widening the complexity of this heterogeneous disease as particular measures, or even personalized, should be considered for each patient. For this reason, accurate and early diagnosis is vital for any cancer, as it will enable the selection of the optimal available procedure or treatment, as well as exponentially improve prognosis (Marcano-Cedeno *et al.*, 2011).

A representative, diverse, and highly extended example of cancer is lung cancer (LC). This kind of cancer has various subtypes and can be histologically classified into small-cell lung carcinoma, adenocarcinoma, squamous cell carcinoma, or large cell carcinoma (the last three types are also known as non-small-cell lung carcinomas) when the tumor has an epithelial origin (Tisch *et al.*, 2012). Currently, about 1.6 million new LC cases are diagnosed worldwide per year (Ramalingam *et al.*, 2011). Also, around 1.4 million LC-related deaths take place, which is the largest amount when compared to any other type of cancer (Jemal *et al.*, 2011) and accounts for approximately 28% of all cancer-related deaths (Peled *et al.*, 2012). In **Figure 4**, a schematic view of the estimated cancer cases and deaths, categorized by types and emphasizing the top three per chart, can be seen for the United States in 2014 for both males and females (Siegel *et al.*, 2014), as well as for 40 European countries in 2012 (Ferlay *et al.*, 2013). This embodies two

characteristic examples, demonstrating that, currently, LC presents one of the highest incidence rates, and clearly the worst mortality.

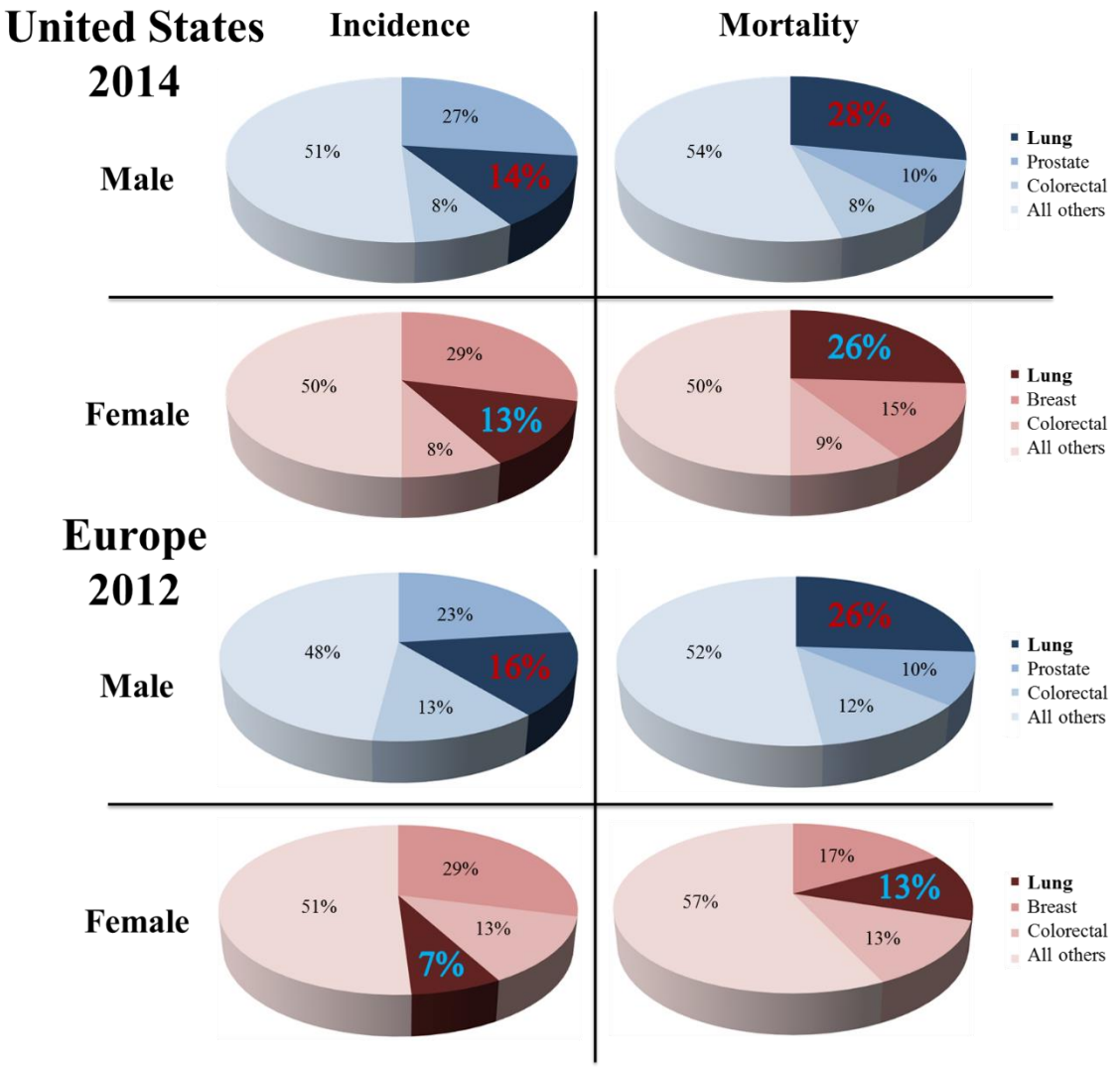

Figure 4. Representation of the estimated incidence and mortality (total deaths) rates for the United States in 2014 and for 40 European countries in 2012, for males and females, highlighting the top three types of cancer in each category (Ferlay *et al.*, 2013; Siegel *et al.*, 2014).

Figure 4 undoubtedly illustrates the impact LC has as it is the second in terms of incidence, but becomes the first in cancer-related deaths for both genders, representing over a quarter of them (except for European women in 2012). Furthermore, the deaths due to LC are around twice the amount of that of the nearest cancer type in the United States and for men in Europe (prostate cancer for men and breast cancer for women) (Ferlay *et al.*, 2013; Siegel *et al.*, 2014). These statistics are caused by two main factors: LC incidence is clearly correlated with smoking tobacco, while the elevated mortality is linked to late diagnosis, which validates the relevance of the present research, as one of its aims is early detection of LC. As a matter of fact, in the United States, the five-year survival rate of LC patients between 2003 and 2009 was around 50% for those that were diagnosed early, when the tumor was still local, opposed to the 4% survival when diagnosed after distant cancerous colonies had been established (Siegel *et al.*, 2014). Therefore, once again, the solution is prevention first and early detection after (Jemal *et al.*, 2011; Siegel *et al.*, 2014). As of today, LC diagnosis mainly relies on x-ray, computed tomography scan, bronchoscopy, or biopsy (Cancer Research UK website, 2016-b). In contrast,

in our research, an attempt to aid in LC early diagnosis is carried out, as mentioned already, through non-invasive breath tests, which would offer complementary information to the one provided by other methods, to discover LC-related patterns created by volatile biomarkers in the early stages of this disease to greatly improve its currently devastating prognosis. The fact that this prognosis is far from being optimal is mainly because during the early stages of LC, it is an asymptomatic disease. For this reason, developing reliable tests that can consistently and accurately locate the disease at these initial stages by merely blowing into a breathalyzer-like device would imply incredible progress in the clinical field, greatly improving survival rates and quality-of-life.

Several research groups from different parts of the world are and have been working on the design of tools with the goal set to diagnose LC through breath analysis (Bajtarevic, *et al.*, 2009; Dragonieri *et al.*, 2009; Peng *et al.*, 2009; Phillips *et al.*, 2007; Poli *et al.*, 2005), and during the present research, it has been looked into in different experiments as well. Nonetheless, we would like to take the application of breath analysis a bit further.

1.3.2) Other Diseases Analyzed – Specifically Labeled Breath?

It may seem logical or easy to understand that a disease such as LC will have a direct impact on the composition of breath, as it particularly affects lungs, which are the organs that produce the air we exhale. On the other hand, non-respiratory diseases such as Parkinson's disease or chronic kidney disease could be harder to perceive as diagnosable through breath analysis. In any case, in this work, we recognize that the metabolic changes that diseases induce throughout the body are somewhat reflected in the volatile compounds in breath in terms of presence and amount (Amman *et al.*, 2014; Buszewski *et al.*, 2007).

In this part of the introduction, a set of worldwide relevant diseases that have been analyzed thorough breath tests during this research will be looked into. Ten completely different and unrelated diseases have been assessed in an attempt to confirm that disease-specific patterns are reflected in a patient's breath, enabling the design of versatile diagnosing devices or, in general, disease detectors. These diseases are asthma, chronic kidney disease, chronic obstructive pulmonary disease, gastric cancer, head and neck cancer, inflammatory bowel disease, multiple sclerosis, Parkinson's disease, preeclampsia, and pulmonary arterial hypertension (**Figure 2**). Some brief notes will be commented next regarding these ten diseases and their background in terms of their diagnosis.

1.3.2.1) Asthma

Asthma (AS) is a chronic inflammatory airway disorder or respiratory disease which affects people irrespective of age, and, when not controlled properly, can greatly hinder the quality-of-life of patients and even cause their death (Bateman *et al.*, 2008). Apparently affecting women more than men (Kynyk *et al.*, 2011), AS possesses a complex pathogenesis and it is defined through a set of clinical, pathological, and physiological traits. It is characterized by chronic inflammation due to airway hyperresponsiveness that originates wheezing (most common symptom) and coughing, as well as breathlessness and chest tightness. These episodes are commonly linked to airflow obstruction inside of the lung, which is usually naturally or artificially

(with treatment) reversible (Bateman *et al.*, 2008). It is thought that both genetic and environmental factors play a role in the risk and development of AS. Genetic factors are more linked to being the causes of the actual development of the disease, while environmental ones, such as air pollution or determined allergens, seem to be related to triggering AS symptoms, not necessarily due to the disorder (Bateman *et al.*, 2008).

The most recent studies estimate that there are around 334 million AS cases worldwide, and its prevalence is seemingly increasing in lower-income countries, which represents a clear reverse in the trend that has been occurring during the last five to ten years, as wealthy countries typically presented higher prevalence rates (Global Asthma Report website, 2016). The burden of AS, which is quantified through its disability and mortality rates, is largest for children within the ages of 10 and 14 and for the elderly (over 75 years old) (Global Asthma Report website, 2016). This disease accounts for approximately 250 thousand deaths per year, and, therefore, its early detection is obviously relevant (Pinnock *et al.*, 2010). The current state of AS diagnosis is based on clinical evidence (symptom evaluation; assessment of medical history and physical examination) and by measuring lung function, airway responsiveness, and allergic status (Bateman *et al.*, 2008). On the other hand, in terms of breath analysis, there are studies that have shown a possible relation between the fraction of exhaled nitric oxide and airway inflammation caused by disorders like AS, although the cost-effectiveness of its measurement to diagnose AS is yet to be validated (Bjermer *et al.*, 2014). In addition, a different analysis, where time-of-flight-secondary ion mass spectrometry was used to measure breath samples of AS patients and healthy controls, has revealed that the amount of particles in exhaled air in the patients was considerably lower than for controls, as well as their unsaturated to saturated phospholipid ratio, showing that breath tests can aid in the diagnosis of this disease (Almstrand *et al.*, 2012).

1.3.2.2) Chronic Kidney Disease

Chronic kidney disease (CKD) is a serious health condition that affects millions worldwide. Its potential adverse consequences range from kidney failure or cardiovascular disease, all the way to premature death (Levey *et al.*, 2005). The definition or development of this disease is based on two parameters, of which at least one takes place: kidney damage or reduced glomerular filtration rate (GFR), which describes the flow rate of filtered fluid through the kidney (Levey and Coresh, 2012). Kidney damage is generally assessed by measuring the amount of protein in urine, specifically albumin. Albumiuria (excess of urinary excretion of albumin) is therefore linked to kidney damage, and is defined as an albumin-to-creatinine ratio greater than 30 mg/g in two out of three independent urine tests (Go *et al.*, 2014; Levey *et al.*, 2005). On the other hand, GFR below 60 mL/min per 1.73 m^2 of body surface area, irrespective of cause, for over three months, is also understood as CKD (Levey *et al.*, 2005).

To get an idea of the extent of this disease, in the United States around 13% of the people suffer from CKD (over 26 million), of which most cases are yet to be diagnosed (Go *et al.*, 2014). Additionally, another 20 million people are at clear risk (Go *et al.*, 2014). These statistics prove that reliable and fast CKD diagnosing tools would be quite handy for this specific part of the clinical sector, as currently the detection is mainly based on GFR assessments (Levey and Coresh, 2012). Recently, some studies have been able to relate breath compounds and certain breath compositions with CKD presence. For instance, the amounts of trimethylamine and pentane in exhaled breath have shown a statistical difference between CKD patients and healthy subjects

(Grabowska-Polanowska *et al.*, 2013). In addition, a different study using cross-reactive gold nanoparticle sensors that interact with breath has set a proof-of-concept which validates the potential ability to attain early detection of CKD and the monitoring of its progression through breath analysis. They were able to correctly distinguish healthy states from early CKD stages 79% of the cases, and classify late CKD stages with 85% accuracy (Marom *et al.*, 2012).

1.3.2.3) Chronic Obstructive Pulmonary Disease

Chronic obstructive pulmonary disease (COPD) is an airway disorder that affects people all around the world, and the main risk factors are smoking and other inhaled exposures such as occupational smoke and/or dust, air pollution, or biomass fuels (Halbert *et al.*, 2006). Although not as evident as smoking tobacco, studies have also shown potential genetic predisposition to having COPD (Eisner *et al.*, 2010; Viegi *et al.*, 2007). The definition proposed by the American Thoracic Society and European Respiratory Society states the following: "*COPD is a preventable and treatable disease state characterized by airflow limitation that is not fully reversible. The airflow limitation is usually progressive and associated with an abnormal inflammatory response of the lungs to noxious particles or gases, primarily caused by cigarette smoking. Although COPD affects the lungs, it also produces significant systemic consequences*" (a highly similar definition is also given by the Global Initiative for Chronic Obstructive Lung Disease) (Viegi *et al.*, 2007). It must be highlighted that a relevant heterogeneity exists regarding the clinical presentation, imaging, physiology, therapy responsiveness, lung impairment, and survival rate of different COPD cases and populations (Han *et al.*, 2010), which complicates accurate diagnosis.

This pulmonary disease represents a global health concern that currently affects around 10% of the population over 45 years old, rising up to an overwhelming 50% when considering heavy smokers (Kirkham and Barnes, 2013), and it causes approximately 2.75 million deaths per year (Calverley *et al.*, 2007). Nevertheless, prevalence and incidence data is not extremely accurate due to the complicated nature of this disease, broad definition, and, therefore, its complex detection and diagnosis, which in many times depends on the physician's criteria (Viegi *et al.*, 2007). For these reasons, a fast and non-invasive breath test to help detect COPD or guide the clinical specialist would clearly be useful for the health sector. In this regard, several biomarkers have been reported to have elevated concentrations in the exhaled breath of COPD patients such as carbon monoxide, 8-isoprostane, hydrogen peroxide, nitrite, and 3-nitrotyrosine (Kharitonov and Barnes, 2010). On the other hand, a research group used gas chromatography-mass spectrometry to identify VOCs in breath from COPD patients and control subjects, reaching a set of six compounds that allowed discriminating the groups with a 93% accuracy (Van Berkel *et al.*, 2010).

1.3.2.4) Gastric Cancer

Despite showing decreasing trends in terms of incidence and mortality, especially in Western countries, gastric cancer (GC) is still a very extended and dangerous disease (Crew and Neugut, 2006; Jemal *et al.*, 2011). Approximately 90% of GC cases or stomach tumors are adenocarcinomas (epithelial tumors that possess a glandular origin and/or glandular traits), and they are classified according to their microscopic morphology alone. They can be well-differentiated or intestinal subtypes, or undifferentiated or diffuse subtypes (Crew and Neugut,

2006; Nobili *et al.*, 2011). In relation to risk factors, it has been shown that a *H. pylori* infection, which affects nearly 50% of the population, is the strongest one. These bacteria lead to chronic inflammation and greatly increase the risk of developing ulcers in the duodenum and stomach as well as GC (Wroblewski *et al.*, 2010). Other observed risk factors are the consumption of high quantities of salt and salt-preserved foods, smoking tobacco, and determined genetic polymorphisms, although this last factor still lacks consistency in the results revealed by research groups (Fock, 2014).

Stomach cancer is the fourth most common type in terms of new cases and the second in deaths per year (only behind LC), accounting for about 8% of total cases and 10% of deceases, of which around 70% occur in developing countries (Crew and Neugut, 2006; Jemal *et al.*, 2011). Showing the highest incidence rates in Asia, Eastern Europe, and South America, GC is developed in men twice as much as in women (Jemal *et al.*, 2011). Even though the mortality of this disease has been decreasing during the past few decades, GC still offers an unfavorable prognosis, mainly due to late detection. This is demonstrated by the decrease in mortality in high-risk locations, such as Japan, where screening has been implemented (national endoscopic surveillance program) for early diagnosis (Crew and Neugut, 2006). Therefore, a breath-based tool that is able to locate GC at its early stages would be a great device to aid during these screening programs to try and detect this disease at curable stages. Nowadays, many studies focus on finding reliable non-invasive approaches based on breath tests to detect *H. pylori* infections, which is useful for GC diagnosis, as it is the main risk factor for its development (Chey and Wong, 2007; Malfertheiner *et al.*, 2002). Nevertheless, directly detecting GC through breath is a different story. A research group has used a set of nanomaterial-based sensors to analyze the breath samples of 130 patients which had either GC, ulcers, or less severe conditions, and they were able to reach correct classification rates that ranged from 77 to 93% (Xu *et al.*, 2013). Also, a different study where selected ion flow tube mass spectrometry was employed to quantify VOCs in breath samples revealed four potential biomarkers (hexanoic acid, phenol, methyl phenol, and ethyl phenol) with significantly distinct concentrations in the breath of esophago-gastric cancer patients and healthy controls (Kumar *et al.*, 2013).

1.3.2.5) Head and Neck Cancer

The term head and neck cancer (HNC) is employed to define a broad set of tumors which may arise from multiple different locations. These anatomic origins comprise craniofacial bones, soft tissues, salivary glands, skin, and mucosal membranes (Pai and Westra, 2009). Around 90% of HNCs are squamous cell carcinomas, implying that most of these tumors have an epithelial source (Pai and Westra, 2009). Common to other types of cancer (e.g., LC), smoking tobacco has been well established as the main risk factor for the development of HNC. This risk is clearly linked to the intensity and duration of the habit, although its cessation does not strictly imply an elimination of HNC appearance (Schlect *et al.*, 1999). While tobacco (and alcohol) account for most head and neck squamous cell carcinomas, it is important to note that a completely different risk factor is behind over 60% of oropharyngeal cancers. It is the human papillomavirus (HPV), specifically HPV-16, a DNA virus which solely infects keratinocytes of the skin or mucous membranes (Leemans *et al.*, 2011; Marur *et al.*, 2010).

Near 500 to 600 thousand new cases of HNC are estimated to develop per year worldwide, including oral cavity, laryngeal, and oropharyngeal sites (Parkin *et al.*, 2005). The current status

of HNC diagnosis could be greatly improved as most cases are diagnosed late. This leads to a lower than 50% cure rate, which could undoubtedly be improved with the implementation of accurate disease screening (Hakim *et al.*, 2011). This data would drastically change if a reliable breath test to diagnose HNC at its early stages would be validated. The patients would clearly benefit from this device, which strengthens the importance of the research carried out here. In the recent past, some scientific papers have been published in this regard. Some research groups have been working with electronic noses (devices which incorporate metal-oxide sensors) to analyze breath samples with noteworthy preliminary results (Leunis *et al.*, 2014; Witt *et al.*, 2012). On the other hand, gold nanoparticle-based sensors have also been employed to classify HNC patients, providing optimistic evidence for this promising approach by reaching around 95% correct HNC patient and control classification rate in a 42 subject study (Hakim *et al.*, 2011). Additionally, using gas chromatography/mass spectrometry, during a feasibility study, some potential breath biomarker candidates for HNC were found, which were ethanol, 2-propenenitrile, and undecane (Gruber *et al.*, 2014).

1.3.2.6) Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) represents a group of idiopathic, chronic, and inflammatory diseases that are subdivided into two more specific classes known as Crohn's disease (CD) and ulcerative colitis (UC), which present both overlapping as well as distinctive clinical and pathological traits (Bernstein *et al.*, 2010; Mowat *et al.*, 2011). CD and UC lead to a characteristic inflammation of the gastrointestinal tract, and appear in genetically prone subjects that have been exposed to determined environmental risk factors (Molodecky *et al.*, 2012; Podolsky, 2008). CD is characterized by patchy and transmural (through the wall of an organ) inflammation, which may affect any region of the gastrointestinal tract. On the other hand, UC is portrayed by diffuse inflammation which is limited strictly to the colon (Mowat *et al.*, 2011).

Currently, incidence and prevalence data, or the geographic trends this disease follows, still requires great research and analysis, especially for developing countries (Molodecky *et al.*, 2012). Nevertheless, it seems that incidence and prevalence is greater in westernized nations in comparison to other regions. Around 10 to 30 new cases per 100 thousand people are diagnosed for each, CD and UC, in countries such as Canada, the United Kingdom, Iceland, or Australia (Molodecky *et al.*, 2012). Currently, the diagnosis of IBD requires a thorough physical evaluation as well as a meticulous review of the patient's clinical background, combined with other tests (blood tests, stool examinations, endoscopies, biopsies, and so on) to aid and assert the detection of the disease (Bernstein *et al.*, 2010). This process would most likely be propelled by a reliable and non-invasive breath test to help in this diagnosis, liberating in certain cases both patients and medical staff from some of the numerous necessary, and, in cases, bothersome (or embarrassing) tests. In other studies that attempt to turn these non-invasive devices into a reality, pentane has been quantified in exhaled breath using selected ion flow tube-mass spectrometry (SIFT-MS) (Dryahina *et al.*, 2013). This has been done because pentane has been considered a potential volatile biomarker for IBD diagnosis as its production is related to inflammatory processes (Dryahina *et al.*, 2013). A different research group has speculated that determining the amount of exhaled nitric oxide may be valuable during the follow-up of CD patients (Malerba *et al.*, 2011). In addition, promising results were obtained by a third group that also used SIFT-MS, this time to analyze entire breath samples to locate patterns that allow diagnosing pediatric IBD (Patel *et al.*, 2014). Finally, some other very recent publications and reviews have been written in the

context of diagnosing IBD and other gastrointestinal diseases through breath analysis employing a wide assortment of analytical alternatives (Aggio and Probert, 2014; Cauchi *et al.*, 2014; Huang *et al.*, 2014; Markar *et al.*, 2015).

1.3.2.7) Multiple Sclerosis

Multiple sclerosis (MS) is a heterogeneous autoimmune neurodegenerative disease that affects the central nervous system, and its main consequence is the inflammatory- and immune-mediated demyelination of determined areas of the brain and the spinal cord white matter, which lead to axon degeneration and, in the end, neuron demise (Glass *et al.*, 2010; Trapp and Nave, 2008). This disease produces defects in sensation as well as in motor, autonomic, visual, and cognitive systems, mainly affecting young adults, and over twice the amount of females than males (Glass *et al.*, 2010). Genetic factors seem to play the main role in the occurrence of MS as the frequency of the disease is clearly increased in relatives of MS patients, and it has been proven that variations or mutations in the major histocompatibility complex (family of cell surface molecules that have a major impact on the immune system) represent the single highest risk factor (Sawcer *et al.*, 2011). Nonetheless, it has been reported that in some countries like Canada or Denmark the MS incidence ratio for women has been increasing in comparison with that of men, leading to the thought of the existence of environmental risk factors as well. As a matter of fact, smoking has shown a certain correlation with the development of MS, as similarities have been found in the sex-related smoking and MS incidence trends (Palacios *et al.*, 2011).

It is estimated that around two million people suffer from MS worldwide, of which nearly half are European (Kingwell *et al.*, 2013). Currently, MS diagnosis is based on clinical and paraclinical laboratory evaluations, which focus on proving the disseminations of lesions in time and space and on excluding other possible diseases. Additionally, magnetic resonance imaging of the central nervous system plays a relevant role supporting (or even replacing) clinical criteria (McDonald *et al.*, 2001; Polman *et al.*, 2011). Nevertheless, this approach is expensive, originating the need to search for reliable biomarker-based diagnosing alternatives. For instance, the analysis of cerebrospinal fluid has revealed potential MS diagnostic evidence (Link and Huang, 2006), although the process is invasive and uncomfortable as a lumbar puncture is required. On the other hand, specifically regarding non-invasive breath-based MS diagnosis, an article has been found reporting evidence of its potential implementation (Ionescu *et al.*, 2011). A cross-reactive sensor array of polycyclic aromatic hydrocarbons and single-wall carbon nanotube bilayers was employed to analyze exhaled breath of MS patients and healthy controls, and the results were promising as around an 80% accuracy in their classification was attained for a 51 volunteer study. Additionally, they discovered two potential MS biomarkers in hexanal and 5-methyl-undecanes, as the amounts of these VOCs showed a significant statistical difference between both studied populations (Ionescu *et al.*, 2011).

1.3.2.8) Parkinson's Disease

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease (AD), as well as the most frequent movement disorder (Glass *et al.*, 2010). PD is, as other neurodegenerative diseases like AD, a proteinopathy that leads to the accumulation of incorrectly folded α-sinuclein, which form intracellular protein aggregations known as Lewy

bodies (Glass *et al.*, 2010). Its main consequence is the death of dopaminergic neurons (producers of the neurotransmitter dopamine) in the substantia nigra pars compacta region of the brain and it is associated with chronic inflammation (Schapira, 2009; Tansey and Goldberg, 2010). All these consequences trigger various clinical features, both motor (e.g., tremor or rigidity) and non-motor related (e.g., cognitive deficits or sleep disorders) (Glass *et al.*, 2010). The etiology of PD, or the risk factors involved in its appearance, is not currently fully understood. Nevertheless, both genetic and environmental factors seem to play a role (Wirdefeldt *et al*., 2011). In addition, it is worth mentioning that in general terms males are affected more than females (Wirdefeldt *et al*., 2011).

This disease possesses an elevated lifetime risk, which is reflected in most prevalence studies. These report that around 100 to 300 people suffer from PD per 100 thousand (Wirdefeldt *et al*., 2011), affecting about 2% of individuals over 60 years old (Glass *et al.*, 2010; Schapira, 2009). The current state of the diagnosis of PD is mainly based on clinical criteria. Primarily, parkinsonian symptoms are assessed and other neurological damage is discarded, as well as history of drugs, toxins, or infections that may potentially emulate similar symptoms (Wirdefeldt *et al*., 2011). Diagnosis of PD is addressed in this fashion due to the scarce information provided by neuroimaging or biomarker analyses (Wirdefeldt *et al*., 2011). For these reasons, the design of a breath-relying test to assist in the diagnosis of PD would be significant for the field, as it may help orient the medical staff in the correct direction. In this context, a scientific paper has been found stating this possibility (Tisch *et al.*, 2013). They employed cross-reactive nanomaterial-based sensors (organically functionalized carbon nanotubes and gold nanoparticles) to attempt and classify PD patients, AD patients, and healthy controls, with impressive results ranging from 78 to 85% correct classification rates. In addition, they discovered several VOCs (biomarkers) through gas chromatography coupled with mass spectrometry that showed statistically different amounts in the three groups (PD and AD patients and controls) (Tisch *et al.*, 2013).

1.3.2.9) Preeclampsia

Preeclampsia (PE), formerly known as toxemia of pregnancy, is a multisystem pregnancy disorder that begins in the placenta and can take place during the second half of the pregnancy, labor, or soon after delivery, potentially affecting both mother and fetus (Redman and Sargent, 2005; Sibai *et al.*, 2005). It is defined by two main systemic disturbances, these being the existence of hypertension and proteinuria (elevated amounts of protein in urine) (Hawfield and Freedman, 2009). Some determined risk factors that increase the odds of developing PE are the existence of previous episodes, obesity, black race, diabetes, multiple gestation, or being below 20 years old (Hawfield and Freedman, 2009; Wallis *et al.*, 2008).

It is reported that around 5% to 8% of pregnancies are complicated by PE (Hawfield and Freedman, 2009). Additionally, data has demonstrated that its incidence rate has increased in the past, between 1987 and 2004 (Wallis *et al.*, 2008), but, nevertheless, a review covering the PE databases from 2002 to 2010 reports that the existing data quality is poor and highlights that the information is not broad enough and too heterogeneous to make an accurate worldwide incidence estimate (Abalos *et al.*, 2013). The diagnosis of PE is based on the evaluation of the two main factors which define this disease: hypertension associated with proteinuria after the 20th week of pregnancy (Sibai *et al.*, 2005). Systolic blood pressure values over 140 mm Hg as well as diastolic ones above 90 mm Hg (in at least two distinct tests, four to six hours apart) combined with greater

than 0.3 grams of protein in urine per day in a formerly normotensive woman are diagnostic indicators of PE (Sibai *et al.*, 2005). Despite the non-invasive nature of these tests, the gynecological field would definitely benefit from a breath test that can potentially assist in the early detection of PE, and there is a research study that has published work in this regard (Moretti *et al.*, 2004). They determined that oxidative stress-related VOCs were clearly higher in PE patients when compared to healthy pregnant controls, potentially enabling an accurate diagnosing device (Moretti *et al.*, 2004).

1.3.2.10) Pulmonary Arterial Hypertension

Pulmonary arterial hypertension (PAH) is a disease that is characterized by high pulmonary arterial pressures and is associated with elevated pulmonary vascular resistance, which can originate right ventricular failure, volume overload (chamber of the heart with excessive amount of blood), deteriorated cardiopulmonary function, and even premature death (Chan and Loscalzo, 2008; Chin and Rubin, 2008; Zhou *et al.*, 2012). The origin of 6% of reported PAH cases are linked to genetic alterations (Newman *et al.*, 2004), while the vast majority are idiopathic cases and appear due to exogenous factors such as chronic hypoxia, viral infections, hemoglobinopathies, or autoimmune vascular disease (Chan and Loscalzo, 2008). In addition, according to statistics there is a gender predilection, as women seem to have greater chances to suffer PAH compared to men (Robles and Shure, 2004).

Different epidemiological studies carried out in France, the United Kingdom, and Ireland have revealed that the prevalence of PAH ranges from around 10 to 50 cases per million people, while the incidence rate is about 1 to 7 cases per year and per million inhabitants (Humbert *et al.*, 2006; Ling *et al.*, 2012; Peacock *et al.*, 2007). As with many diseases, early diagnosis of PAH is beneficial, as it enables the implementation of targeted therapies prior to the appearance of relevant right heart failure (Hoeper *et al.*, 2013). Nonetheless, detection of idiopathic PAH is based on a diagnosis of exclusion (made by a process of elimination of other diseases such as, for example, human immunodeficiency virus, connective tissue disease, or congenital heart disease) (Galie *et al.*, 2009), which clearly indicates that a breath test that could assist in its detection would be significantly helpful. In this context, there are some research works that are related to breath analysis to diagnose PAH. For instance, there are some scientific articles which explain the analysis of the compounds in exhaled breath condensate to discriminate between healthy subjects and PAH patients (Mansoor *et al*, 2014; Warwick *et al.*, 2012). On the other hand, two other papers work with direct exhaled breath using gold nanoparticle sensor arrays (Cohen-Kaminsky *et al.*, 2013) and ion flow tube-mass spectrometry (Cikach *et al*, 2014) to process the samples, and both were successful in terms of distinguishing healthy controls from PAH patients (about 92% accuracy for the former and 83% for the latter).

Now that cancer and LC, as well as a set of ten other diseases have been defined, it seems clear that the design of devices which through breath analysis are able to reliably detect different diseases and assist in decision making, would imply a great leap forward in the medical field. The prognosis of many diseases incredibly improves with early detection, thus emphasizing the importance of these non-invasive and safe disease-detecting systems. Nonetheless, in order to reach this goal, adequate and consistent analytical approaches to analyze complex gaseous matrices such as breath ought to be correctly selected and tuned. The technology and equipment employed in this research will be covered in the following part of the present thesis.

1.4) Analytical Equipment and Technology – Processing Exhalation

Once acknowledged that the clinical status of a human being can be to an extent reflected in exhaled air (Amman *et al.*, 2014; Buszewski *et al.*, 2007), the potentiality behind breath tests for disease detection and early diagnosis is immense. The fact that breath analysis contains underlying information that can provide an insight of the clinical status of a person, converts the need of perfecting this approach into a must, as the procedure would be straightforward, non-invasive, and user-friendly for the patient. To achieve this, the equipment to acquire this "exhaled data" should be correctly selected, perfected, and optimized.

In this research, three main analytical approaches were employed. Each one was used in different experimental procedures to analyze breath samples or gaseous mixtures, and they were proton transfer reaction-mass spectrometry and two different cross-reactive sensor arrays based on functionalized silicon nanowire field-effect transistors and functionalized gold nanoparticles. The basis and background of the three methodologies used, together with their applications and current involvement with breath analysis will be described next.

1.4.1) Proton Transfer Reaction-Mass Spectrometry

Proton transfer reaction-mass spectrometry (PTR-MS) is a technique that is employed nearly exclusively to detect VOCs in air (Blake *et al.*, 2009). It has rapidly evolved, turning into a quick and sensitive sensing device for monitoring VOCs (Capellin *et al.*, 2013). It is becoming a more than feasible alternative to the more commonly employed gas chromatography-mass spectrometry (GC-MS) for the detection and quantification of VOCs (Blake *et al.*, 2009; Ligor *et al.*, 2009). GC-MS is a highly reliable and sensitive technique, but nonetheless has several drawbacks, such as requiring between minutes and tens of minutes to fully separate the VOCs and, in many cases, preconcentration steps (Blake *et al.*, 2009). Both of these issues are solved through PTR-MS, which practically enables real-time identification and quantification of VOCs without preconcentrating, with sensitivities clearly breaching the pptV scale (Blake *et al.*, 2009; Ligor *et al.*, 2009).

PTR-MS is a type of direct injection mass spectrometry method that was developed in the 1990s in Professor Werner Lindinger's laboratory in Austria, initially reaching the on-line measurement of components at concentrations as low as 1 ppb (Hansel *et al.*, 1995). This technique is based on the chemical ionization through proton transfer of a gaseous sample contained in a drift tube (reaction chamber), which is a conducting enclosure at a constant potential so charged particles suffer no change in velocity inside it, enabling a stable reaction time for the ions as they flow through the tube (Lindinger *et al.*, 1998). The proton source or donor is typically H_3O^+, and if its concentration is essentially unchanged when reacting with the sample, the concentration of the acceptor molecules (VOCs) can be determined rapidly and with a great sensitivity after combining reaction kinetics with mass spectrometry (Blake *et al.*, 2009). H_3O^+ is selected as it can originate proton-transfer reactions with many VOCs whilst not reacting with the common molecules in air such as O_2 and N_2, as these present lower proton affinities than water (Zhan *et al.*, 2013).

The three most important components in a PTR-MS system are the ion source (hollow cathode), which produces H_3O^+, the drift tube, where the VOCs in the sample experience the non-dissociative proton transfer, and the analyzing system or mass spectrometer, which leads to the

detection and quantification of the VOCs (Ionicon website, 2016-a). A schematic representation of a PTR-MS system can be seen in **Figure 5** (Hansel *et al.*, 1995; Ionicon website, 2016-a).

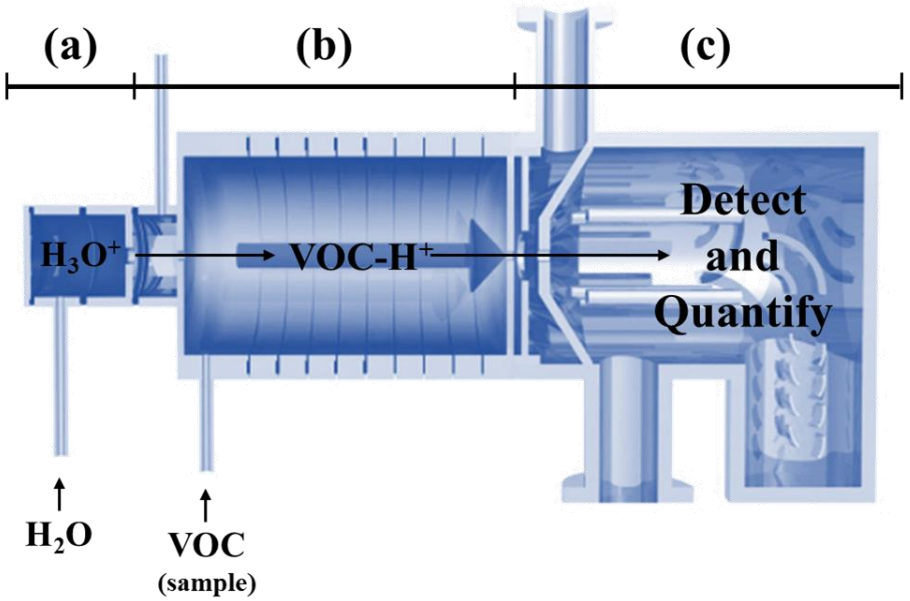

Figure 5. Schematic representation of a PTR-MS system. The three main parts are shown: **(a)** Ion source, where water molecules are converted into H_3O^+; **(b)** Drift tube, where the VOCs in the sample are ionized; **(c)** Mass spectrometer and detector, where the VOCs are identified and quantified (Hansel *et al.*, 1995; Ionicon website, 2016-a).

The considerable advantages that PTR-MS brings to the table allows its implementation in a wide range of fields and applications. For instance, this technology can be employed within the environmental scope, to locate air contaminating sources (Rogers *et al.*, 2006) or to assess indoor air quality (Kolarik *et al.*, 2010), all the way to applications in food technology, for example, to analyze the complex VOC-profile that coffee emanates (Yeretzian *et al.*, 2003). Nevertheless, what truly interests us, within the context of this research, is the utilization of PTR-MS in the medical and clinical area (Zhan *et al.*, 2013). This methodology has been employed, among others, for urine analysis, *in vivo* skin studies, quality control of medical devices, and, most importantly, breath analysis for disease diagnosis (Zhan *et al.*, 2013).

Regarding PTR-MS combined with breath analysis, different potential non-invasive procedures are being worked on, with the goal set to, for instance, determine blood cholesterol levels (Karl *et al.*, 2001) or detect LC (Bajtarevic *et al.*, 2009). In the present work, as mentioned in previous sections, an attempt to differentiate LC patients from healthy controls has been carried out, and one of the analytical options selected has been PTR-MS, due to its great sensitivity, ability to operate at real-time, and because it enables the discovery of breath biomarkers to diagnose diseases as well as their quantification at very low concentrations (Smith *et al.*, 2014).

1.4.2) Cross-Reactive Sensor Arrays – Gathering Gaseous Fingerprints

To begin this section, prior to analyzing the two specific types of sensor arrays synthesized and employed, it is important to fully understand the meaning of cross-reactivity and cross-reactive sensors. To define them, a brief description of conventional sensors will be given. Typically, sensors are designed to detect specific compounds through a "lock-and-key"

interaction, which resembles, for instance, the way enzymes interact with their substrates. In other words, sensors are created to act as a "lock" against determined "key" analytes, reaching highly selective systems (Tisch and Haick, 2010-a; Tisch and Haick, 2010-b). This methodology comes in handy when particular molecules within a fixed or regulated background have to be detected or quantified with high sensitivities. Nevertheless, when working with mixtures, the design of specific sensors would be required for each compound which needs to be detected. In addition, these sensing devices struggle when chemically similar compounds are present in mixtures, as cross-reactivity can occur (Tisch and Haick, 2010-b). Therefore, to overcome these issues, sensors or sensor arrays can be designed to take advantage of this cross-reactivity, allowing the analysis of complex matrices of gases (electronic noses) or liquids (electronic tongues) as fingerprint-like responses can be obtained enabling the assignment of patterns to determined types of samples (Röck *et al.*, 2008; Tisch and Haick, 2010-b). These are known as cross-reactive sensors, and they can be found graphically represented and compared with common selective sensors in **Figure 6**.

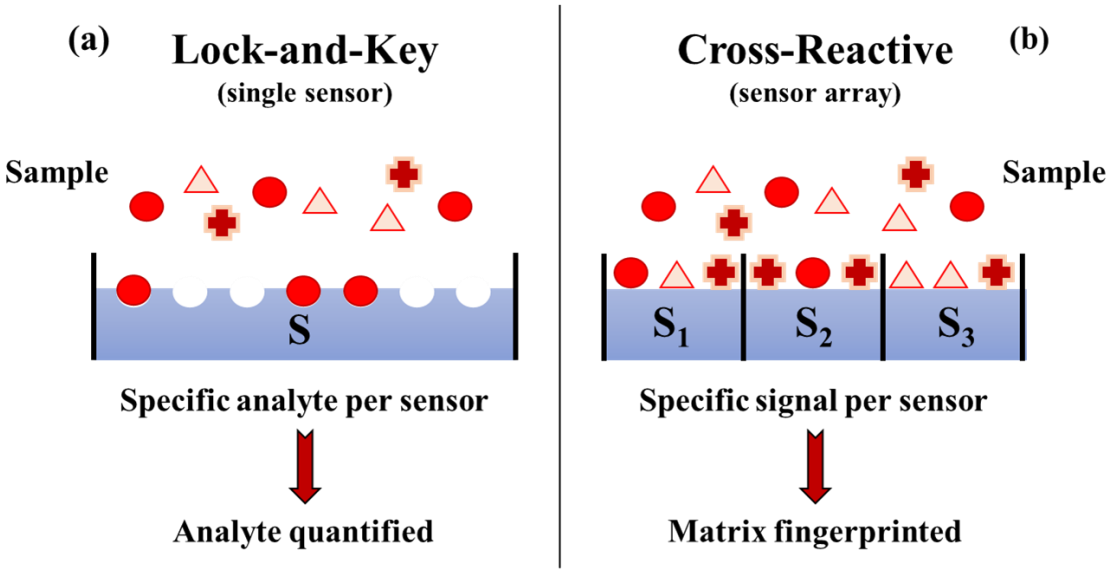

Figure 6. Comparison regarding sensors for **(a)** specific analytes (lock-and-key) and **(b)** cross-reactive sensor arrays. The first ones are commonly employed to detect and quantify particular molecules, while the second ones are used to extract information about an entire mixture of compounds.

As can be deduced, the main difference between the two kinds of sensors defined is that specific lock-and-key sensors excel when the quantification of a determined compound in a matrix is required, while cross-reactive sensors enable a global analysis of complex samples that may contain hundreds or thousands of types of molecules, without the need of fully understanding their nature or amount. A perfect example of these kind of samples are exhaled breaths, reason why cross-reactive sensors have been selected to carry out the analyses during the present research. The idea is to attain specific patterns from breath samples of patients of different diseases, as well as healthy controls, which would empower the design of non-invasive diagnosing tools based on breath tests. In other words, breath is analyzed as a whole with the goal set to search, discover, and interpret the fingerprints (or "breathprints") that define different diseases. Several scientific articles show revealing results in this regard (Hakim *et al.*, 2011; Shehada *et al.*, 2015; Tisch and Haick, 2010-a), signifying that this path should be further explored.

Two distinct types of cross-reactive sensor arrays have been synthesized to analyze gaseous matrices and breath samples, and they are functionalized silicon nanowire field-effect

transistors and functionalized gold nanoparticles. They will both be looked into in the next subsections.

1.4.2.1) Silicon Nanowire Field-Effect Transistor Sensor Arrays

The first type of cross-reactive sensors synthesized were field-effect transistor (FET) sensors, which are becoming a popular and useful selection within chemical and biochemical contexts as reliable detecting and quantifying systems (Paska and Haick, 2009). Nonetheless, prior to getting into the actual sensors, let us look into the basic traits concerning FETs. Invented and patented by Julius Edgar Lilienfeld in the 1920s (Kleint, 1998), FETs operate as capacitors in which a first plate plays the role of a conducting channel that connects two ohmic contacts known as source and drain electrodes (**Figure 7**). The relative amount of charge carriers within the channel depends on the voltage applied to a second plate, which is the gate electrode (Horowitz, 1998). FETs, and particularly metal-oxide-semiconductor FETs (MOSFETs), were present in around 90% of all the semiconducting devices and equipment in the market at the beginning of this century (Sze, 2001). The semiconductor contained inside a MOSFET is a material formed by silicon and thermally grown SiO_2 (**Figure 7**), which thanks to the rapid development of technology, is ideal to create nanoscale sensing devices (Sze, 2001).

A specific kind of FET-based nanosensor is the silicon nanowire (SiNW) FET sensor (**Figure 7**). It acquires this name as silicon nanowires are employed to connect source and drain electrodes in the FET sensor, providing the powerful sensing properties (Cui *et al.*, 2003; Shehada *et al.*, 2015). SiNWs are highly adaptable as their stability and electrical attributes can be altered via molecular engineering. The surface of this nanomaterial can be covalently bonded to or functionalized with several organic compounds (**Figure 7**) (Shehada *et al.*, 2015), such as alkyl side chains (Blase and Serra-Fernández, 2008) or biochemical macromolecules (Chen *et al.*, 2011). This provides a great versatility to these sensors, as they can be tuned or even tailored to fulfill specific applications and extract complete patterns from determined gaseous matrices. The different SiNW FET sensors employed were functionalized with multiple organic chains, and it will be covered in further sections.

In the case of the present thesis, the exceptional traits of cross-reactive SiNW FET sensors have been exploited to analyze controlled gaseous mixtures containing known amounts of specific VOCs as well as real breath samples. An array of these sensors has been designed and employed in an attempt to locate and describe patterns that determined VOCs or breath samples may potentially produce.

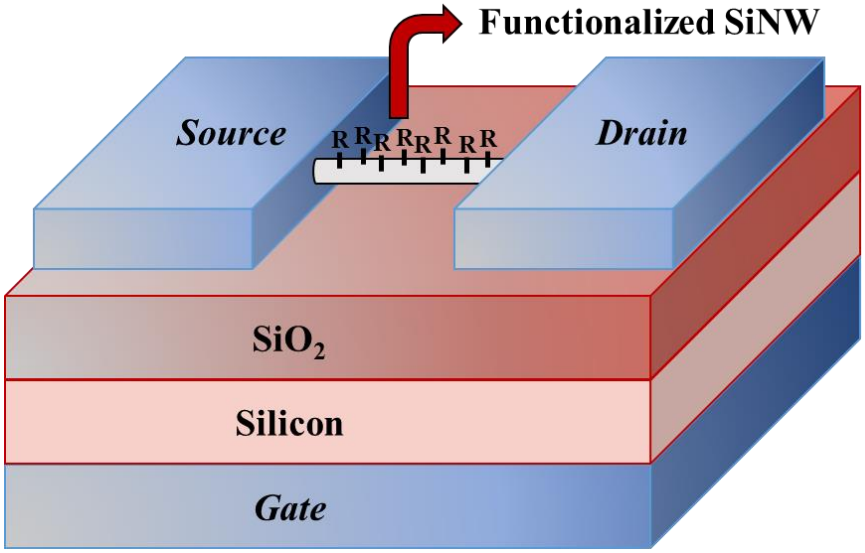

Figure 7. Graphical representation of a SiNW FET sensor. The most important components can be seen: the three electrodes (source, drain, and gate), the semiconducting material (silicon and thermally grown SiO_2), and the functionalizable SiNW (Shehada *et al.*, 2015).

1.4.2.2) Gold Nanoparticle Sensor Arrays

An alternative set of cross-reactive sensors based on gold nanoparticles (GNPs) have also been synthesized and employed to analyze breath samples. These materials, also known as colloidal gold, possess appealing optical, electronic, thermal, and catalytic properties (Guo and Wang, 2007). Back in the day, around the 4th century, GNPs began being used as a method to stain glass. One of the most representative examples of this is the Roman Lycurgus Cup (**Figure 8**), which presented somewhat surprising optical effects due to the colloidal gold employed during its manufacture (Daniel and Astruc, 2004). Therefore, it is possible that the Ancient Romans were the "unexpected" founders of what is currently known as nanotechnology.

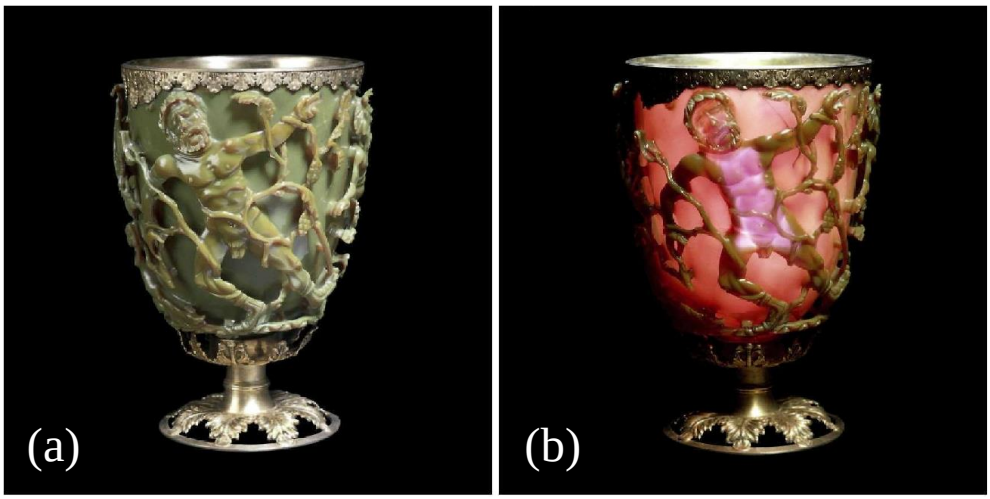

Figure 8. Images of the Lycurgus Cup under different light exposures. **(a)** It is seen green when it reflects light (light source outside) and **(b)** ruby red when it transmits light (light source inside). This phenomenon is caused by the presence of GNPs in the glass (Daniel and Astruc, 2004).

Technological advance, combined with the extraordinary properties of GNPs, have enabled their implementation in a wide variety of fields ranging from physics and chemistry to biology or biomedicine (Guo and Wang, 2007). Being the most stable among all subtypes of metallic nanoparticles (Daniel and Astruc, 2004), GNPs own a broad set of traits, such as possessing a large surface-to-volume ratio, being chemically functionalizable, and having customizable physical and chemical properties, which transform them into an excellent option for the design of cross-reactive sensor nanoarrays (Haick, 2007). A representation of a typical functionalized (monolayer-capped) GNP-based sensor can be seen in **Figure 9** (Tisch and Haick, 2010-a). Once again, common to SiNW FETs, GNPs are very adaptable or flexible as their sensitivity and selectivity can be modified and customized to optimally accomplish specific sensing tasks through their functionalization (Nakhleh *et al.*, 2014).

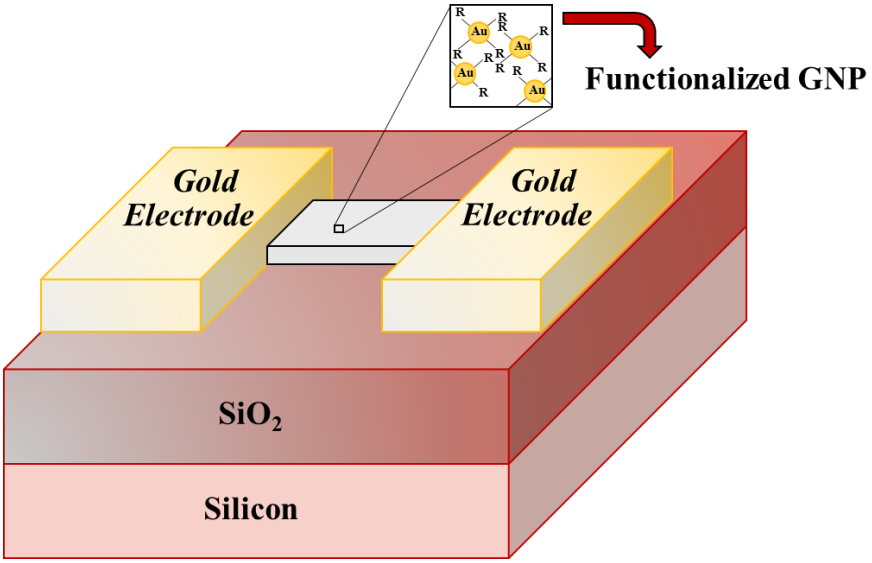

Figure 9. Schematic illustration of a GNP based sensor. The monolayer-capped GNPs are on a film contained between two gold electrodes. The necessary semiconducting materials are also shown (silicon and SiO_2) (Tisch and Haick, 2010-a).

As well as with the SiNW FET sensors, a GNP based sensor array has been synthesized and employed to retrieve relevant information from breath samples, in an attempt to locate patterns that may be able to describe determined diseases and allow their diagnosis non-invasively.

The goal of the analysis with these cross-reactive sensors is to establish these analytical methodologies as viable approaches to translate the information contained in breath into data that can be used to create reliable diagnosing tools for the medical field. If it is proven that these sensors can be used to discover robust patterns that define determined diseases, it would imply that early diagnosis is achievable, and that perfecting these tools would save incredible amounts of money and especially lives.

The three sophisticated analytical techniques described (PTR-MS, SiNW-FET sensors, and GNP sensors) clearly own great characteristics for breath analysis. PTR-MS, on one hand, can be employed to locate important volatile biomarkers that are representative of specific diseases, while cross-reactive sensors, on the other, can provide fingerprint-like information regarding the bulk of the breath sample, with the intention of locating disease-specific patterns. Although they are clearly different, what these approaches do have in common is that they lead to significant amounts of numerical data that ought to be adequately interpreted for these studies

to be of use. In the next and final section of the introduction, we will look into the different mathematical algorithms that have been employed to convert the great amounts of raw data obtained into actual tools or models that, in the end, are the true judges of the feasibility of the methods employed and of the assumption that breath can reflect the clinical condition of a human being (Amman *et al.*, 2014; Buszewski *et al.*, 2007). We are only trying to extract, understand, and take advantage of this "volatile" information.

1.5) Mathematical Analysis – Breath into Numbers

In many occasions during scientific research, enormous amounts of data and databases are produced during experimentation, and the present research is no exception. Regardless of the quality and quantity of this data, or the high-tech, state-of-the-art analytical equipment employed to gather it, if the results are not properly analyzed and interpreted, all the time and resources will have been spent in vain. In this section, the different algorithmic tools employed to extract and understand the relevant information contained within the databases obtained will be described. Fundamentally, it will cover two themes: feature selection and non-linear modeling based on artificial neural networks. These two mathematical tools can be combined in two-step analyses, where the first one is employed to identify the most relevant independent variables to solve classificatory or estimative problems, and the second to create reliable models based on computational artificial intelligence which will employ these variables to carry out the desired classifications or estimations. Both of these significant sets of algorithmic "problem solvers" will be described next, as well as their background inside the scientific theme.

1.5.1) Feature Selection – Where is the Useful Information?

Feature selection (FS) is a statistical procedure that is meant to discover a subset of relevant attributes from a larger dataset for the construction of a succeeding model (Guyon and Elisseeff, 2003). The goal of this process can be summarized into three points: (**a**) improving the performance of mathematical models such as estimators, classifiers, or predictors, while avoiding overfitting; (**b**) originating faster models with lower computational requirements (shorter training times); (**c**) aiding in the comprehension of the basic process behind the production of the data (Chandrashekar and Sahin, 2014; Guyon and Elisseeff, 2003). There are many occasions during research that databases which contain hundreds, thousands, or even hundreds of thousands of variables are generated. In many of these circumstances, these variables or features noticeably surpass the amount of available samples. An obvious example of this is gene expression microarray analysis, where typical amounts of variables (genes) range from 6 to 60 thousand, while the amount of samples (people) is, in some of the greatest of scenarios, only in the hundreds (Guyon and Elisseeff, 2003; McLachlan *et al.*, 2004). Therefore, selecting or identifying the most significant genes before attempting to design a mathematical model to solve, for instance, a disease classification, becomes essential. In the present work, although not so drastic, the amount of variables provided by the analytical equipment employed (PTR-MS and cross-reactive sensor arrays) is still elevated when compared to the amount of gaseous or breath samples analyzed, thus requiring a prior "filtering" phase, which is carried out via FS.

Although the word "filtering" may seem to have been selected arbitrarily, it was chosen because this preliminary analysis of the data was carried out through different filter-based FS

algorithms, which analyze the variables independently, and discard the least informative ones for a determined task. In other words, filter methods do not take into account potential redundant information that different variables may possess, the reason why they are mainly used as a fast pre-processing tool (Zhang *et al.*, 2011). In contrast, the other main family of FS methods are known as wrappers, which do compare variables when performing the selection, allowing the elimination of correlated variables. Nevertheless, the computational time greatly increases with the amount of variables, and as these calculations have been employed as an initial analysis, wrapper methods were left aside (McLachlan *et al.*, 2004; Zhang *et al.*, 2011).

To begin analyzing the data that the analytical equipment provides, five different filter-based FS algorithms have been employed, all of which possess their own traits and selection criteria. Their purpose is to locate the most relevant variables (volatile compounds in the case of the PTR-MS studies or sensing features during the cross-reactive sensor array analyses) which are able to solve a particular problem. The set of five FS algorithms employed can be seen in **Table 3**, with highlights regarding their main characteristics.

Table 3. Set of five filter-based FS methods employed to begin the data analysis, as well as the main traits of each algorithm and a reference where it has been used and/or detailed.

Feature Selection Method	Main Statistical Characteristics	Reference
χ^2 Score	This test is applied as a FS to attain an estimate about the independence of two events. It is employed to determine whether the occurrence of a specific feature and of a specific class are independent or not.	Liu and Setiono, 1995
Fisher's discriminant ratio	This strategy is based on analyzing the mean and deviation of the values of the variables of different classes. Bigger mean differences between classes and lower value scatters within a class imply a better score for the feature.	Wang *et al.*, 2011
Kruskal-Wallis	It is a non-parametric approach that relies on the comparison of the medians of the groups to rank the discriminative power of the features.	Kruskal and Wallis, 1952
Relief-F	To determine the quality of features, Relief-F searches for the nearest neighbors of a sample within its same class and the different existing classes, in order to give priority to those features or variables which better distinguish the sample from their nearest neighbors in different classes.	Wu *et al.*, 2013
Information gain	This FS method measures the dependence of a feature and its class label through the information theory. It measures the variation of entropy (system disorder) when the feature is present or not.	Dhir *et al.*, 2007

The use of these data filters on large databases provides an ordered list of variables in terms of discriminative power, according to different statistical criteria corresponding to each one of the FS algorithms. This initial data analysis phase enables locating the relevant and useful information-containing features for the construction of further mathematical models with a constricted set of variables. These models represent the second step of the data treatment, and they are based on artificial neural networks.

1.5.2) Artificial Neural Networks – Giving Applicability to Volatolomics

Once the most important features is identified to solve a given matter, and it is necessary to further treat the data using suitable mathematical models to reach a final application. Typically, linear models such as partial least squares or principal components analysis are implemented, due to being simple, straightforward, and easy to create and interpret (Aroca-Santos *et al.*, 2015; Frank and Friedman, 1993). Nevertheless, there are many cases during research in which these basic approaches are just not enough to solve complex systems with loads of information, and a clear example of this is breath analysis. The thousands of compounds present in breath make it obvious that the mathematical modeling task is not going to be effortless, and the reason why the more sophisticated non-linear artificial neural networks (ANNs) have been the chosen algorithms.

Back in the 1940s, and inspired by the mechanism of biological neurons, simplified versions known as "artificial neurons" were presented as models of their biological brothers that were capable of carrying out computational activities (McCulloch and Pitts, 1943). This discovery implied that the seed for the development of an entirely new subsection within computational artificial intelligence had been planted and it was ready to grow. Years later, these artificial units were further arranged into more complex models known as perceptrons, which instigated great interest due to their ability to recognize simple patterns and represent the first ANN (Rosenblatt, 1958). Nevertheless, the expansion and enthusiasm that these algorithms were generating was hindered by the publication of a set of deficiencies or limitations within these perceptrons (Minsky and Papert, 1969). It was not until the 1980s that the interest in ANNs began to exponentially rise again, due to in great measure the discovery of the back-propagation learning algorithm for multilayer perceptrons (MLPs), which enable resolving systems that are much more complex than those that simple perceptrons can handle (Jain *et al*, 1996).

These, at the time, revolutionary algorithms are the most employed and implemented kind of ANN due to their reliability and relatively simple training process. MLPs have been widely implemented in a broad number of fields, ranging from chemical (Cancilla *et al.*, 2014-a; Roosta *et al.*, 2012), biomedical (Guo *et al.*, 2010; Webster *et al.*, 2009), and food technology research (Aroca-Santos *et al.*, 2015), to industry (Geem and Roper, 2009) or even economics (West *et al.*, 2005). In all these fields, MLPs have been able to model intricate systems, leading to compelling applications of diverse natures. All these reasons have guided MLPs into the core of the mathematical treatment of the present thesis, turning them into the main character of the entire data analysis.

A MLP is a supervised type of algorithm, which implies that it requires target data to fulfill its training process. In other words, samples or data points have to be labeled with their dependent variable values in order to properly complete the optimization of the mathematical tool (Basheer and Hajmeer, 2000). In contrast, there are other ANNs that follow an unsupervised training procedure, the most common and famous being the self-organizing maps (SOMs) or Kohonen's networks (presented by Teuvo Kohonen in the 80s) (Kohonen, 1989). SOMs, alike any unsupervised model, only require independent variables to be trained as they attempt to cluster different samples into groups according to the relations that may exist between the variables of the samples (Basheer and Hajmeer, 2000).

Returning back to MLPs, as their name suggests, they are characterized by a layered topology or architecture. There are three different kinds of layers that form these algorithms: input, hidden, and output layers. The input layer is formed by nodes, which are strictly responsible

for introducing the independent variables into the model. There will be as many nodes as independent variables employed. On the other hand, the hidden and output layers are both comprised of neurons ("artificial neurons"), which are the actual processing units where the non-linear calculations take place. The amount of hidden neurons should be optimized for each MLP to function properly, as low amounts may lead to models with a poor learning capability, and high amounts may tend to produce over-fit models (only accurate for the data used to train it, implying low generalization ability). Finally, the amount of neurons in the output layer will coincide with the quantity of dependent variables that will be estimated (Cancilla *et al.*, 2014-a). In **Figure 10**, an example of a typical MLP can be seen to fully understand their topology.

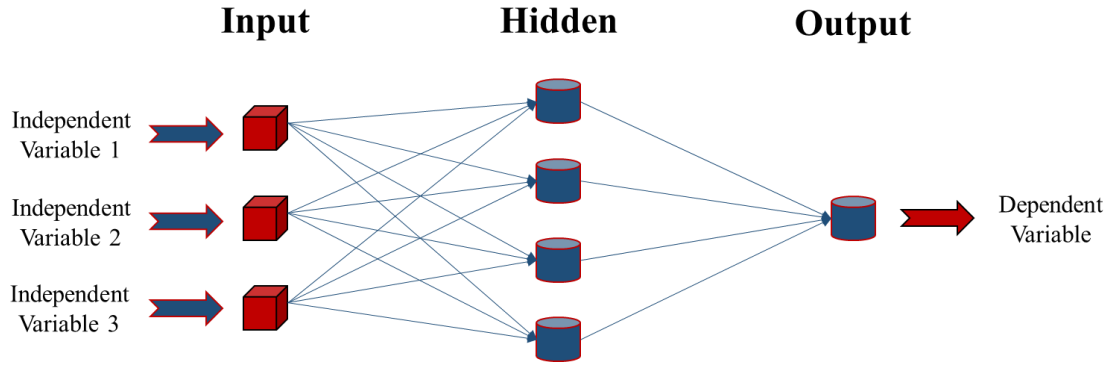

Figure 10. Representation of a hypothetical MLP with the following topology: three input nodes, four hidden neurons, and a single output neuron.

As can be seen in **Figure 10**, every unit in a determined layer is connected to all the units in its neighboring layers (represented by the blue arrows in **Figure 10** that connect nodes with neurons or neurons with other neurons). Each one of these connections is governed by a weighted coefficient or weight. Initially acquiring a random value between zero and one, these weights are optimized during the training or learning process of a MLP, with the goal set to increase the accuracy of the designed model, whether it acts a classifier or an estimator (Cancilla *et al.*, 2014-a). Once these weights have been optimized, and the model has been statistically validated correctly, the mathematical tool is theoretically ready to offer reliable results.

In our specific scenario, the MLPs will be trained with data obtained from the PTR-MS studies or the cross-reactive sensor arrays, with a diverse set of goals which will be further detailed in the following blocks of the thesis. Now that the main traits and characteristics of the mathematical tools that have been used to analyze and process the information contained in exhaled breath have been covered, all of the primary steps in this research have been described. To conclude this introduction, a brief summary of the key "pillars" that support this work is shown in the next and last subsection.

1.6) Summary and Objective

Along this introduction, the five principal themes or fields touched during this research have been presented: **breath, biomarkers, diseases, analytical equipment, and mathematical analysis**. They have been described and linked together to reach one main goal: **obtain a proof-of-concept that it is possible to design reliable and non-invasive disease diagnosing tools based on breath analysis through cutting-edge analytical methodologies and intelligent mathematical modeling**. A graphical summary of all the phases is shown in **Figure 11**.

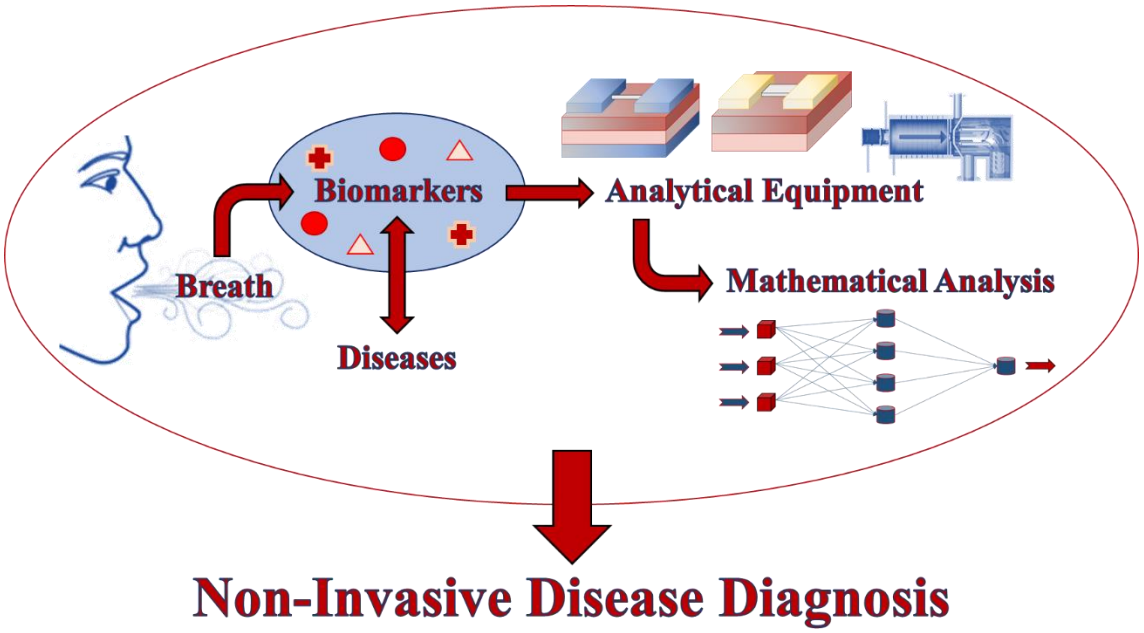

Figure 11. Graphical summary showing the main themes covered during the research as well as the primary goal.

These five main blocks are vital to this research as they can be used to answer the most important questions: "where", "why", "what", "who", "how", and "when". "Where" is answered by **breath**, as it is *where* metabolic information about a patient can be found. "Why" is answered by **biomarkers**, as they are the reason *why* breath contains discriminative information regarding different diseases. "What" is answered by **diseases**, as they are *what* is needed to be diagnosed as early as possible. "Who" is answered by the analysts, *who* use the **analytical equipment** that is in charge of translating the information that breath provides. "How" is answered by the **mathematical analysis**, as it is *how* the information is deciphered and interpreted to attain the final diagnosing tools. And, finally, "when" is answered by… Well, there are no more protagonists to answer *when*, yet the answer is clear: **early diagnosis to save lives is needed NOW**.

As all the pieces of the puzzle have been presented and described, it is time to begin with the next section of the thesis, where the different equipment and methodologies employed as well as more details regarding the specific protocols followed, both analytical and mathematical, will be thoroughly explained.

2) Materials and Methods

Now that the background of the main themes that cover this research have been described, it is time to look into the specific methodologies which have been employed. In this section, a detailed description of the breath collection method and protocol, the analytical equipment utilized, and the mathematical algorithms used can be found.

2.1) Breath Gathering

As mentioned in the introduction, the biological matrix that has been relied on to extract metabolic information from different groups of patients was breath. For this reason, the development of a reliable system, that can capture breath samples in a stable fashion, must be employed. The breath collecting equipment used in all the studies that involved breath analysis was developed in the laboratory for nanomaterial based devices (LNBD) at Technion (Haifa, Israel) (LNBD website, 2016), and a schematic representation highlighting all its main components can be seen in **Figure 12**.

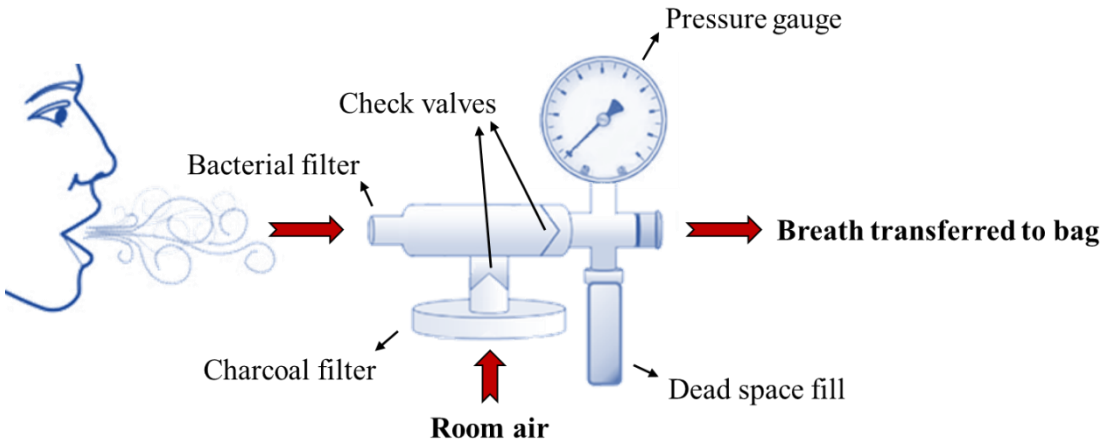

Figure 12. Representation of the breath collecting system.

During breath collection, it is necessary to retrieve the alveolar breath from the subjects, free of exogenous and contaminating VOCs. The VOCs that truly provide metabolic or disease-related information are the endogenous VOCs (Mukhopadhyay, 2004), the reason why it is crucial to clear or separate as many interfering compounds as possible. Therefore, a series of precautions have been followed in order to collect in a controlled manner the exhaled alveolar breath samples. All the breath tests for a particular study were carried out in the same room, or, in other words, under the same atmosphere, to rule out potential location influences or confounding factors on the measurements. Also, people that had consumed food, drinks, or tobacco within two hours from the time of the test, were excluded to avoid samples that may be heavily contaminated. Finally, a meticulous "lung wash" to eliminate ambient contaminants was performed. It consisted of a 3-5 minute procedure in which the tested individual had to continuously inhale through a mouthpiece containing two different filters, bacterial and charcoal (activated carbon to adsorb contaminating VOCs), which eliminated about 99.99% of compounds in inspired air and a great part of other exogenous VOCs and bacteria (Risby and Solga, 2006). This mouthpiece was acquired from Eco Medics, Duerten, Switzerland (Peng *et al.*, 2009).

Immediately after the lung wash, the subjects exhaled into the breath collecting device (**Figure 12**) against 10-15 cm H_2O of pressure (7-11 mm Hg) to ensure closure of the velum (soft palate) and avoid potential nasal contaminants. Exhaled breath is composed of respiratory dead space air (volume of inhaled air that does not undergo gas exchange) and the relevant alveolar air, and, therefore, these parts had to be separated. The device possesses two exit ports, one for the dead space air, which is guided into a plastic bag, and another for the alveolar breath. The second port directs the alveolar air into a chemically inert 750 mL Mylar sampling bag (Eco Medics). Both bags are filled in a single step, which implies that bags did not have to be changed in the process. The content of the second bag (alveolar air) was transferred into a Tenax TA and Carboxen-1080 glass adsorbent tube or into a two-bed ORBOTM 420 Tenax TA sorption tube (Sigma-Aldrich, Saint Louis, MO, USA) using a vacuum pump. The safely sealed tubes containing the breath samples were stored at 4ºC and controlled relative humidity until their further analysis, which was always carried out considerably before three weeks, which is when it has been determined that these samples start to deteriorate as the results they provide begin showing a correlation with the storage time beyond this point (Peng *et al.*, 2008; Peng *et al.*, 2009).

2.2) Proton Transfer Reaction-Mass Spectrometry

Proton transfer reaction-mass spectrometry (PTR-MS) is an exceptionally sensitive methodology that is intended for real-time identification and quantification of VOCs, without the need to preconcentrate the volatile compounds in the gaseous samples (Blake *et al.*, 2009; Ligor *et al.*, 2009). In this section, the details and particular specifications of the instrument employed will be covered.

During this research, a "high-sensitivity PTR-MS" from Ionicon Analytik (Innsbruck, Austria) was employed (Ionicon website, 2016-b). It is a type of PTR-QMS, where the "Q" stands for quadrupole, as the detection and quantification of the VOCs is based on quadrupole mass spectrometry (Dawson, 1976). In **Figure 13**, a representation of the equipment can be found.

It is worth mentioning that when compared to standard PTR-MS systems, the one employed during this research possesses about a six times lower detection limit (six times more sensitive), at around 5 pptV. This is the reason why it is referred to as a high-sensitivity PTR-MS. Regarding the quadrupole mass spectrometer, the model that had been implemented into the device was a Pfeiffer QMG 422 (Pfeiffer Vacuum, Germany). The operational conditions employed during the measurements, which led to the real-time detection of VOCs, can be seen in **Table 4**.

These conditions were tuned and finally selected to reach the optimal performance of the system. The PTR-MS operated in vacuum, and the dwell time for each mass (time it takes to detect and quantify each compound, which are separated according to their mass/charge ratio (m/z)) was set to maximize efficiency, as it is long enough to determine the concentrations of the volatile compounds as well as quick enough to enable repetitions and online measurements.

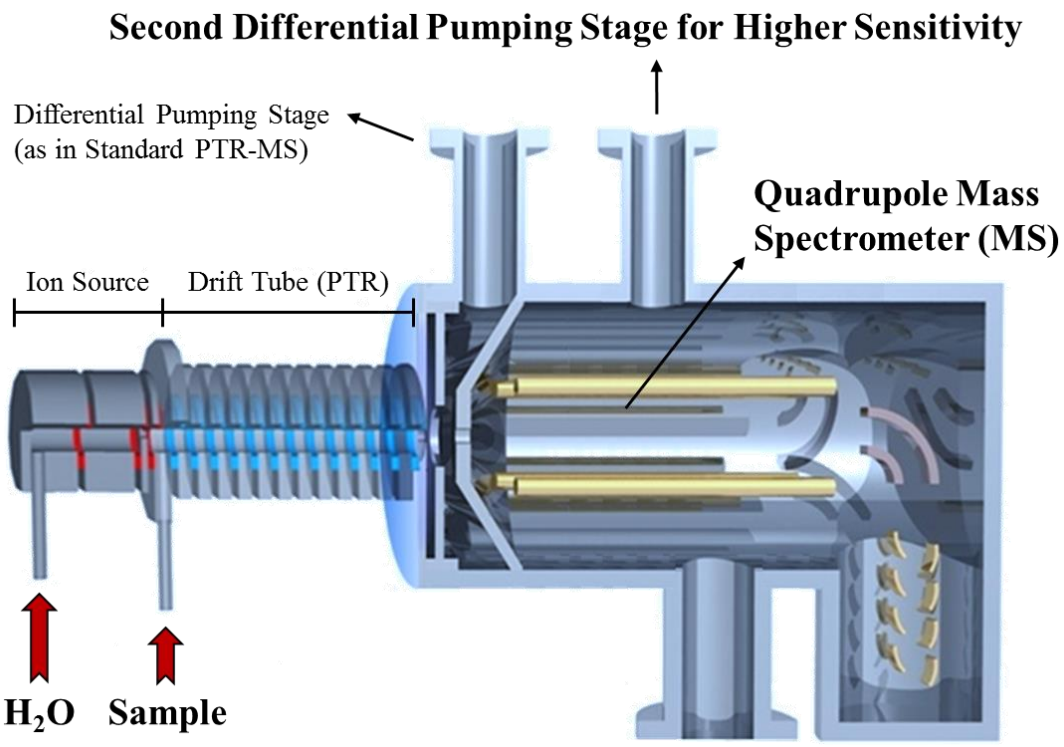

Figure 13. Schematic representation of the high-sensitivity PTR-MS device employed, which contains a quadrupole-based mass spectrometer (Ionicon website, 2016-b).

Table 4. Operating conditions of the PTR-QMS.

Parameter	Value
Drift tube pressure	2.2 mbar
Drift tube temperature	60ºC
Drift tube voltage	600 V
Sample inlet temperature	80ºC
Detector pressure	$9.2\text{-}9.2\cdot10^{-6}$ mbar (vacuum)
Dwell time for each mass	500 msec

2.3) Cross-Reactive Sensor Arrays

An alternative approach to analyze the gaseous matrices is through the use of cross-reactive sensors, which provide global information of the entire sample, as a whole, rather than offer evidence from particular compounds like typical lock-and-key sensors (see **Figure 6** in section 1.4.2 for graphical comparison) (Röck *et al.*, 2008; Tisch and Haick, 2010-b). Furthermore, the use of arrays, instead of single sensors, provides a greater amount of potentially

useful and complementary data. This leads to a "fingerprinting" methodology, which labels entire samples with recognizable patterns. As already mentioned, two different approaches have been followed, and they will be presented in the next subsections.

2.3.1) Silicon Nanowire Field-Effect Transistor Sensor Arrays

The relevant steps during the fabrication of these sensors are the silicon nanowire (SiNW) synthesis, the SiNW field-effect transistor (FET) preparation, and the molecular functionalization of the SiNW. These three steps are described next.

Although the protocol followed to synthesize the p-type SiNWs and to fabricate the SiNW FETs has been optimized and described in the past (Assad *et al.*, 2012; Wang and Haick, 2013-a; Wang and Haick, 2013-b), the most relevant steps are the following. The SiNWs were prepared on a silicon wafer (thin slice of semiconducting material) through chemical vapor deposition (SiH_4 and B_2H_6 as precursor gases (B/Si ratio 1:20000) and gold as the growth catalyst), which originated nanowires with limited dimensions of 40±8 nm in diameter and of 8.5±1.5 µm in length. The as-grown SiNWs (without being further modified) majorly consisted of single-crystalline silicon cores coated with a layer of native SiO_x of 5±1 nm. These as-grown materials were then first treated in buffered hydrofluoric acid for 15 seconds and afterwards with a solution of $KI/I_2/H_2O$ (mass ratio 4:1:40) for 2 minutes to remove gold (catalyst and potential surface contaminants) and the layer of SiO_x. Next, the SiNWs were dispersed in ethanol using ultrasonication during 6 seconds, and the resulting SiNW suspensions were spray coated onto a previously cleaned silicon substrate known as "p-Si(100)" (p-type; 0.001 Ω cm resistivity; 300 nm SiO_2; 10 titanium nm/200 gold nm gate electrode). Additionally, the source and drain electrodes (30 titanium nm/110 gold nm) were added onto the sprayed SiNWs using photolithography (Karl Suss MA6 Mask Aligner) and lift-off processes, and were placed to create a channel with a 2 µm width, where the SiNWs are exposed and connect these two electrodes (Assad *et al.*, 2012; Wang and Haick, 2013-a).

Once the SiNW FETs were fabricated, the nanowires were molecularly functionalized with organic chains to attain 12 different sensors. First, the surfaces of the devices were activated through a 30 or 60 second oxygen plasma treatment, and after this phase, three different approaches were followed to reach the final SiNW FET sensors:

1. **Two-step silane-acyl chloride modification**: this synthesis method begins with a 60 minute immersion in a 20 mL solution of 3-aminopropyl-triethoxysilane (APTES) (10 mM) in dehydrated ethanol at room temperature. This process resulted in APTES-terminated SiNW FETs, which were then meticulously rinsed with acetone, ethanol, and isopropanol, and dried using a flow of N_2. At this point, the SiNW FETs are ready to undergo the molecular modifications, which were carried out by submerging them for 17 hours in a solution of a specific acyl chloride (10 mM) in chloroform containing catalytic amounts of triethylamine (Wang and Haick, 2013-a). The particular acyl chlorides employed are the chemicals that determine the final and unique nature of each sensor. Seven acyl chlorides were employed, which led to seven distinct sensors (**S1**-**S7**). These compounds were 5-phenylvaleric chloride ($C_{11}H_{13}ClO$; **S1**), 1,4-butanedicarbonyl chloride ($C_6H_8Cl_2O_2$; **S2**), methyl adipoyl chloride ($C_7H_{11}ClO_3$; **S3**), hexanoyl chloride ($C_6H_{11}ClO$; **S4**), heptanoyl chloride ($C_7H_{13}ClO$; **S5**), decanoyl chloride ($C_{10}H_{19}ClO$; **S6**), and dodecanoyl chloride ($C_{12}H_{23}ClO$; **S7**). It must be noted

that the SiNW FETs that were treated with 1,4-butanedicarbonyl chloride were afterwards immersed into hot water (90ºC) for two hours in order to hydrolyze the acyl chloride end group. Finally, the functionalized SiNW FETs were rinsed with acetone, ethanol, and isopropanol, as well as dried employing a flow of N_2, just like after being treated with APTES.

This method led to the synthesis of seven different SiNW FET cross-reactive sensors (**S1-S7**), each one possessing its own chemical properties and structure. In **Figure 14**, a representation of the final surface of the SiNW FETs can be seen, as well as the particular head of the organic chains bonded to each sensor (**S1-S7**). As well, back in the introduction section, in **Figure 7** (section 1.4.2.1), a graphic illustration showing the appearance of these sensors can be found.

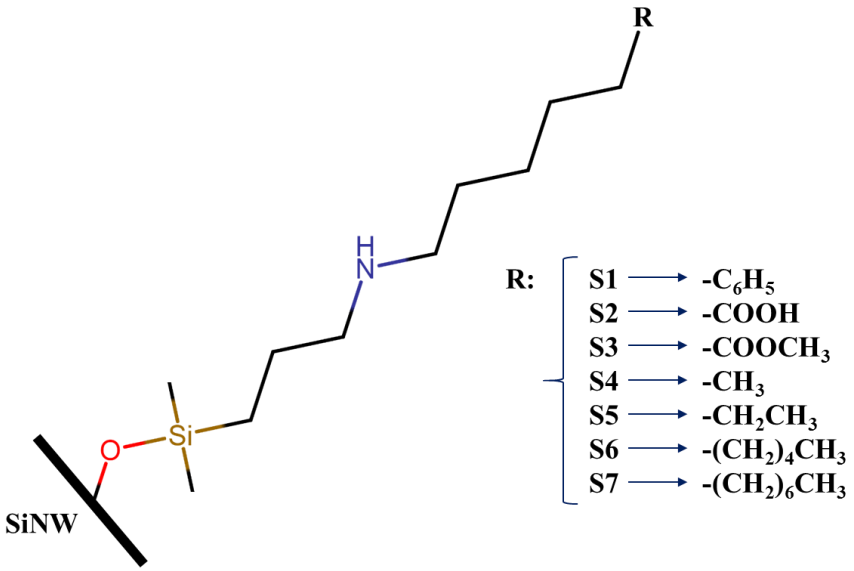

Figure 14. Scheme of the molecularly functionalized surfaces of seven of the SiNW FET sensors synthesized.

The chains that are linked to the surface of the SiNWs form monolayers, and depending on the head of the chain (Wang and Haick, 2013-a), as well as its length (Wang and Haick, 2013-b), they will possess different chemical properties and therefore originate distinct responses when interacting with gaseous samples. Sensors **S1** through **S4** possess similar chain lengths, but different functional groups (heads), which enables the comparison of the sensing capability according to them (Wang and Haick, 2013-a). While **S1** and **S4** are electron-donating, **S2** and **S3** are electron-withdrawing, which will clearly have an effect over the sensing process. On the other hand, sensors **S4** through **S7** have the same functional group, yet increasing chain lengths, which allows assessing the effect of the amount of carbon atoms in the backbone of the chains on the sensor-sample interaction (Wang and Haick, 2013-b).

2. **Single-step silane modification**: to carry out this type of synthesis, the activated surfaces of the SiNW FETs were immersed 10 mL of a 2 mM silane/chloroform solution for 45 minutes at room temperature, followed by a sequential chloroform, acetone, ethanol, and isopropanol rinsing process. At the end, the functionalized sensors were dried using a N_2 flow. In this case, four different silanes were used to reach the final devices. These silanes were trichloro(3,3,3-trifluoropropyl)silane

($CF_3CH_2CH_2SiCl_3$; **S8**), trichloro(phenethyl)silane ($C_6H_5CH_2SiCl_3$; **S9**), (3-bromopropyl)trichlorosilane ($C_3H_6BrCl_3Si$; **S10**), and APTES ($C_9H_{23}NO_3Si$; **S11**). It must be noted, that during the synthesis of **S11** (functionalized with APTES), it was carried out with ethanol instead of chloroform during both the functionalizing and rinsing steps (Wang and Haick, 2013-a).
This second methodology originated four very different SiNW FETs with unique sensing properties (**S8-S11**).

3. **Two-step silane-monomer modification**: this last approach or synthesis method starts with the immersion of the activated devices in 10 mL of a 2 mM trichloro(3,3,3-trifluoropropyl)silane/chloroform solution for one hour. Next, the resulting surface is successively rinsed with chloroform, acetone, ethanol and isopropanol, and then dried with a flow of N_2. Once it is completely dry, an anthracene ($C_{14}H_{10}$) monomer solution in tetrahydrofuran (THF; catalytic amount) is drop-casted onto the surface and preserved in a vacuum oven overnight at 55ºC. Finally, it is rinsed with THF, acetone, ethanol, and isopropanol, and again dried with N_2, leading to the completion of the synthesis of the last, anthracene-functionalized, SiNW FET sensor (**S12**) employed during this research (Wang and Haick, 2013-a).

As a precaution, every one of the molecularly modified SiNW FET sensors (**S1-S12**) were characterized through X-ray photoelectron spectroscopy to evaluate the amount of functionalized sites (monochromatized X-ray Al Kα 1486.6 eV source; Thermo VG Scientific, Sigma Probe, England;) and ellipsometry to determine the thickness of the molecular layers (spectroscopic ellipsometer, M-2000 V; J. A. Woollam Co., Inc., Lincoln, NE, USA) (Wang and Haick, 2013-a; Wang and Haick, 2013-b). Finally, some of the sensors (**S1-S7**) were loaded into a steel chamber of approximately 170 cm^3, enabling the extraction of information from every sensor at the same time, or, in other words, operating as a cross-reactive sensor array. The chemical compounds required during the synthesis of the sensors were purchased from Sigma-Aldrich.

2.3.2) Gold Nanoparticle Sensor Arrays

In addition to SiNW-based FET sensors, a broad series of functionalized gold nanoparticle (GNP) sensors (sensor array) have also been designed to perform breath analysis. Their synthesis process will be explained in the current section.

The functionalized GNPs were prepared following an established protocol based on a modified two-phase method (Brust *et al*., 1995), which resulted in monolayer-capped 5-nm gold nanoparticles. The difference between sensors relied on the organic capping layer (molecular functionalization), which provided unique chemical properties to each of the sensors synthesized in the end and enables the extraction of different and, perhaps, complementary information. The synthesis of the monolayer-capped GNP began with the transfer of $AuCl_4^-$ from a 25 mL aqueous solution of $HAuCl_4 \cdot xH_2O$ (31.5 mM) to an 80 mL toluene solution (organic phase) using the phase transfer catalyst tetraoctylammonium bromide (34.3 mM). Once the organic phase had been separated, thiols in excess were added into the solution. Each particular thiol (hexanethiol, decanethiol, dodecanethiol, octadecanethiol, 3-ethoxythiophenol, and 4-chlorobenzenemethanethiol) led to a specific molecular functionalization, and combined with number of layers of thiol, their thickness, density, and percentage of area covered, it originated

the synthesis of 34 different GNPs with their own chemical and sensing properties. All of these design variables will allow the determination of the best sensors and chemistry for specific applications. Then, the thiol-containing solution was vigorously stirred for 10 minutes, and, afterwards, 25 mL of an aqueous solution containing a great excess of ice-cooled $NaBH_4$ (0.4 M) was included due to being a versatile reducing agent. This chemical reaction was carried out at room temperature and stirred for a minimum of 3 hours, which led to a dark brown solution which contained thiol-capped GNPs (organically functionalized). The solvent was removed in a rotatory evaporator, and the final GNPs were thoroughly washed with ethanol and toluene. Finally, the functionalized GNPs were purified through repeated extractions to eliminate free thiol ligands (Peng *et al.*, 2009). The chemical compounds mentioned were all acquired from Sigma-Aldrich.

To finally develop the GNP sensors, the following steps were fulfilled. First of all, for each sensor (chemiresistor), 10 pairs of circular interdigitated gold electrodes were deposited by an electron-beam evaporator TFDS-870 (Vacuum Systems and Technologies, Petah Tikva, Israel) onto a segment of quality silicon wafer covered with 300 nm of thermal oxide (Silicon Quest International, Reno, NV, USA). The external diameter of the entire circular electrode was 3000 µm, while the gap between the two contiguous electrodes, as well as their own width, were both 20 µm. Once the electrodes were ready, the previously synthesized functionalized GNPs were first dispersed in chloroform by sonicating, and then drop-casted onto the surface of the electrodes. When the solution was covering them, they were blown dry with nitrogen. This phase was repeated several times until a resistance of around 1 MΩ was reached. Afterwards, the devices were dried for two hours at room temperature and finally baked in a vacuum oven at 50ºC all night (Peng *et al.*, 2009). At the end, as a precaution, the sensors were characterized similar to how the SiNW FET sensors were (*vide supra*). A schematic representation of a GNP sensor, like the ones synthesized in this research, can be seen in **Figure 15**. The 10 pairs of circular interdigitated gold electrodes can be found, as well as an amplification of the functionalized nanoparticles.

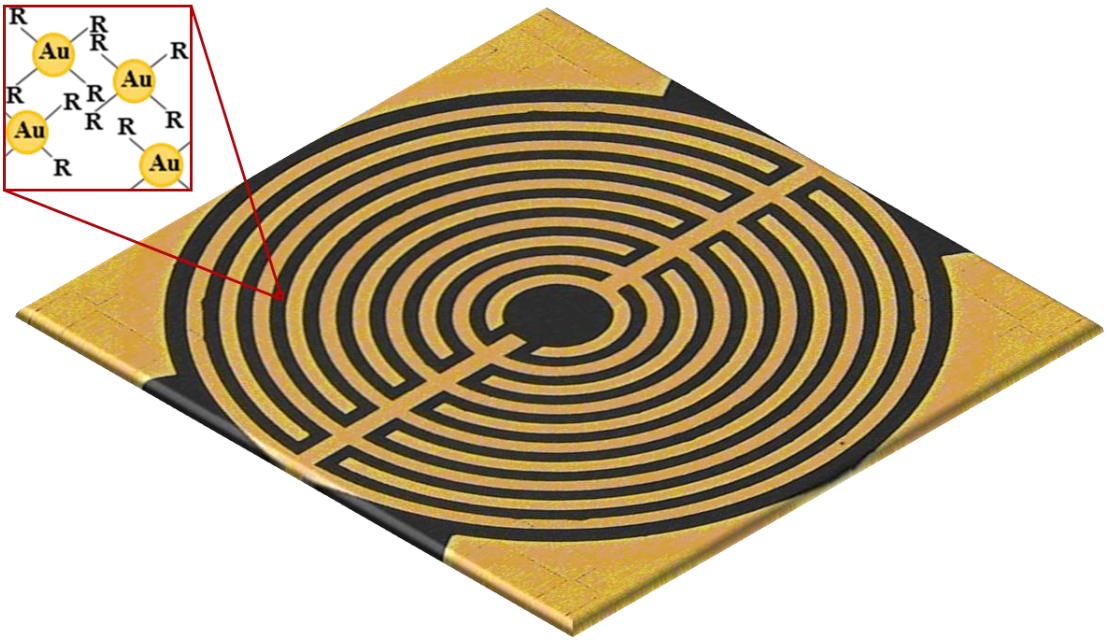

Figure 15. Schematic representation of the molecularly functionalized GNP sensors synthesized.

Once the 34 sensors were prepared, they were mounted onto a custom polytetrafluoroethylene circuit board and located in a stainless steel chamber (100 cm^3 volume) to create the final sensor array.

Both of the types of cross-reactive sensor arrays that have been designed, as well as the analyses carried out with PTR-MS, provide great amounts of information when a gaseous sample is processed. This data will be treated and modeled appropriately to create the tools desired during this research.

2.4) Mathematical Tools and Analysis

In this section, an insight regarding the set of mathematical tools that have been employed in this research will be given, as they play a basic role in any analysis which leads to the creation of substantial amounts of data. They are divided into two fundamental groups: feature selection (FS) algorithms and multilayer perceptrons (MLPs). The first group of algorithms will be in charge of locating the most relevant independent variables from databases to carry out further modeling tasks, which will be done by the non-linear MLPs. A schematic representation of this two-phase calculation can be seen in **Figure 16**. It must be noted that prior to these calculations, statistical outliers were located and removed from the databases to avoid potential alterations during the FS-MLP two-phase calculation (when the value of a variable for a specific sample was lower than Q_1-3×IQR or greater than Q_3+3×IQR, where Q_1 and Q_3 represent the first and the third quartile values, respectively, and IQR symbolizes the interquartile range, it was considered a statistical outlier).

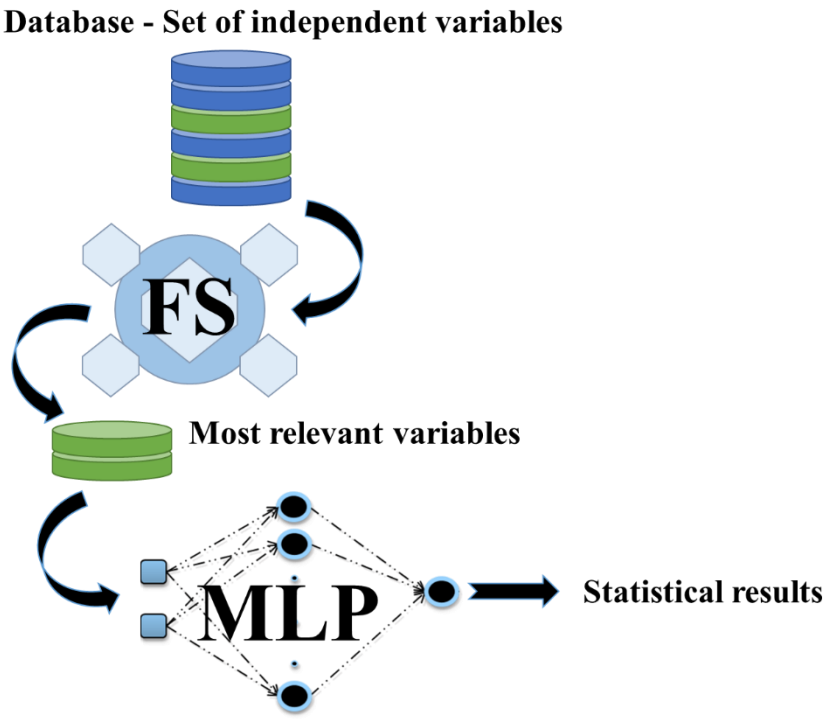

Figure 16. Schematic illustration of the main calculations carried out during the present research. They are based on an initial feature selection (FS) procedure followed by a multilayer perceptron (MLP) modeling phase.

Now that the two basic groups of algorithms that have been used in this thesis have been presented, the first one will be thoroughly described in the next subsection.

2.4.1) Feature Selection

During this research, different analytical approaches have been utilized to gather information concerning gaseous or breath samples, which were PTR-MS and cross-reactive sensor arrays. These methodologies originate large databases, with many variables, after processing the samples. A consequence of this is the need to establish a preliminary mathematical procedure to eliminate noisy signals and/or irrelevant information. In the present case, filter-based feature selection (FS) algorithms have been selected to accomplish this task (Zhang *et al.*, 2011).

Five different supervised data filters have been employed, each one with their own particular mathematical algorithms and selection criteria. In the case of PTR-MS data, the FS process located specific m/z (potential biomarkers) that had the greatest discriminative power. On the other hand, for databases originated during the cross-reactive sensor array studies, they determined the best sensing features for a successive modeling task. The way these filter-based FS methods operate is by ordering the variables, according to their specific mathematical calculations, in terms of discriminative power. They can be employed, for instance, to classify samples into two groups or classes which are properly labelled, as the FS algorithms employed are supervised (e.g., patients with different diseases). This enables the selection of only the most useful independent variables, which in the end allow reducing computational load and time, as well as improving prediction performance of further mathematical modeling (Chandrashekar and Sahin, 2014).

As mentioned, five different FS methods have been used during the data analysis (see **Table 3** in section 1.5.1), and they are based on χ^2 score (Liu and Setiono, 1995), Fisher's discriminant ratio (Wang *et al.*, 2011), Kruskal-Wallis' analysis (Kruskal and Wallis, 1952), relief-F algorithm (Wu *et al.*, 2013), and information gain test (Dhir *et al.*, 2007). The essential mathematical traits and calculations that are used by these data filters to quantify the discriminative power of different variables will be described in the following subsections.

It must be noted that the first four FS methods were carried out using a software package that has been programmed in Matlab language during this research (the code has been created manually), and Matlab version 7.0.1.24704 (R14) was employed to perform the calculations. The information gain test was carried out through Orange 2.7 Data Mining software (Demsar *et al.*, 2013).

2.4.1.1) χ^2 Score

χ^2 (chi^2) calculations, based on classic χ^2 statistics, are employed in FS as a test of independence to evaluate if a class or group label depends on the value of a specific feature or not. The χ^2 score obtained by a feature that has r samples and C classes is calculated using **Equation 1**.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{C}\frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (1)$$

In this equation, n_{ij} represents the number of samples with the i^{th} feature value, and μ_{ij} is obtained through **Equation 2**.

$$\mu_{ij} = \frac{n_{*j}n_{i*}}{n} \quad (2)$$

Where n_{*j} symbolizes the amount of samples in class j, n_{i*} is the number of samples with i^{th} value for a specific feature, and n is the amount of total samples. Relatively low χ^2 scores for a specific feature imply that the values of that feature possess discriminative information for the classes labeled, indicating that it is an adequate variable for following classification models (Liu and Setiono, 1995; Zhao *et al.*, 2011). The χ^2 scores can be ordered to determine the variables or features with the most discriminative power.

2.4.1.2) Fisher's Discriminant Ratio

This filter-based FS method uses linear calculations to determine the discriminative power of a variable. It is widely employed due to its relative simplicity, and it operates by searching for a line that can separate the data samples into their corresponding classes the best way possible. In mathematical terms, the Fisher score is obtained through **Equation 3**, which is Fisher's discriminant ratio (FDR).

$$FDR = \frac{((\overline{x_1}) - (\overline{x_2}))^2}{Var(x_1) + Var(x_2)} \quad (3)$$

In this function, $\overline{x_1}$ and $\overline{x_2}$ represent the means of the values of a certain feature for classes x_1 and x_2, respectively, while $Var(x_1)$ and $Var(x_2)$ are the variances of these datasets. Therefore, a variable that possesses an elevated discriminative power according to this test, will reflect relatively high FDR values, as the means of each group should be different, and the samples within each class should not be overly scattered (Wang *et al.*, 2011). These results will directly enable the location of the best features for further modeling phases.

2.4.1.3) Kruskal-Wallis Test

The Kruskal-Wallis FS method relies on non-parametric calculations to rank the features by comparing the medians of the different classes. It is able to interpret non-linear relations between the values of the variable evaluated and the class label, and determines whether the medians of the values of a feature of two or more classes are equal or not to rank them in terms of discriminative capacity. The calculation carried out is shown in **Equation 4**.

$$KW = \frac{12}{N(N+1)}\sum_{i=1} n_i(\overline{r}_i)^2 - 3(N+1) \quad (4)$$

Where N is the amount of observations or samples in all the groups, n_i is the number of observations in group i, and $\bar{r}_i$ represents the mean of the ranks of observations in group i. It is a complex calculation that leads to the comparison of the medians of the values of the different classes assessed, enabling the selection of those features with greater discriminative power (Kruskal and Wallis, 1952; Zhao *et al.*, 2011).

2.4.1.4) Relief-F Algorithm

This FS method is based on evaluating features by the extent of their ability to distinguish the values of instances or samples that are near to each other. When analyzing a sample value, it seeks for the nearest neighbors, one per class (same and different), and adjusts the feature weighting vector to enable ranking variables according to their ability to discriminate neighbor samples from other classes. The function used to obtain the relief-F score is shown in **Equation 5**.

$$R_F(f_i) = \frac{1}{2}\sum_{t=1} d(f_{t,i} - f_{NM(x_t),i}) - d(f_{t,i} - f_{NH(x_t),i}) \quad (5)$$

In this equation, $f_{t,i}$ represents the value of the sample analyzed ($\mathbf{x_t}$) of a specific feature ($\mathbf{f_i}$), while $f_{NM(x_t),i}$ and $f_{NH(x_t),i}$ are the values of the i^{th} feature corresponding to the nearest neighbors of different and same classes, respectively. Finally, $d(\cdot)$ is the function employed as a distance measurement between the sample and the nearest neighbors (Wu *et al.*, 2013; Zhao *et al.*, 2011).

2.4.1.5) Information Gain Test

Information gain is a commonly employed type of filter FS method that is based on the entropy or, to a certain extent, the uncertainty linked to a determined variable. The calculation it uses to determine the discriminative power of a feature is carried out with **Equation 6**.

$$IG(X,Y) = H(X) - H(X|Y) \quad (6)$$

Where H symbolizes entropy, $H(X)$ is the entropy of a particular variable (**X**), and $H(X|Y)$ is the entropy of the same variable after considering the class label (**Y**). These entropies are calculated through **Equations 7** and **8**.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (7)$$

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \quad (8)$$

In these equations, $P(x_i)$ denotes the marginal probability density function for a specific variable (**X**), and $P(x_i|y_j)$ is the conditional probability of a class (**Y**) given the analyzed variable (**X**). If the observed values of **X** are related to those of **Y** (class label or target variable), and $H(X|Y)<H(X)$, then the variable will be linked to the class label, or, in other words, will possess discriminative power according to this FS approach. The maximum value of IG(X,Y) is 1, and

high scores imply elevated ranks and discriminative potential (Dhir *et al.*, 2007; Novakovic, 2009; Zhao *et al.*, 2011).

Now that all of the FS methods employed have been mathematically described, it is time to begin explaining the algorithms used to design the models that utilize the features or independent variables located by these five data filters. These are multilayer perceptrons, the most commonly implemented kind of supervised artificial neural network.

2.4.2) Multilayer Perceptron

The FS algorithms utilized enable identifying the most relevant features or independent variables in a database that can fulfill a following modeling task. In this research, artificial neural networks, and, more specifically, feed-forward multilayer perceptrons (MLPs) have been selected, as they are powerful non-linear algorithms (Knoerzer *et al.*, 2011). MLPs are supervised mathematical tools which rely on non-linear interpolation to perform estimations, identifications, or classifications. This means that the operational window of a MLP is limited by the range of the values of the variables employed to train the model, meaning that if the model is tested with data outside this range, it will not be reliable as it is being forced to extrapolate (Torrecilla *et al.*, 2008-a).

These mathematical tools are reliable algorithms that have been employed in a wide variety of fields, such as chemistry, nanotechnology, food technology, biogeology, or biomedicine. In particular, some examples of where MLPs have been successfully employed in these five fields, respectively, are to accurately estimate the physicochemical properties of chemical compounds such as ionic liquids (Cancilla *et al.*, 2015), to sensitively detect xylene isomers at low ppm levels with ambipolar diketopyrrolopyrrole-based FET sensors (Wang *et al.*, 2016), to systematically identify different adulterations in olive oil samples (Aroca-Santos *et al.*, 2015), to predict the presence of moisture in microbially colonized halite rocks (Wierzchos *et al.*, 2015), or to classify patients to diagnose diseases like malaria (Webster *et al.*, 2009).

As mentioned in the introduction, MLPs possess a layered topology or architecture, with a set of units in each layer (nodes in the input layer and neurons in the hidden and output layers) (see **Figure 10** in section 1.5.2). The input layer is represented by the independent variables (one node per variable; in this case, the variables potentially selected by the FS algorithms) that the MLP uses to perform its calculations. On the other hand, the amount of neurons or calculation centers in the hidden layer must be optimized appropriately (*vide infra*), while the ones in the output layer are determined by the number of dependent variables defined (variables to be estimated) (Cancilla *et al.*, 2014-a). Once a specific topology has been established, the preliminary model can be trained.

2.4.2.1) Training a Multilayer Perceptron

Training a MLP is synonymous to optimizing the weights it contains. The variation in the amount of weights in a MLP solely depends on the quantity of units contained within it, which at the same time, is determined by the database modeled. The reasoning behind this, is that every unit (node or neuron) in a layer will be connected to all of the units in neighboring layers (not with units in the same layer), and each one of these connections is controlled by a weight, which

initially possesses a random value between zero and one (Cancilla *et al.*, 2014-a). Therefore, the amount of weights in a MLP, which changes with the topology, can be calculated through **Equation 9**.

$$\#W = IN * HN + HN * ON \quad (9)$$

In the equation above, *IN* is the number input nodes that the model has (independent variables), *HN* denotes the amount of hidden neurons, and *ON* represents the number of output neurons (dependent variables). It is important to highlight that there are additional weights, which are linked to biases (not included in **Equation 9**), which help shift the results offered by an activation function (*vide infra*) to become more accurate. There is usually one bias that enters all the neurons in the hidden layer, and a second one that enters the neurons in the output layer (Demuth *et al.*, 2005). During the training process, the weights are iteratively modified during a series of training cycles or epochs to reach their optimal values. The goal of these calculations is to attain weights that offer lower errors during an estimation or smaller misclassification rates during a classifying task (Knoerzer *et al.*, 2011).

During a training cycle, two successive calculations take place in every neuron (hidden and output) in order to attain a final result (estimated value). This result is evaluated by the MLP and compared with the existing real value, which is available as these algorithms are in fact supervised (Cancilla *et al.*, 2014-a). Then, through back-propagation, the weights are modified in order to lower the errors of the estimated results in the next training cycle (Kröse and van der Smagt, 1996). The mentioned calculations are carried out by two different mathematical functions. Initially, the activation function processes the data that enters a neuron through **Equation 10** (Cancilla *et al.*, 2015).

$$x_k = \sum_{j=1} w_{jk} y_j \quad (10)$$

Where w_{jk} represents the value of the weight that connects units j and k, y_j denotes the signal which enters the particular neuron (from an input node or a hidden neuron), and x_k is the solution of the activation function (Knoerzer *et al.*, 2011). As can be noticed, the activation function collects and processes all of the information that enters a single neuron from all the units in the previous layer.

The solution provided by the activation function (**Equation 10**) is used to perform the second calculation, which is done by the transfer function. This mathematical step is in charge of limiting the range of the values emitted by the activation function and perform the mathematical interpolations. Various options are available such as linear functions or other non-linear ones like hyperbolic tangent or sigmoid functions. In the present research, the selected transfer function was the sigmoid function (**Equation 11**), and its choice was purely based on the success attained in the recent past by our research group (Aroca-Santos *et al.*, 2015; Cancilla *et al.*, 2014-a; Cancilla *et al.*, 2014-b; Cancilla *et al.*, 2015).

$$y_k = \frac{1}{1 + e^{-x_k}} \quad (11)$$

In this equation, x_k and y_k are the answers of the activation and sigmoid (transfer) functions, respectively (Knoerzer *et al.*, 2011). As mentioned, the sigmoid function provides the

non-linearity to the MLP, as well as restricts the data values to a range that covers from 0 to 1. It can be seen in **Figure 17**.

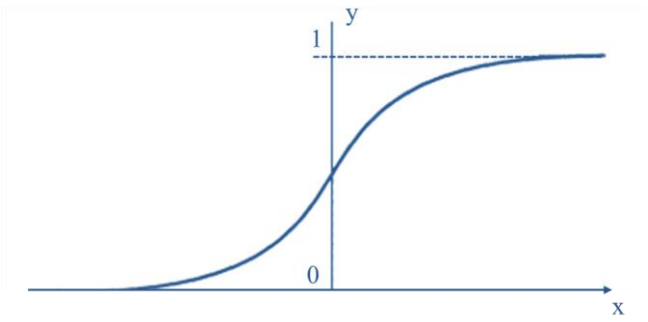

Figure 17. Mathematical representation of the sigmoid function (**Equation 11**).

The answer provided by the transfer function is the result of a neuron, which may be inputted into another neuron, or represent the final answer of the MLP model if it is originated by a neuron from the output layer (dependent variable).

It must be noted that a requirement that has to be met prior to begin the training or learning process is that the database must be prepared accordingly. In first place, it is a good habit to normalize all the variables (dependent and independent) in order to have the data limits coincide with those set by the transfer function and its interpolation boundaries, which in this case is between zero and one, as the sigmoid function has been employed (Lawrence, 1992).

Secondly, the database has to be randomly divided into at least two different datasets, which are the training and verification datasets (containing around 80-85% and 15-20% of the data points, respectively). This is key to achieve a properly trained model, because if every sample or data point is used to train it, an over-fit mathematical tool will be obtained, meaning that it will not be able to operate reliably with data that is external to the data used to train it, as it will have modeled trends which are intrinsic to that data such as experimental error or even noise (Torrecilla *et al.*, 2013). In other words, the verification dataset is employed by the MLP during the learning process, in every single training cycle, to verify that the error is not only lower for the data in the training dataset, which should happen as the weights are modified, but also for data external to it (verification dataset). In theory, without a verification dataset, the error rate can be lowered practically to zero, reaching an over-fit model (Cancilla *et al.*, 2015). The verification dataset denies this phenomenon, because when the error or misclassification rate provided by the MLP for samples from this external data increases for a determined number of training cycles in a row (six in this case), the learning process ends, and the weights, as well as the MLP, can be thought of as optimized. Avoiding being over-fit is equivalent to being able to generalize for data that is different from the samples used to train the model, which is the true goal of a MLP, and any other mathematical model (Cancilla *et al.*, 2014-a). In **Figure 18** a graphical representation of the typical development of training and verification errors with the training cycles is shown. In this case, the learning process would stop at training cycle 7, as it is the point where the error in the verification dataset begins to increase, and continues for six cycles in a row.

The process explained leads to a correctly trained, yet preliminary MLP model, as the algorithm is far from being optimized. There are several parameters, in addition to the weights, that must be properly selected or optimized to accomplish a fully operating and trustworthy MLP, and they will be described next.

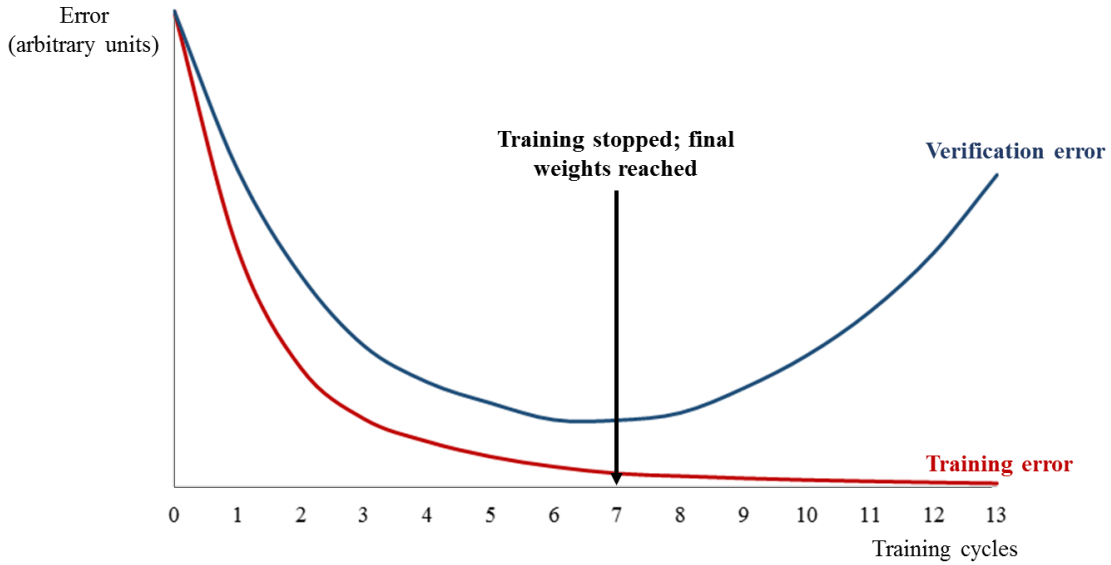

Figure 18. Chart that represents the typical development of the error concerning both datasets involved in the training process of a MLP: training dataset (red) and verification dataset (blue).

2.4.2.2) Optimizing a Multilayer Perceptron

Many parameters in a MLP have to either be selected or optimized so as to reach a proper, reliable, and accurate non-linear model. They will be looked into in the following subsections, which will describe all the required steps to achieve a useful and fully optimized MLP. These steps are optimizing the hidden neuron number, selecting an adequate training function, and optimizing a set of MLP parameters (learning coefficient and its modifying parameters).

2.4.2.2.1) Hidden Neuron Number

The hidden neuron number (HNN) represents the amount of neurons that are present in the hidden layer of a MLP (see **Figure 10**, section 1.5.2). Artificial neurons are the actual calculation centers of the algorithm, so optimizing their amount is crucial for the non-linear model (Gnana-Sheela and Deepa, 2013). A MLP that possesses a low HNN may have a hampered learning capability, and, therefore, may not be able to adequately interpret the existing relations between the variables, resulting in non-accurate models. On the other hand, if the HNN is excessive, this could lead to over-fit systems that are not able to generalize well for data that is external to the learning or verification datasets (Cancilla *et al.*, 2015).

There are various methods to optimize the HNN (Gnana-Sheela and Deepa, 2013), but in the present work a heuristic approach has been selected, testing HNNs within a logical range that would never lead to models containing less than double the amount of data points compared to the amount of weights, which can be calculated through **Equation 9** (Aroca-Santos *et al.*, 2015; Cancilla *et al.*, 2014-a). An elevated weight-to-sample ratio tends to originate over-fit models (Torrecilla *et al.*, 2013). It is worth noting that all of the MLPs designed only contained a single hidden layer in an attempt to lower the computational load (testing with more hidden layers was not necessary considering the amount of data points in the databases).

2.4.2.2.2) Training Function

The training algorithm is in charge of modifying the weights in order to reach more accurate estimations. In other words, it is in charge of optimizing the weights. There are many available training functions to carry out this process and, for this reason, it is essential to select an adequate one to fulfill a determined task. Specifically, 14 different training functions were considered, as each one possesses its own advantages and characteristics. They can be seen in **Table 5** (Torrecilla *et al.*, 2008-b).

Table 5. Different assessed training functions and their main traits (Torrecilla *et al.*, 2008-b).

Training Function	Subclasses	Brief Description
Gradient descendent BP*		Slow response. Employed during incremental-mode training
Gradient descendent with momentum BP	Gradient descent with variable learning rate	In general terms, it is faster than trainGD. Also used for incremental-mode training
Gradient descendent with momentum and adaptive linear BP		
Gradient descendent with adaptive learning rate BP		Faster than trainGD, but only operates in batch-mode training
Random-order incremental update	Resilient BP	Fast optimizing algorithms with very low storage needs. Operates in simple batch mode.
Resilient BP		
Fletcher-Powell conjugate gradient BP	Conjugated gradient descent	Lowest storage requirements of the conjugate gradient algorithms
Polak-Ribiere conjugate gradient BP		Larger storage needs that trainCGF. Faster optimization for particular cases
Powell-Beale conjugate gradient BP		Larger storage needs that trainCGP. Commonly faster optimization
Sealed conjugate gradient BP		No line search required. Reliable general-purpose algorithm
BFGS quasi-Newton BP	Quasi-Newton algorithm	Higher computation needs than gradient descent algorithms. Usually optimizes faster
One-step Secant BP		In between conjugate gradient algorithms and the quasi-Newton algorithm
Levenberq-Marquardt BP	Levenberq-Marquardt	Fastest for moderate sized databases
Bayesian regularization	Automated regularization	Tends to create models that generalize well

*BP stands for back-propagation.

Initially, it is difficult to predict the best training function for a specific problem, as it depends on many factors such as the amount of data points, weights, and biases, the intricacy of the problem, whether the MLP is intended for a classification (pattern recognition) or an estimation (function approximation), and so on (Mathworks website, 2016). For these reasons,

and based on our recent past experience, the training functions that have been primarily evaluated during this research were Levenberq-Marquardt back-propagation (trainLM) and Bayesian regularization (trainBR), as they have shown to be the most reliable and fast alternatives in most cases during our work (Aroca-Santos *et al.*, 2015; Cancilla *et al.*, 2014-a; Cancilla *et al.*, 2014-b; Cancilla *et al.*, 2015). TrainLM is the fastest training algorithm for moderate-sized MLPs, possessing a memory reduction feature for when the training dataset is large. On the other hand, trainBR is a modified version of trainLM that originates models that generalize well (not over-fit) and facilitates locating the optimal topology (Demuth *et al.*, 2005; Torrecilla *et al.*, 2008-b).

2.4.2.2.3) Multilayer Perceptron Parameters

Finally, to conclude the optimization of a MLP, certain parameters must be optimized as well. They are the Marquardt adjustment parameter (Lc), the decrease factor for Lc (Lcd), and the increase factor for Lc (Lci) (Demuth *et al.*, 2005). The Lc parameter is analogous to the learning coefficient in classic back-propagation algorithms (Palancar *et al.*, 1998), and its value is decreased and increased by Lcd and Lci, respectively, until the changes on Lc result in a worsened statistical performance for the MLP model. Lc is employed by the MLP during the back-propagation phase of the calculations that take place during each epoch, which results in the modification of the values of the weights (Demuth *et al.*, 2005). The calculation in which it participates is covered in **Equation 12**.

$$w(t+1) = w(t) + Lc \cdot MPE \cdot y_k(1 - y_k) \cdot y_{jk} \qquad (12)$$

In this equation, *w* stands for the weight value, *t* denotes the epoch, iteration, or training cycle, *Lc* is the aforementioned training coefficient, *MPE* is the mean prediction error between the real and estimated value (see **Equation 13**), y_k represents the result of the transfer function of a unit in layer *k* of the MLP (result of the direct calculation), and y_{jk} symbolizes the result of the transfer function of a unit from layer *j* that connects to the previous unit in layer *k* (for example, *j* can represent a neuron in the hidden layer, while *k* can stand for a neuron in the output layer).

$$MPE = \frac{1}{N}\sum_{k=1}^{N}\frac{|r_k - y_k|}{r_k} \times 100 \qquad (13)$$

Where r_k represents the real value, y_k is the estimated value provided by the model for the corresponding real value, and *N* is the amount of data points evaluated.

After understanding where Lc is involved, it is time to define how it is has been optimized, together with Lcd and Lci. In order to optimize these three parameters, a meticulous experimental design based on the "Box-Wilson Central Composite Design 2^3 + star points" was performed. The range of values tested went from 0.0005 to 1 for Lc and Lcd, and from 2 to 150 for Lci (Cancilla *et al.*, 2014-a; Torrecilla *et al.*, 2008-b). This thorough experimental design was carried out using the software Statgraphics Centurion XVI, while all of the other MLP-related calculations were achieved with Matlab version 7.0.1.24704 (R14) (Demuth *et al.*, 2005).

Now that all of the necessary steps to reach a completely optimized MLP have been covered, it is required to adequately validate these mathematical models, to be able to assure that

they can generalize well and be reliably applied to data that is external to the one used to train it. This will be looked into in the next and last subsection of the design of MLPs.

2.4.2.3) Validating a Multilayer Perceptron

Alike any type of arithmetical model, MLPs have to be statistically validated once they have been optimized. In the present thesis, every model that has been trained, has been validated using the following mathematical procedures: k-fold cross-validation (Cancilla *et al.*, 2014-a; Soleymani *et al*., 2011) and/or internal validation (Cancilla *et al.*, 2015; Cancilla *et al.*, 2014-b).

2.4.2.3.1) K-Fold Cross-Validation

During a k-fold cross-validation, the own verification dataset is employed to test the MLP. As this dataset is not involved in the weight modification process, it is a legitimate approach. It is based on the random division of the global database into *k* parts (or folds) containing the same amount of data points, and using *k*-1 segments as the training dataset, and the extra one as the verification dataset. This process is carried out *k* times, swapping the verification dataset for a new one in each new test. The final statistical performance of the model is evaluated by averaging the results from all *k* tests (Cancilla *et al.*, 2014-a; Soleymani *et al*., 2011), which is usually the MPE (**Equation 13**) for an estimator or the correct classification rate for a classifier. A graphical illustration of this mathematical validation method can be seen in **Figure 19**.

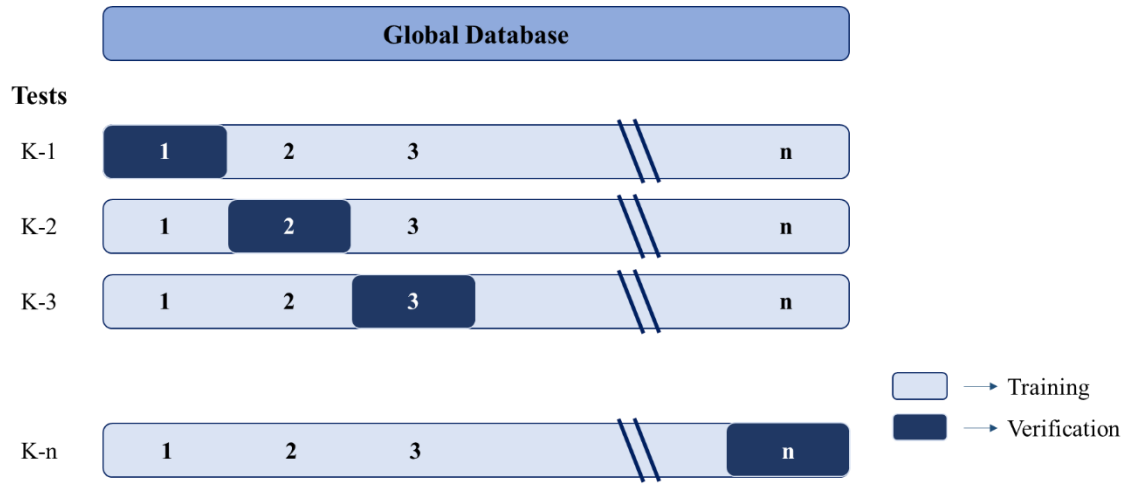

Figure 19. Representation of the k-fold cross-validation (k = n) method to verify the quality of mathematical models (each block represents the global database, with a unique training/verification separation).

All of the k-fold cross-validations that were performed throughout this research were k = 6, as it leads to an adequate data segmentation to properly train MLPs (83% training samples and 17% verification samples).

When a model provides accurate results during this validation method, it typically implies that the model can generalize well and, therefore, be reliable for data that is different from the one used in the training or verification datasets.

2.4.2.3.2) Internal Validation

In this second validating approach, the subdivision of the database is different than in the prior case. Now, the database is divided into three different kind of datasets, rather than two. The global database will be randomly divided into training, verification, and simulation datasets, each containing around 70%, 20%, and 10% of the data points, respectively. In this scenario, the learning phase of the model is carried out using the training and verification datasets, as explained above. Once the entire optimization process concludes, the model is tested with the simulation dataset that had been initially separated (Cancilla *et al.*, 2015; Cancilla *et al.*, 2014-b). This mathematical method can be seen depicted in **Figure 20**. Three individual tests can be observed, as it was the procedure followed during this research. The final score of the internal validation is the average of the three tests which possessed different data points in each of the three simulation datasets. Once again, the performance of each test is typically evaluated through the MPE (**Equation 13**) for an estimating MLP, or through the correct classification rate for a classifying one.

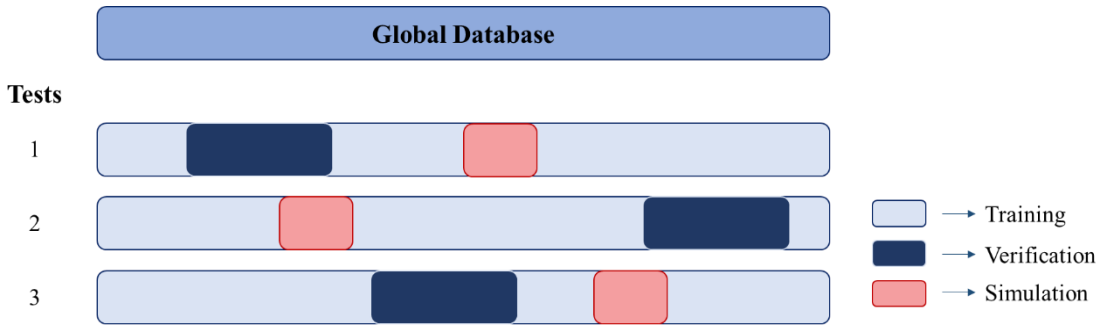

Figure 20. Representation of the internal validation approach, with three individual tests, to assess the statistical quality of mathematical models (each block represents the global database, with a unique training/verification/simulation separation).

In this case, if the statistical performance of the model is suitable, it is more than likely that the model can generalize well because it uses data that is completely separate from the training process (the simulation dataset is neither in training nor verification).

2.5) Summary

With the previous subsection, all of the analytical equipment and mathematical tools employed during this research have been presented and described. Nonetheless, before beginning with the results and discussion section, to get a fast and clear view of all the tools that have been utilized in this thesis, a graphical summary of the materials and methods can be seen in **Figure 21**.

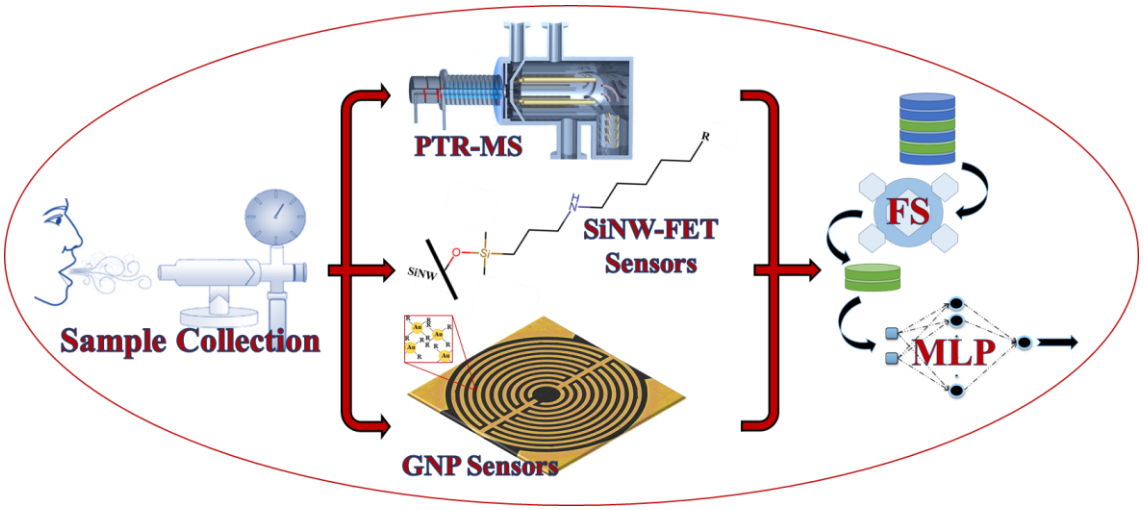

Figure 21. Graphical summary of the materials and methods that have been employed during this research and thesis.

As can be seen, the flow of every study carried out follows the same dynamic. First, the gaseous samples are collected, then, they are processed by one of the analytical approaches described (PTR-MS, SiNW FET sensors, or GNP sensors), and, finally, the data is treated using the mathematical tools covered (FS and/or MLP). Therefore, we are ready to begin analyzing the different experiments and results obtained.

3) Results and Discussion

Now that all the different tools that have been employed during this research, in the form of analytical equipment and mathematical algorithms and models, have been thoroughly described, it is time to present the results of the experiments that have been carried out. This section will be divided into a set of five main subsections. The first four will each cover a separate experimental section and its results. The first one involves the use of silicon nanowire field-effect transistor sensors and neural networks to set a proof-of-concept regarding the existing relation between the volatile organic compounds in a gaseous matrix and the signals provided by the sensors. The second through fourth experiments include the use of real breath samples to classify patients according to their disease. The second experiment employs silicon nanowire field-effect transistor sensors combined with neural networks (analogous to the first experiment), to classify patients with lung cancer, gastric cancer, asthma, or chronic obstructive pulmonary disease. The third one, describes the use of proton transfer reaction-mass spectrometry and intelligent modeling to identify lung cancer patients during an oral glucose tolerance test. The fourth and final experiment attempts to distinguish multiple patients with different diseases (chronic kidney disease, head and neck cancer, inflammatory bowel disease, multiple sclerosis, Parkinson's disease, preeclampsia, or pulmonary arterial hypertension) from groups of healthy controls using gold nanoparticle-based sensors and neural networks. On the other hand, the fifth and last subsection of the results and discussion will provide a global analysis of the results obtained and a comparison with previous studies. Therefore, we will begin with the description of the first experiment next.

3.1) Identifying and Quantifying Volatile Organic Compounds in Gaseous Mixtures through Silicon Nanowire Field-Effect Transistor Sensors and Neural Networks

In this first work of the research, the goal is to develop a method to selectively identify volatile organic compounds (VOCs) in a determined gaseous sample, as well as to estimate their concentration. To do so, specific samples of different VOCs were prepared, with known concentrations. The main goal of this experiment is to prove that the methodology employed leads to a successful end, which is the location and interpretation of the relationship there is between a specific VOC, as well as its amount, and the signals produced by silicon nanowire field-effect transistor (SiNW FET) sensors. The resulting data will be finally processed by neural networks, particularly multilayer perceptrons (MLPs), to reach VOC identifying and quantifying mathematical tools.

3.1.1) Obtaining the Data

In this first phase, the description of the gaseous samples of the VOCs prepared and processed, as well as the sensors and sensing feature extraction procedure will be meticulously presented.

3.1.1.1) Gaseous Samples

Eleven different VOCs were utilized to prepare every gaseous sample of this study, which were single- and multi-component mixtures. The molecules can be seen in **Figure 22**.

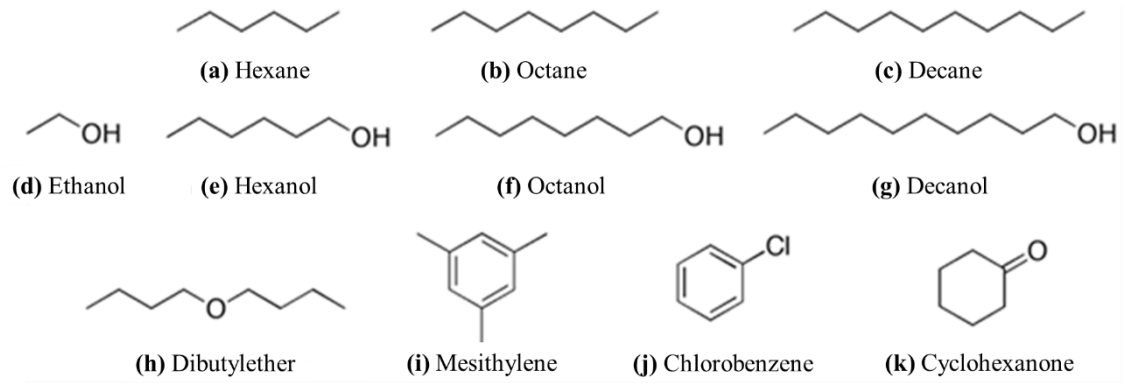

Figure 22. Eleven VOCs used to prepare the samples used in this experiment.

Of the compounds employed during this study, there are four non-polar VOCs (**Figure 22; a-c, h**) and seven polar ones (**Figure 22; d-g, i-k**). The non-polar compounds are alkanes (**Figure 22; a-c**) and an ether (**Figure 22; h**), while the polar ones are alcohols (**Figure 22; d-g**), benzenes (**Figure 22; i, j**), and a ketone (**Figure 22; k**). The molecules analyzed possess different chemical properties and structure, which allows evaluating the behavior of the SiNW FET sensors when interacting with diverse gaseous matrices.

The 11 VOCs analyzed are chemically similar to compounds that have been identified in breath samples of patients with different diseases, and, in other words, are comparable to biomarker candidates in breath (Buszewski *et al.*, 2007; Peng *et al.*, 2009). Therefore, if proven worthy, this methodology can be extrapolated to the identification and quantification of true volatile biomarkers as a proof-of-concept would be established.

The samples that have been employed are divided into two categories: single-component samples and multi-component mixtures. Every VOC (**Figure 22**) was used to prepare four single-component samples at different concentrations, which, in increasing order, were p_a/p_o = 0.01, 0.02, 0.04, and 0.08 (where p_a and p_o represent the partial pressure and vapor pressure of the VOC, respectively). On the other hand, hexane, octane, and hexanol were used to prepare all possible binary and ternary mixtures at a fixed concentration of each VOC (p_a/p_o = 0.08; three binary and one ternary mixture). The single-component samples were employed to determine the selectivity of the sensors towards the VOCs by using a MLP to identify and quantify them, whereas the multi-component samples were used to verify if the compounds that form binary and ternary mixtures also provide specific signals or patterns which can enable the identification of the elements of the samples (also through a MLP).

3.1.1.2) SiNW FET Sensors and Sensing Features

As mentioned in the materials and methods section, different SiNW FET sensors, each with unique molecular functionalizations (see **Figure 14** in section 2.3.1), were synthesized, and seven of them (**S1-S7**), which were produced through a two-step silane-acyl chloride modification process (*vide supra*), were introduced into a steel chamber to create a sensor array. In other words,

every sensor will interact with the gaseous sample at the same time in order to ensure that they are all under the same operating conditions. The names that are used to refer to each sensor are shown in **Table 6** (as well as **Figure 14**).

Table 6. Seven SiNW FET sensors in the chamber.

Name Given	Molecular Functionalization (-R in Figure 14)
S1	$-C_6H_5$
S2	-COOH
S3	$-COOCH_3$
S4	$-CH_3$
S5	$-CH_2CH_3$
S6	$-(CH_2)_4CH_3$
S7	$-(CH_2)_6CH_3$

All of the VOC samples were processed identically. Dry airflow was introduced for 30 minutes into the chamber, followed by another 30 minutes of VOC flow. The flowrates of air and VOC samples were 5 L per minute, and signals were registered during the entire time. Every one of the sensors leads to the same kind of signals from which a set of four sensing features are calculated and employed as independent variables in the following MLP models. The SiNW FET sensors interact with the VOCs through different non-covalent interactions which are dipole-dipole interactions, induced dipole-dipole interactions, and a tilt of the molecular layer that is originated due to the diffusion of the VOCs (Wang and Haick, 2013-a). The mentioned sensing features are extracted from source-drain current (I_{ds}) vs. gate voltage (V_g) curves which are a result of the sample exposure to the SiNW FET sensors (V_g ranged from 40 to -40 V and the source-drain voltage (V_{ds}) was 2 V). A typical example of these kind of curves can be seen in **Figure 23**, where all of the calculated features are shown as well (voltage threshold (V_{th}), I_{ds} at -40 V (I_{on}), subthreshold swing (SS), and charge carrier (hole) mobility (μ_h)).

The features are extracted as follows: (**a**) **V_{th}** is obtained by extrapolating the linear fit of the I_{ds} vs. V_g curve to where I_{ds} equals zero. (**b**) **I_{on}** is extracted by determining the source-drain current at -40 V (highest current). (**c**) **SS** is achieved by calculating the slope of the linear regime of the logarithmic I_{ds} vs. V_g curve. (**d**) **μ_h** is calculated using **Equation 14** and represents the velocity of the charge carriers in motion due to the influence of the electrical field (Wang and Haick, 2013-a).

$$\mu_h = \frac{ln[(2t_{ox} + R_{NW})/R_{NW}]}{2\pi\varepsilon_{ox}} \frac{L_{NW}}{V_{ds}} \frac{\delta I_{ds}}{\delta V_g} \quad (14)$$

Where t_{ox} is the thickness of the gate oxide (300 nm), ε_{ox} represents the relative dielectric permittivity of the oxide (obtained by multiplying the vacuum permittivity (ε_o) by the dielectric constant of the oxide (ε_r); $\varepsilon_o = 8.854e^{-12}$ F/m; $\varepsilon_r = 3.7$) , R_{NW} is the radius of the nanowire (40 nm),

L_{NW} symbolizes the length of the FET channel (2 µm), V_{ds} denotes source-drain voltage (2 V), and the expression $\delta I_{ds}/\delta V_g$ is the liner fitting slope of the linear region of the I_{ds} vs. V_g curve (Wang *et al.*, 2014).

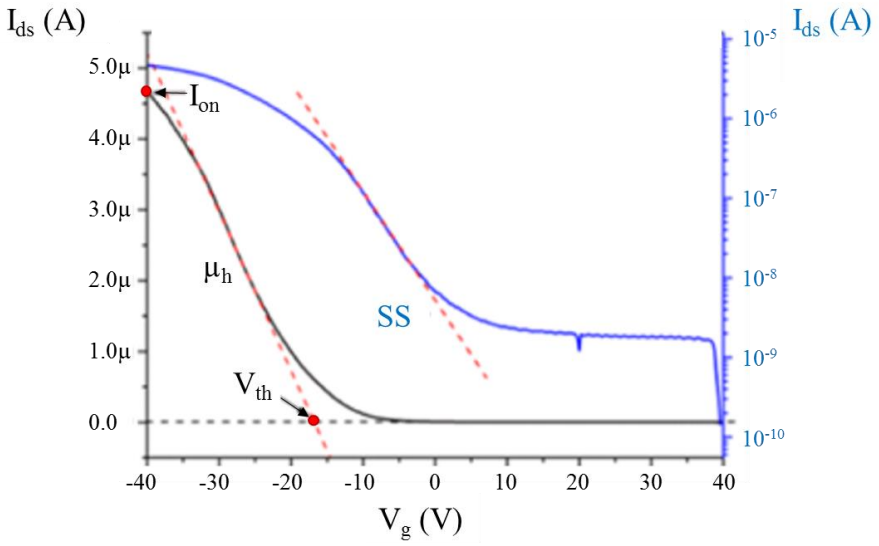

Figure 23. Typical example of an I_{ds} vs. V_g curve originated by a SiNW FET sensor. The black line is a linear scale, while the blue one is logarithmic. The four sensing features that have been extracted can also be seen.

The idea behind the extraction of these parameters is that every VOC at a specific concentration, should present a fingerprint-like combination of the four features, enabling VOC identification and quantification during the mathematical modeling phase. Three distinct types of MLPs will be designed and optimized using the data provided by the SiNW FET sensors, each with their own particular goal. Their design, optimization phases, and statistical performances will be presented next.

3.1.2) Mathematical Treatment

In this section, the results attained in the three different types of MLP-based models will be presented, as well as their repercussion in the field. They are intended to (**a**) identify the VOCs using data from single-component samples, (**b**) quantify the VOCs using the same samples, and (**c**) identify the VOCs from multi-component mixtures. It must be noted that models were independently optimized for data retrieved from single sensors, in an attempt to determine if it is possible to use them individually to fulfill the desired task. In other words, seven MLP models (one per sensor) were developed for every experiment, leading to a total of 21 mathematical tools.

3.1.2.1) Identification of VOCs in Single-Component Samples

In this first analysis, the goal was to use the data obtained from the SiNW FET sensors after processing the single-component samples to create models that are able to distinguish signals from particular VOCs, regardless of their concentration. The global database, from all seven sensors (see **Table 6**), contained a total of 709 data points which were obtained from samples of the VOCs at different concentrations (p_a/p_o = 0.01, 0.02, 0.04, and 0.08). Samples were randomly

measured twice or three times to possess a greater database and to verify that the results and methodology are repetitive (repeatability was confirmed). Therefore, each of the seven MLPs (one per sensor) was designed, trained, optimized, and validated using approximately 100 measurements of the 11 VOCs (see **Figure 22**).

The MLPs were in this case classifiers, as their purpose was to discriminate data points or instances that came from samples containing different VOCs. In order to design a classifying MLP, discrete outputs or dependent variables must be defined, and in this scenario, each of the 11 VOCs had a specific binary 1x10 vector assigned. These vectors can be seen in **Table 7**.

Table 7. Vectors assigned to each VOC for classifying intent. They will be used as dependent variables when training the MLPs.

VOC	Output Vector									
Hexane	1	0	0	0	0	0	0	0	0	0
Octane	0	1	0	0	0	0	0	0	0	0
Decane	0	0	1	0	0	0	0	0	0	0
Ethanol	0	0	0	1	0	0	0	0	0	0
Hexanol	0	0	0	0	1	0	0	0	0	0
Octanol	0	0	0	0	0	1	0	0	0	0
Decanol	0	0	0	0	0	0	1	0	0	0
Dibutylether	0	0	0	0	0	0	0	1	0	0
Mesitylene	0	0	0	0	0	0	0	0	1	0
Chlorobenzene	0	0	0	0	0	0	0	0	0	1
Cyclohexanone	0	0	0	0	0	0	0	0	0	0

In order to reach comparable results, the network parameters and topology of all seven MLPs were maintained constant (amount of inputs, hidden neurons, and outputs) after verifying that they were adequate. The parameters and functions employed can be found in **Table 8** and the architecture of the MLPs is depicted in **Figure 24**.

Table 8. MLP parameters and functions employed in the seven classifiers.

MLP Parameters	Selection or Value
Transfer function	Sigmoid
Training function	TrainBR
Lc	0.001
Lcd	0.1
Lci	10

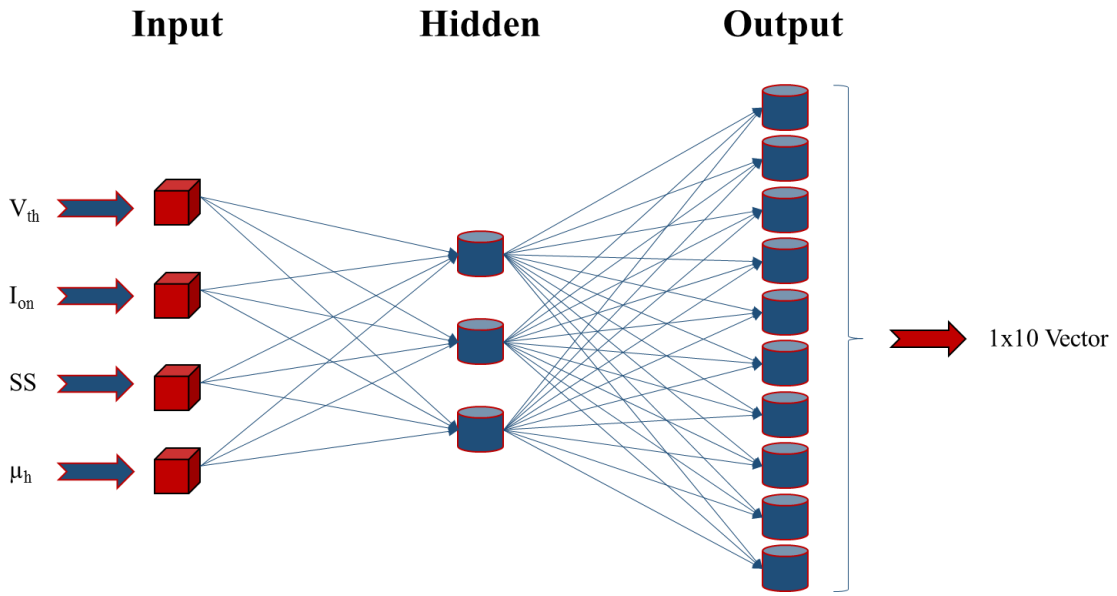

Figure 24. Topology of the MLPs used to identify the VOCs from single-component samples. The final architecture is 4-3-10 (input-hidden-output). The output 1x10 vector corresponds with the values shown in **Table 7**.

As can be seen, the independent variables employed correspond with the four sensing features extracted, which are employed to classify the VOCs. The accuracy or statistical performance of classifying models can be evaluated in various ways, and, in this case, the Euclidean distance (ED) has been employed (Palomar *et al.*, 2009). It is used to compare the target vectors (see **Table 7**) with those estimated by the MLPs, and, in the end, allows assessing the recognition power of each individual SiNW FET sensor. ED is calculated through **Equation 15**.

$$ED = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{15}$$

In this equation, x_i stands for a value in the target vector, y_i is the corresponding value of the estimated vector, and k is the amount of elements in the vector (10 in this case). Therefore, small EDs imply higher accuracies or correct classification rates. In **Table 9**, the VOCs that the MLPs are able to correctly identify are shown individually for each sensor. These results were obtained after evaluating the EDs of the trained MLPs during an internal validation (see **Figure 20**, section 2.4.2.3.2).

Table 9. Correctly and incorrectly classified VOCs by the MLP models corresponding to each SiNW FET sensor.

Sensor	VOCs Correctly Classified	VOCs Incorrectly Classified
S1	Hexane, decane, ethanol, hexanol, octanol, decanol, mesitylene, chlorobenzene, and cyclohexanone	Octane and dibutylether
S2	Hexane, octane, decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, chlorobenzene, and cyclohexanone	-
S3	Hexane, octane, decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, chlorobenzene, and cyclohexanone	-
S4	Hexane, octane, decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, chlorobenzene, and cyclohexanone	-
S5	Hexane, octane, decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, chlorobenzene, and cyclohexanone	-
S6	Decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, and chlorobenzene	Hexane, octane, and cyclohexanone
S7	Hexane, decane, ethanol, hexanol, octanol, decanol, dibutylether, mesitylene, chlorobenzene, and cyclohexanone	Octane

As can be deduced from the results, four of the seven sensors perfectly distinguish all single-component VOC samples (**S2** through **S5**), irrespective of concentration, while the other sensors at least correctly classify 8/11 VOCs (worst case for **S6**). This proves that the four sensing features extracted are able to characterize the VOC samples properly. In other words, the patterns calculated from the I_{ds} vs. V_g curves provided by SiNW FET sensors are VOC-specific and, therefore, the sensors are selective towards different volatile molecules.

Analyzing the specific performance of particular sensors, it is possible to compare the results offered by sensors with similar chain lengths, yet different functional groups (**S1** through **S4**) (Wang and Haick, 2013-a), and vice versa, sensors with common functional groups, but different chain lengths (**S4** through **S7**) (Wang and Haick, 2013-b). From the first group, only **S1**, which is functionalized with a phenyl group (see **Table 6**, section 3.1.1.2), fails to recognize some of the VOCs (octane and dibutylether). A possible reason for this may be the low adsorption power of non-cyclic compounds on phenyl groups (Wang and Haick, 2013-a). On the other hand, selectivity seems to decrease as sensors are functionalized with longer alkyl chains (**S4** and **S5** originate data that leads to more accurate models than **S6** and **S7**).

To sum up, single functionalized SiNW FET sensors and neural networks can be used to perfectly distinguish up to 11 chemically similar VOCs in single-component gaseous samples, regardless of their concentration (which was within $p_a/p_o = 0.01$ and $p_a/p_o = 0.08$). Now, let us see if these concentrations can also be determined using the same data provided by the sensors.

3.1.2.2) Quantification of VOCs in Single-Component Samples

In this phase of the experiment, using the data retrieved in the prior analysis (709 data points), the quantification of VOCs has been attempted. As mentioned above, the single-component samples possessed concentrations ranging from $p_a/p_o = 0.01$ to $p_a/p_o = 0.08$, and, therefore, it will be determined if the signals produced by the SiNW FET sensors are, apart from being component-dependent, also linked to VOC amount. Once again, seven independent MLPs have been developed, one per every sensor utilized, to assess the discriminative power of every individual sensing device.

The nature of the mathematical algorithms trained in this case is different than in the last study. Now they are estimators instead of classifiers and, for this reason, the dependent variables used, as well as the evaluation of their statistical performance, is different. Particularly, rather than a binary vector containing 10 outputs, there is only a single target or dependent variable which is employed to estimate the concentration of the samples.

Once again, with the intention to obtain comparable results, the architecture and the different network parameters were stabilized for all seven MLPs, after confirming they were suitable. These parameters are located in **Table 10**, and the topology of the estimators can be seen in **Figure 25**.

Table 10. MLP parameters and functions employed in the seven estimators.

MLP Parameters	Selection or Value
Transfer function	Sigmoid
Training function	TrainBR
Lc	0.01
Lcd	0.3
Lci	5

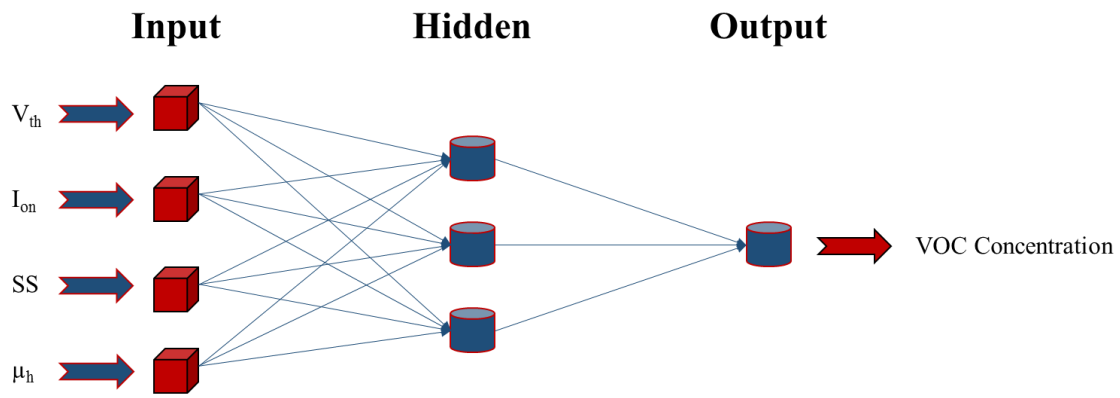

Figure 25. Topology of the MLPs used to quantify the VOCs from single-component samples. The final architecture is 4-3-1 (input-hidden-output).

In these tools, the statistical performance has been evaluated using the MPE (see **Equation 13**, section 2.4.2.2.3), which is a common calculation to assess the accuracy of estimating tools (Cancilla *et al.*, 2014-a). In **Table 11** the different MPEs obtained for the estimation of the concentration of the VOCs of all seven MLPs is shown, enabling the comparison of the accuracy of the different SiNW FET sensors. These are the results of an internal validation procedure (see **Figure 20**, section 2.4.2.3.2).

Table 11. MPEs of the concentration estimation for every VOC analyzed. The results cover all the MLPs designed, which correspond to each SiNW FET sensor employed in this experiment. The best estimation for each VOC is marked in bold.

VOCs	MPE (%)						
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Hexane	5.7	5.4	**<0.1**	5.3	4.8	4.4	3.7
Octane	4.7	1.1	**0.2**	0.9	4.9	1.3	8.7
Decane	2.8	2.2	1.3	2.9	3.9	1.9	**0.8**
Ethanol	3.9	0.4	**0.1**	2.9	4.2	1.7	0.3
Hexanol	2.1	2.4	2.4	0.1	0.3	1.1	**<0.1**
Octanol	2.7	1.5	2.4	1.9	2.3	2.2	**1.4**
Decanol	1.7	3.1	0.7	**0.1**	2.2	1.2	3.8
Dibutylether	3.2	5.7	2.7	1.5	**<0.1**	1.3	3.9
Mesitylene	5.1	3.6	**0.7**	9.6	4.1	3.1	9.7
Chlorobenzene	1.8	1.9	3.2	**0.5**	5.4	1.5	6.0
Cyclohexanone	2.9	1.9	2.3	0.5	**<0.1**	1.3	1.9

As can be observed, no quantification exceeds a 10% MPE, proving that this approach can be reliably employed to estimate the concentration of VOCs in single-component gaseous samples. As a matter of fact, every compound can be quantified with MPEs below 1.5% using one of the SiNW FET sensors employed, and in many cases with errors of approximately 0.1%. Specifically, **S3** (ester functionalization) offered the best results for the estimation of the concentration of non-polar VOCs (alkanes and mesitylene), while sensors that end in methyl groups (**S4** through **S7**) appear to be the best to quantify polar VOCs like alcohols.

Therefore, a very accurate system has been attained with the combination of these sensors and properly designed and trained MLPs, demonstrating a true relation between the sensing features extracted and the quantity of a VOC in a sample (within $p_a/p_o = 0.01$ and $p_a/p_o = 0.08$). Finally, the last part of this experiment is intended to identify VOCs in multi-component samples.

3.1.2.3) Identification of VOCs in Multi-Component Samples

The last part of this experiment was carried out using multi-component samples of three chemically similar VOCs (hexane, octane, and hexanol at $p_a/p_o = 0.08$) to determine if the followed approach is valid for the identification of compounds in mixtures. All three possible binary mixtures and the ternary mixture were prepared. Also, the data corresponding to the single-component samples of these VOCs were also employed to train this MLP model, leading to a total of 60 data points. Therefore, we had measurements of seven different kinds of samples, which were each encoded with an individual 1x3 binary vector that will become the dependent variables of the classifying model. The samples as well as their identifying vector can be seen in **Table 12**.

Table 12. Vectors assigned to each type of sample employed in the multi-component analysis. They will be utilized as dependent variables during the MLP optimization process.

Sample	Output Vector		
Hexane	1	0	0
Octane	0	1	0
Hexanol	0	0	1
Hexane-Octane	1	1	0
Hexane-Hexanol	1	0	1
Octane-Hexanol	0	1	1
Hexane-Octane-Hexanol	1	1	1

In this case, only data from **S2** was used to develop the model. The optimized network parameters as well as the functions employed are located in **Table 13**, whereas the architecture of the final MLP is represented in **Figure 26**.

Table 13. MLP parameters and functions employed in the multi-component analysis.

MLP Parameters	Selection or Value
Transfer function	Sigmoid
Training function	TrainBR
Lc	0.001
Lcd	0.1
Lci	10

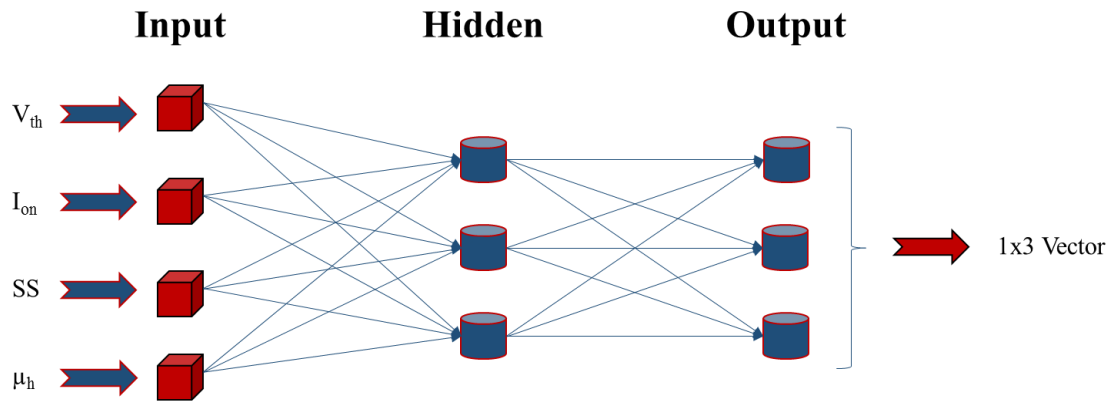

Figure 26. Architecture of the MLP utilized to classify the types of multi-component samples. The final architecture is 4-3-3 (input-hidden-output). The output 1x3 vector corresponds with the values shown in **Table 12**.

The same four sensing features are employed, and, once again, the ED (**Equation 15**) was used to evaluate the statistical performance of the classifier. For every single classification during an internal validation (see **Figure 20**, section 2.4.2.3.2), the ED was lower than 10^{-3}, which means that the MLP model operated with a correct classification rate 100%. It was able to distinguish the three individual VOCs (just as in the first part of this study) as well as locate those specific samples which correspond with particular binary and ternary mixtures. In other words, an approach based on combining a (single) functionalized SiNW FET sensor and intelligent algorithms such as ANNs is able to selectively identify VOCs in mixtures, which clearly opens the door to the analysis of more complex gaseous mixtures such as real breath samples.

To summarize the results of this experiment, using a set of seven molecularly functionalized SiNW FET sensors to process gaseous samples of 11 different VOCs, and treating the data with MLP models, leads to a reliable methodology that is able to accurately classify and quantify single-component samples and identify the compounds in multi-component ones. The consequence of this is that it is possible to establish a proof-of-concept that there is a real relationship between the composition of a gaseous sample and the signals provided by the cross-reactive sensors employed, as non-linear models have been able to very accurately carry out the desired tasks. All of this clearly favors future research in this line, because if it is possible to link

the signals that specific gaseous matrices originate with determined compounds and their amounts, it may be possible to relate particular patterns found in breaths of patients with determined diseases to that disease, and achieve non-invasive diagnosing tools. These three experiments have led to the publication of a scientific article in a prestigious journal which covers the mentioned results (Wang *et al.*, 2014).

In the second experiment of the current thesis, real breath samples will come into play, as they will be analyzed using a different set of functionalized SiNW FET sensors in an attempt to classify patients with different diseases as well as healthy controls.

3.2) Silicon Nanowire Field-Effect Transistor Sensors to Process Exhaled Breath Samples from Patients with Various Diseases to Classify them Via Neural Network Modeling

In this study, contrary to the previous one, breath samples obtained from humans will be gathered and analyzed with the aim set to distinguish those that come from different patients with either lung cancer (LC), gastric cancer (GC), chronic obstructive pulmonary disease (COPD), or asthma (AS) from others that are originated by healthy control subjects. This experiment is based on the fact that breath exhaled by patients possessing a specific disease is different in terms of volatile organic compound (VOC) composition to the breath produced by someone healthy or sick with a different disease (Bajtarevic, *et al.*, 2009; Dragonieri *et al.*, 2009; Peng *et al.*, 2009; Phillips *et al.*, 2007; Poli *et al.*, 2005).

Once the breath samples were gathered, they were processed using a set of differently functionalized silicon nanowire field-effect transistor (SiNW FET) sensors. Afterwards, the acquired databases were mathematically analyzed and prepared for the following two-phase calculations based on feature selection (FS) algorithms and multilayer perceptrons (MLPs) (see **Figure 16**, section 2.4). These MLPs were designed using data from individual SiNW FET sensors to classify the different diseases (LC, GC, COPD&AS, and healthy controls) through multiple binary classifiers. Once the models are fully optimized and validated, their statistical performance will allow determining the best sensor to carry out this task, as well as reaching useful tools, if accurate, that can aid in the breath-based non-invasive detection of these serious diseases.

3.2.1) Breath Samples and Population Study

The breath samples were all gathered as explained in section 2.1 from 374 volunteers that were able to sign a written informed consent. Relevant clinical data regarding the participants can be seen in **Table 14**.

Table 14. Clinical traits of the population study of every subgroup involved in this analysis.

Data	LC	GC	COPD&AS	Controls
Amount of participants	149	40	56	129
Gender (male/female)	86/63	28/12	35/21	51/78
Age ± SEM*	65 ± 11	60 ± 10	71 ± 11	65 ± 9
Smoking status (current or past/never)	30/119	15/25	15/41	22/107

*SEM stands for standard error of the mean (it is the standard deviation divided by the square root of the sample size).

It is worth noting that data regarding the staging of LC and GC cases was also attained (34 LC patients were staged as I or II, 110 were staged as III or IV, and 5 were unknown; 12 GC patients were staged as I or II, 24 were staged as III or IV, and 4 were unknown). The information from **Table 14** and concerning disease staging was gathered from self-report surveys done by the

volunteers as well as from hospital data. Next, the specific SiNW FET sensors employed will be described, as well as the extracted sensing features.

3.2.2) SiNW FET Sensors and Sensing Features

During this experiment, six different SiNW FET sensors were synthesized as described in section 2.3.1 to carry out the breath analysis, and were all located in a steel chamber to act as a sensor array. The employed sensors (**S5** and **S8**-**S12**), their synthesis process, and their surface functionalizations can be seen in **Table 15**.

Table 15. Six SiNW FET sensors in the chamber and their synthesis process.

Name Given	Synthesis Process	Molecularly Functionalized With
S5	Two-step silane-acyl chloride modification	3-Aminopropyl-triethoxysilane (APTES) + heptanoyl chloride
S8	Single-step silane modification	Trichloro(3,3,3-trifluoropropyl)silane
S9		Trichloro(phenethyl)silane
S10		(3-Bromopropyl)trichlorosilane
S11		APTES
S12	Two-step silane-monomer modification	Trichloro(3,3,3-trifluoropropyl)silane + anthracene

Every breath sample of this experiment was processed identically using this sensor array. The sensors were exposed to the breath samples for 15 min, after being under vacuum for also 15 min. During the sample exposure time, continuous measurements were gathered in the form of source-drain current (I_{ds}) vs. gate voltage (V_g) curves, where V_g ranged from 40 to -40 V and the source-drain voltage (V_{ds}) was 2 V (as shown in the **Figure 23**, in section 3.1.1.2 of the previous experiment). A set of sensing features were extracted from these curves, but in this case, instead of directly obtaining information from single or averaged I_{ds} vs. V_g curves (like single measurements of voltage threshold (V_{th}), I_{ds} at -40 V (I_{on}), subthreshold swing (SS), or hole mobility (μ_h)), specific I_{ds} values were calculated at different V_g to gather multiple sensing features per exposure. Each fixed V_g led to the calculation of four unique sensing features, which were the peak (maximum or minimum), the middle, the end, and the area under the curve of the resulting I_{ds} plateau. A hypothetical example of this feature extraction process can be seen represented in **Figure 27**.

Therefore, as explained, the different sensing features that have been extracted are based on a specific V_g value. A set of 19 voltages and/or related parameters have been used, which are 14 specific V_g (-39.8 V, -35 V, -30 V, -25 V, -20 V, -15 V, -10 V, -5 V, -0.2 V, 5 V, 10 V, 20 V, 30 V, and 39.8 V), V_{th}, μ_h, I_{on} (highest current), lowest current (I_{off}), and I_{on}/I_{off}. This leads to a total of 76 sensing features from a single sensor per sample analyzed, as four features are extracted from each of these 19 parameters (peak, middle, end, and area under curve).

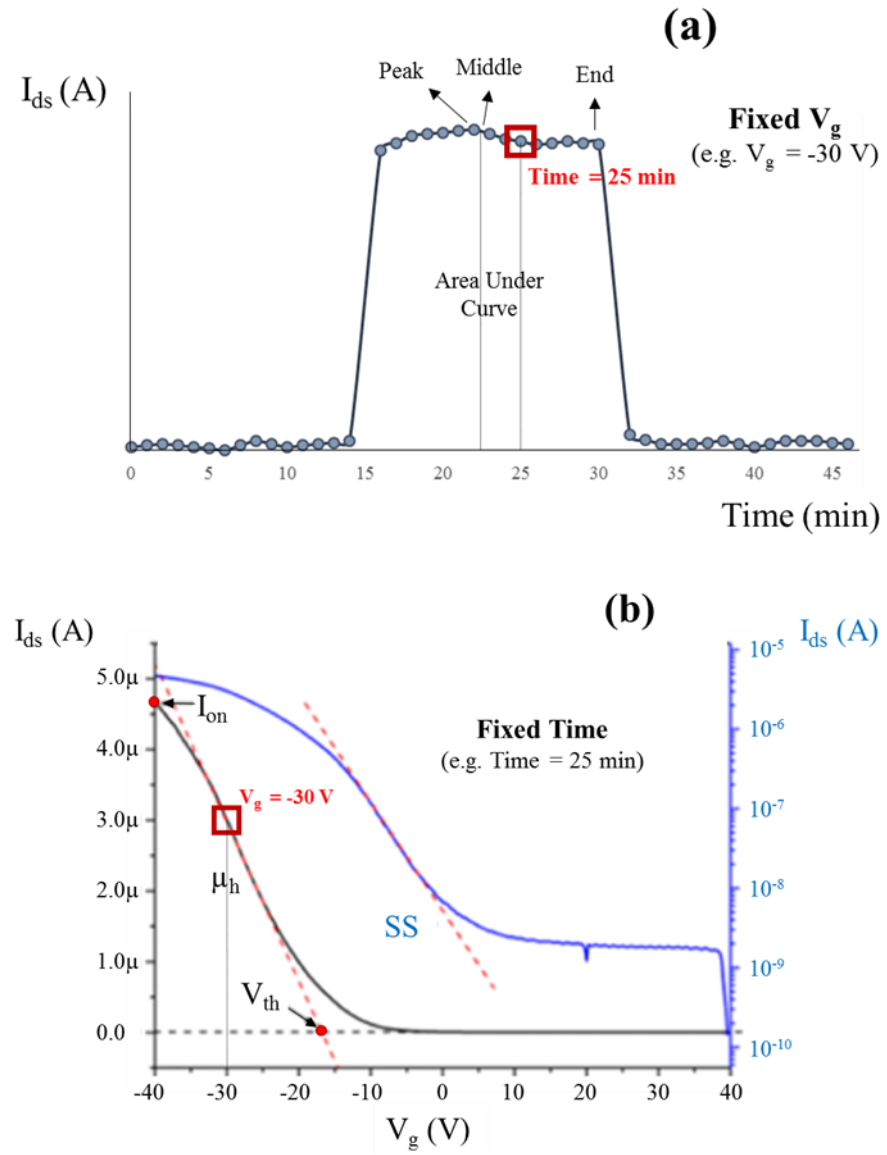

Figure 27. Charts such as these were employed to calculate the sensing features of this experiment (they represent a particular example). **(a)** Representation of different source-drain current (I_{ds}) values at fixed gate voltages (V_g). In this hypothetical example, V_g was set at -30 V. As can be seen, there is an I_{ds} plateau, and it is used to extract four features (peak, middle, end, and area under curve). **(b)** I_{ds} vs V_g graph from which the current values at specific times are gathered. In this case, it represents the measurement carried out at 25 min (during sample exposure). The red squares in both charts symbolize the same I_{ds}, or, in other words, the current that is extracted from **(b)** to represent **(a)**.

As a result of this process, a large database is created, where each of the volunteers' breath samples (374 samples) are characterized by 76 sensing features per sensor of the array. This leads to the next phase of the experiment, which is the mathematical analysis of the databases in order to reach disease classifying algorithmic tools and locating the best sensor from the array.

3.2.3) Mathematical Treatment

In order to achieve useful tools from the large databases produced, a proper mathematical analysis and treatment must be carried out. This procedure is mainly divided into two calculation phases, a filter-based FS to locate those sensing features with the greatest discriminative power, followed by a modeling phase using non-linear MLPs to classify the samples (as can be seen in **Figure 16**, section 2.4). This process will be carried individually for the databases originated by each of the six sensors, enabling the identification of the best sensor to distinguish samples from patients with different diseases (LC, GC, and COPD&AS) or healthy controls through a set of binary classifiers (MLPs).

3.2.3.1) Feature Selection

A total of six databases containing 76 sensing features were available, as six differently functionalized SiNW FET sensors were employed to process the breath samples. As there are four distinct groups (LC, GC, COPD&AS, and healthy controls) that are going to be classified through binary classifying MLP models, the possible combinations of two different groups are six. Therefore, there are 36 resulting databases that have each been run through the filter-based FS algorithm Relief-F (section 2.4.1.4). The samples within each of these databases were labeled with either a zero or a one, to differentiate the two groups being classified (these labels become the dependent variables of the samples).

The criteria set to determine the amount of selected sensing features was to reduce the database until there was a 1:10 independent variable/sample ratio. At least 10 samples per variable are desired to avoid over-fitting effects during this stage as well as the following modeling one (Torrecilla *et al.*, 2013). The amount of features selected during the FS calculations are presented in **Table 16**.

Table 16. Amount of variables selected by the Relief-F FS algorithm to distinguish breath samples from different diseases. The amount may differ within a binary classifier due to the elimination of samples that were catalogued as statistical outliers.

SiNW FET Sensor	Amount of Features Selected for each Binary Classifier					
	LC vs GC	LC vs COPD&AS	LC vs Control	GC vs COPD&AS	GC vs Control	COPD&AS vs Control
S5	18	20	27	9	16	18
S8	18	20	24	9	16	18
S9	18	20	27	9	16	18
S10	15	20	20	9	15	18
S11	18	20	27	9	16	18
S12	18	20	27	9	16	18

These selected sets of sensing features were each used to train individual MLP models during the second phase of the calculations.

3.2.3.2) Multilayer Perceptrons

A total of 36 MLPs were developed and trained, and their statistical performances (correct classification rate (%)) were compared to determine the most suitable sensor to classify the diseases through breath analysis. For every specific binary classifier, the MLPs had to be comparable, and, therefore, they were trained using the same parameters, training and transfer functions, and topology. This information can be seen in **Table 17**.

Table 17. MLP parameters, functions, and architecture of the different binary classifiers.

MLP Parameters and Topology	Selection or Value					
	LC vs GC	**LC vs COPD&AS**	**LC vs Control**	**GC vs COPD&AS**	**GC vs Control**	**COPD&AS vs Control**
Transfer function	Sigmoid					
Training function	TrainLM					
Lc	0.001					
Lcd	0.1					
Lci	10					
Inputs	15 or 18	20	20, 24, or 27	9	15 or 16	18
Hidden neurons	4					
Outputs	1					

It must be noted that the amount of hidden neurons was set so the maximum value possible that still originated at least a 2:1 ratio of samples/weights to avoid over-fit models (Cancilla *et al.*, 2015).

In order to compare the models and determine the best SiNW FET sensor to distinguish the diseases, a k-fold cross-validation test (k=6) (see **Figure 19**, section 2.4.2.3.1) was performed for each of the 36 MLPs to evaluate their generalization capability and accuracy. To assess the statistical performance of these binary classifiers, a threshold value was set to determine how well the models classify the samples. For example, a threshold of 0.5 would imply that all results provided by the MLP below that value would be seen as "0", while the ones above 0.5 would be considered "1". Initially, the thresholds were all set at 0.5, as it is the middle value of the possible dependent variables (zero or one). The results attained during the validations can be seen in **Table 18**.

Table 18. Statistical performance of the binary classifiers (MLPs) from each group combination and SiNW FET sensor used, in terms of correct hits (%), according to a k-fold cross-validation (k = 6). Best performances within a classifier marked in bold.

Sensor	LC vs GC	LC vs COPD&AS	LC vs Control	GC vs COPD&AS	GC vs Control	COPD&AS vs Control
	Statistical Performance*					
S5	**100/100/100**	**100/75/95**	91/75/84	**100/100/100**	**88/100/97**	80/55/65
S8	94/80/92	97/75/93	81/83/82	100/90/95	75/92/88	60/55/57
S9	94/100/95	88/88/88	91/83/88	**100/100/100**	88/88/88	**60/77/70**
S10	100/86/97	82/75/80	75/75/75	100/90/95	75/96/91	53/68/62
S11	91/100/92	**91/100/93**	**94/88/91**	**100/100/100**	75/96/91	**67/73/70**
S12	100/60/95	91/88/90	81/83/82	**100/100/100**	63/92/85	60/67/64

*The results are given in terms of correct % $group_1$/correct % $group_2$/correct % total.

As can be deduced, the results confirm the usefulness of this approach. Many of the designed tools operate fairly accurately (many breaching the 90% mark), except for the COPD&AS vs control classifier, which barely reaches 70% in the best cases (this fact can be explained as there is a wider heterogeneity within this population, as two diseases have been combined into a single group and the control samples are inherently heterogeneous). From the results, it is possible to say that any of the six SiNW FET sensors synthesized combined with neural network modeling can lead to accurate and non-invasive detection of severe diseases such as LC and GC through breath analysis.

If a smaller set of sensors had to be selected to design a medical device for breath analysis, probably **S5** and **S11** would be chosen (molecularly functionalized with APTES + heptanoyl chloride and with APTES, respectively; see **Table 15**), as they have shown the best statistical performances in general. For this reason, the MLP models corresponding to these two sensors have been fully optimized (as described in section 2.4.2.2). The resulting hidden neuron number, learning coefficients, and final validation results (k-fold cross-validation (see **Figure 19**, section 2.4.2.3.1) and internal validation (see **Figure 20**, section 2.4.2.3.2)) with their optimal thresholds, which were calculated to achieve the best possible performances for both groups classified (best possible multiplied percentage; correlated with results from receiver operating characteristic curves (Kumar and Indrayan, 2011)), are gathered in **Tables 19** and **20**, one for each sensor.

The results of these optimized MLP models reveal accurate tools to distinguish the defined groups using this approach. A single functionalized SiNW FET sensor (**S5** or **S11**) that is used to analyze breath samples is capable of producing data that is clearly representative of LC, GC, or COPD&AS patients, as the models are able to accurately classify them. The statistical performance offered by both validation procedures ensure as well the generalization capability of the system, especially from the results of the internal validations, which employ data that is completely unrelated to the training process. It is worth highlighting that the internal validations

for all six binary classifiers designed with the data provided by **S11** showed accuracies above 86%.

Table 19. Optimized MLP parameters and architecture of the different binary classifiers for **S5**, as well as the statistical performance of a k-fold cross-validation and an internal validation.

MLP Parameters and Topology	**Optimized Value/Statistical Performance**					
	LC vs GC	**LC vs COPD&AS**	**LC vs Control**	**GC vs COPD&AS**	**GC vs Control**	**COPD&AS vs Control**
Lc	0.001	0.001	0.001	0.001	0.001	0.001
Lcd	0.1	0.1	0.001	0.1	0.1	1
Lci	10	10	2	10	10	100
Inputs	18	20	27	9	16	18
Hidden neurons	4	4	3	4	4	4
Outputs	1					
K-fold cross-validation performance (correct % $group_1$/correct % $group_2$/**correct % total**)	100 100 **100**	100 75.0 **95.0**	87.2 83.8 **85.6**	100 100 **100**	88.0 100 **97.0**	64.3 76.0 **72.4**
K-fold cross-validation threshold	0.50	0.50	0.49	0.50	0.50	0.52
Internal validation performance (correct % $group_1$/correct % $group_2$/**correct % total**)	97.7 92.8 **96.5**	95.3 60.0 **84.1**	95.3 73.2 **84.5**	100 95.2 **96.7**	75.0 100 **94.1**	63.6 89.1 **84.2**
Internal validation threshold	0.40	0.46	0.51	0.66	0.45	0.50

To sum up, in this second experiment, a relation between the compounds in the breath of sick and healthy people and their clinical status has been successfully found, interpreted, and taken advantage of to create mathematical tools that are able to distinguish and detect different relevant diseases like LC, GC, COPD, and AS. The SiNW FET sensor and intelligent modeling combination has proven to be a worthy approach to non-invasively evaluate the clinical status of patients, opening the door to the implementation of such devices during disease screening procedures. These results have been submitted to a relevant journal in the nanotechnology field for publication (Shehada *et al.*, 2016).

The following experiment of this thesis will presented during the next subsections. It involves the use of proton transfer reaction-mass spectrometry to analyze the breath of different LC patients as well as a high-risk population for LC, during an oral glucose tolerance test, to try and design LC-detecting MLP models and locate potential volatile LC biomarkers.

Table 20. Optimized MLP parameters and architecture of the different binary classifiers for **S11**, as well as the statistical performance of a k-fold cross-validation and an internal validation.

MLP Parameters and Topology	**Optimized Value/Statistical Performance**					
	LC vs GC	**LC vs COPD&AS**	**LC vs Control**	**GC vs COPD&AS**	**GC vs Control**	**COPD&AS vs Control**
Lc	1	0.001	0.001	0.001	0.001	0.001
Lcd	0.001	0.1	0.1	0.1	0.001	0.001
Lci	2	10	10	10	2	2
Inputs	18	20	27	9	16	18
Hidden neurons	4	4	4	4	3	4
Outputs	1					
K-fold cross-validation performance (correct % group$_1$/correct % group$_2$/**correct % total**)	98.6 97.5 **98.4**	91.0 100 **93.0**	94.0 88.0 **91.0**	100 100 **100**	80.0 98.4 **94.1**	69.6 83.7 **79.4**
K-fold cross-validation threshold	0.69	0.50	0.50	0.50	0.58	0.39
Internal validation performance (correct % group$_1$/correct % group$_2$/**correct % total**)	97.9 88.9 **96.5**	87.2 93.8 **88.9**	92.5 81.8 **86.9**	100 94.4 **96.7**	85.7 100 **96.1**	78.6 90.7 **87.7**
Internal validation threshold	0.32	0.43	0.40	0.61	0.56	0.38

3.3) Detecting Lung Cancer during an Oral Glucose Tolerance Test through Breath Analysis Using Proton Transfer Reaction-Mass Spectrometry and Intelligent Modeling

In this experiment, breath samples from lung cancer (LC) patients were obtained, studied, and compared to others from controls that had been identified as high-risk individuals for LC. Additionally, in the present study, the tests were carried out during the course of an oral glucose tolerance (OGT) test to assess the role that glucose metabolism plays on the volatile compounds found in breath.

The breath samples were analyzed using proton transfer reaction-mass spectrometry (PTR-MS) and, therefore, quantitative information regarding specific volatile organic compounds (VOCs) was attained. Afterwards, the data was treated using a two-step calculation procedure based on feature selection (FS) and multilayer perceptrons (MLPs), as explained in the materials and methods section (see **Figure 16**, section 2.4), after necessary preliminary mathematical calculations were done. In the next subsections, population characteristics, OGT test, sample gathering, PTR-MS analysis, and mathematical treatment will be described and the results given and discussed.

3.3.1) Population Traits

Initially, 48 Israeli participants were included in the study, but due to technical difficulties, eight of the samples they produced were excluded from the analysis. From the remaining 40 people, 18 were LC patients and 22 were controls which were identified as high-risk for LC and were being treated in pulmonology clinics at the time of the breath sampling. Age, gender, and medical and smoking history of the sick and control groups were comparable, and this information can be seen in **Table 21**.

Table 21. Relevant information from the population study.

Data	LC Patients	Controls
Amount of participants	18	22
Gender (male/female)	10/8	15/7
Age ± SEM*	63 ± 15	61 ± 15
Body mass index ± SEM*	25 ± 5	25 ± 5
Participants with allergies (pollen or dust/other)	2/8	4/4
Participants with exposure to asbestos**	2	1
Smoking status (current/past/never)	5/6/7	5/7/10

*SEM stands for standard error of the mean.

**Exposure to asbestos, which are a set of naturally appearing silicate minerals, has been linked to the development of certain LC cases (Nicholson *et al.*, 1982).

The data from **Table 21** was obtained from self-report surveys (filled in by the participants) and from different databases at hospitals. Regarding the histology of the LC cases, there were two small cell and 16 non-small cell cases. These last 16 were divided into 12 adenocarcinomas, two squamous cell cancers, and two others. In addition, it is worth noting that 17 out of the 18 cases were in stages three or four (no reliable information regarding staging for the last case).

Finally, there was a set of exclusion criteria that was followed to reach the final population study: (**a**) patients that were uninterested in participating or were not able to sign the according consent form, (**b**) LC patients who had begun treatment prior to the research, (**c**) those who were unable to complete the needed steps during the research and/or the follow up visits, and (**d**) individuals (patients or controls) who suffered from diabetes.

3.3.2) Oral Glucose Tolerance and Breath Tests

All the participants (patients and controls) were asked to drink a 273 mL solution containing 75 grams of glucose and water, after having fasted for six hours, and glucose levels in blood were determined before and after the OGT test (average glucose levels and standard deviation for LC patients: 101.2 ± 25.1 mg/dL pre-OGT test and 166.1 ± 58.3 mg/dL post-OGT test; average glucose levels and standard deviation for controls: 94.1 ± 12.1 mg/dL pre-OGT test and 155.2 ± 46.3 mg/dL post-OGT test). Two exhaled breath samples were gathered per individual, one before the test, and another 90 minutes (lay period) after the OGT test. Therefore, two breath samples were obtained per participant, and they will be referred to as pre-glucose uptake and post-glucose uptake samples, leading to a total of 80 breath samples. These samples were obtained as explained in the materials and methods section 2.1, and directly loaded onto the PTR-MS system within one to three hours of their collection.

3.3.3) PTR-MS Analysis

As mentioned before in section 2.2, PTR-MS is a very sensitive system which is able to identify and quantify VOCs, without requiring any preconcentrating procedures, sample preparation, or chromatography (Blake *et al.*, 2009; Ligor *et al.*, 2009). In this study, bar scan mode was used to achieve clean information about specific VOCs, which are characterized by mass/charge ratios (m/z). The concentration of the volatile molecules was measured 20 times (20 cycles) for each sample, to ensure that the results and methodology were statistically repetitive and robust (it was confirmed). The m/z measured were 21 and every option (every natural number) between 32 and 180, both included.

3.3.4) Mathematical Treatment

In this section, the procedures followed to reach the final mathematical tools are presented. First of all, the initial data analysis and preparation was carried out. Once the databases were ready, two distinct analyses, with different goals, were performed using FS and MLPs.

3.3.4.1) Preliminary Analysis and Database Preparation

In this phase, the results provided by the PTR-MS system were organized and made comparable to be able to proceed with the further mathematical analysis. The steps carried out were as follows:

1. **Locate and eliminate statistical outliers** from the measurements (as described at the beginning of section 2.4).
2. **Calculate the means of every m/z** for every sample (without outliers) to possess representative data.
3. **Discard m/z that possess relatively high values** such as m/z 21, which corresponds to $H_3{}^{18}O^+$, and has around 10^7-10^8 times greater values than most of the remaining m/z. Nine out of the initial 150 m/z were consequently excluded.
4. **Normalize the data per participant** between zero and one in order to make all the samples comparable.

After completing these four steps, the database is prepared to be treated in order to reach the desired mathematical models. In **Figure 28** two examples of the resulting PTR-MS data can be seen after undergoing these four preliminary phases.

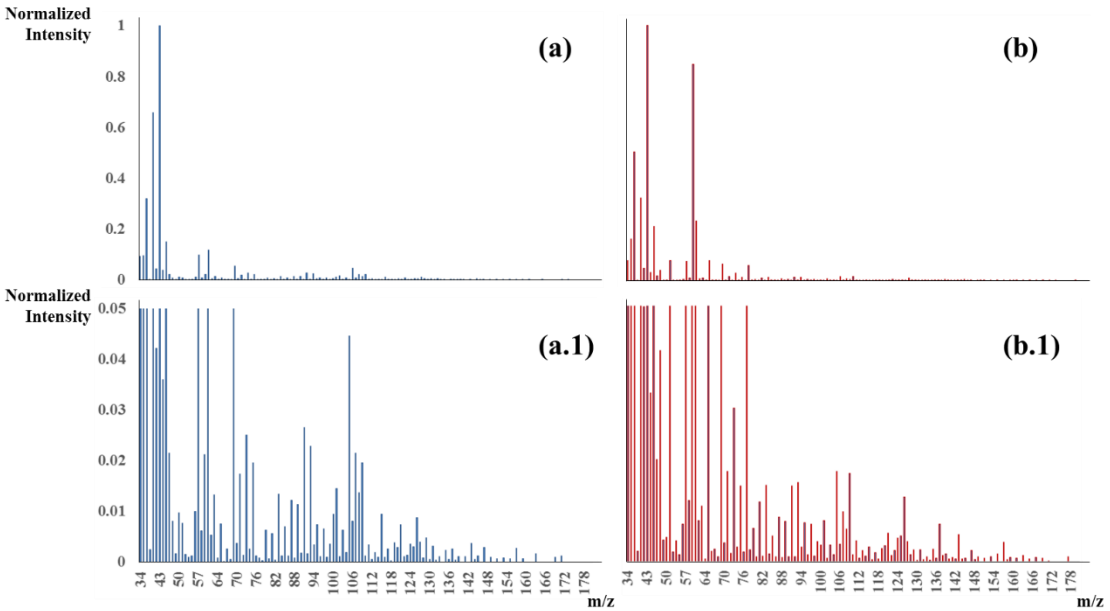

Figure 28. Graphical representation of two samples from the PTR-MS study after the preliminary analysis. **(a)** Example corresponding to a control individual (pre-glucose uptake), and **(a.1)** is its amplified version (amplified ordinate axis; allows a better evaluation of the intensity values of VOCs presenting low concentrations); **(b)** example belonging to a LC patient (pre-glucose uptake), and **(b.1**) is its amplified version (amplified ordinate axis).

In these representations, thanks to bar scan mode employed during the measurements and the preliminary analysis and calculations performed, single intensity values are linked to specific m/z or VOCs in the breath samples. These m/z will become the independent variables of the models that will be presented next.

Two main algorithmic tools have been developed in this research, each one possessing its own purpose. The first one was designed using all 80 samples individually, regardless of the

glucose uptake, and only labeling them with their medical status (LC vs. control). Therefore, every participant originated two breath samples which were used to create a classifier to distinguish those samples that come from LC patients or controls, in order to reach a diagnosing system. This study will be looked into next.

3.3.4.2) Distinguishing LC Patients from Controls Regardless of Glucose Uptake

The database for this study contained 80 data points, two per participant (pre- and post-glucose uptake; 44 from controls and 36 from LC patients), with a total of 141 m/z which are the independent variables of the system. These 141 m/z represent different volatile molecules present in the breath samples, and hopefully some of them will enable the discrimination of controls from LC patients. If accurate models are achieved, it may reveal potential m/z with discriminative or diagnostic power, indicating that they might be possible volatile biomarker candidates for LC diagnosis. On the other hand, every data point was labeled according to the participant's status. Every control was assigned a zero, while every LC patient was given a one. These labels are the dependent variable of the analysis, as they will be employed to classify the samples.

The first step of this analysis involves the use of the five filter FS algorithms presented previously (section 2.4.1). They were used to determine the m/z (independent variables) with the greatest discriminative power to separate breath samples from controls and LC patients. In **Table 22** the top eight variables selected by each FS method can be found (eight was selected to maintain a 1:10 ratio of independent variables/samples during the subsequent MLP modeling, avoiding potential over-fitting effects (Torrecilla *et al.*, 2013)).

Table 22. Variables (m/z) selected by the FS algorithms to distinguish breath samples from both populations (repeated m/z marked in bold).

FS Algorithm	m/z Selected
χ^2	152, 160, 162, 166, 168, 175, 176, and 178
Fisher	60, 61, **62**, 119, **125**, **126**, **147**, and **148**
Kruskal-Wallis	**62**, 78, **125**, **126**, 140, **147**, **148**, and 161
Relief-F	41, 42, 43, 69, 70, 120, **126**, and 173
Information gain	50, **62**, 107, 109, **125**, **126**, **148**, and 164

As can be seen, various m/z are selected in more than one test (the selected m/z (volatile compounds) will be covered afterwards), even though each algorithm employs its own mathematical criteria to select the features. The variables selected by these algorithms were directly employed as inputs in five different yet comparable MLP models. Therefore, the statistical results provided by these non-linear binary classifiers will indicate the best m/z combination to distinguish breath samples from controls and LC patients, and may reveal potential volatile biomarker candidates for LC diagnosis (irrespective of glucose consumption).

So as to reach analogous results, the set of network parameters as well as the architecture of the five MLPs were stabilized once determined that they were suitable. These parameters and mathematical functions can be found in **Table 23** and the topology of the MLPs is represented in **Figure 29**.

Table 23. MLP parameters and functions employed in the five classifying models (glucose consumption not considered).

MLP Parameters	Selection or Value
Transfer function	Sigmoid
Training function	TrainLM
Lc	0.001
Lcd	0.1
Lci	10

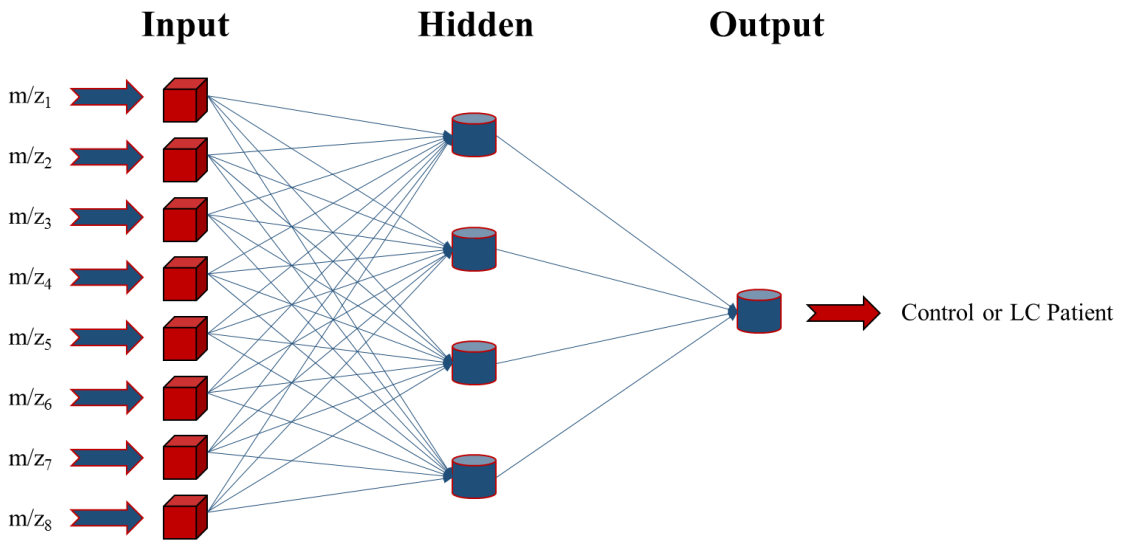

Figure 29. Topology of the MLPs used to distinguish breath samples from controls and LC patients, regardless of glucose consumption. The final architecture is 8-4-1 (input-hidden-output).

To evaluate the statistical performance of the classifiers, a threshold value was set to be able to compare the estimated values with the real labels. These thresholds were optimized according to the results obtained during the validation procedures, to achieve the best possible performances for both groups classified (best possible multiplied percentage). In this study, for every MLP, a k-fold cross-validation (k = 6) (see **Figure 19**, section 2.4.2.3.1) and an internal validation (see **Figure 20**, section 2.4.2.3.2) were carried out. In all of the validation tests, the samples were divided randomly into the required training, verification, and simulation datasets, but equivalently for every different MLP. This implies that the results from the different models

are comparable, as the same samples are evaluated during the same kind of validations. The final results are shown in **Table 24** in terms of accuracy (correct hits (%)).

Table 24. Statistical performance of all five MLPs, in terms of correct hits (%), according to two validation procedures (k-fold cross-validation and internal validation). Best results are marked in bold.

MLP	**k-Fold Cross-Validation** (n = 80)				**Internal Validation** (n = 24)			
	Th*	**Control** (n = 44)	**LC** (n = 36)	**Total**	**Th**	**Control** (n = 14)	**LC** (n = 10)	**Total**
χ^2	0.50	70.4%	36.1%	55.0%	0.50	78.6%	50.0%	66.7%
Fisher	0.50	84.1%	83.3%	83.8%	0.50	92.8%	80.0%	87.5%
Kr.-Wa.	0.45	88.6%	86.1%	87.5%	0.50	78.6%	80.0%	79.2%
Relief-F	0.50	**90.9%**	**91.7%**	**91.2%**	0.60	92.8%	**90.0%**	**91.7%**
Info gain	0.55	**90.9%**	77.8%	85.0%	0.52	**100%**	80.0%	**91.7%**

*__Th__ symbolizes threshold.

As can be seen, except for the MLP model that uses the m/z selected by χ^2 FS method, all of the other variable sets lead to considerably accurate mathematical tools. As a matter of fact, the variables selected by χ^2 possess a great amount of zeroes (around 90% of the values), due to the relatively simple type of calculations it performs. All of the VOCs it selects possess elevated m/z ratios (on the right of the charts in **Figure 28**), which are the least abundant molecules. For this reason, the models are not accurate when only employing these kind of m/z to discriminate the two kinds of samples. On the other hand, when looking at **Table 14**, it can be seen that several of the m/z provided by Fisher's discriminant ratio, Kruskal-Wallis test and information gain test overlap (up to five m/z). This may be the reason why the statistical performances of these models are very similar. Finally, the MLP which employs the m/z selected by relief-F appears to be the most accurate considering both of the validation tests carried out (only model with an accuracy over 91% for both tests).

As the results from the model that uses the m/z selected by relief-F FS method are the most accurate, the parameters and topology of this model have been fully optimized (as described in section 2.4.2.2). The results of this process can be seen in **Table 25**, as well as the statistical results of both a k-fold cross-validation (k = 6) and an internal validation. The final results of the validations have been graphically represented in **Figure 30**.

Therefore, it can be concluded that it is possible to accurately distinguish breath samples originated from LC patients and high-risk controls, regardless of the individuals' glucose consumption, by combining PTR-MS analysis and intelligent mathematical modeling. This could signify that potential interfering signals originated by glucose metabolism may be avoided using this approach, facilitating the sample gathering procedure and perhaps evading certain confounding factors. The accurate results provided by two independent validation tests (93.8% and 95.8% for the k-fold cross-validation and internal validation, respectively) show that the relationship exists between the m/z selected and the clinical status of the patient. As well, these

tests ensure the generalization capability and applicability of the models, as they avoid over-fit systems.

Table 25. Optimized MLP parameters and architecture of the binary classifier that distinguishes LC patients from healthy controls regardless of their glucose uptake, as well as the statistical performance of a k-fold cross-validation and an internal validation.

MLP Parameters and Topology	Optimized Value or Statistical Performance
Lc	1
Lcd	0.001
Lci	2
Inputs	8 (m/z 41, 42, 43, 69, 70, 120, 126, and 173)
Hidden neurons	3
Outputs	1
K-fold cross-validation performance (correct % controls/correct % LC patients/**correct % total**)	93.2/94.4/**93.8**
K-fold cross-validation threshold	0.50
Internal validation performance (correct % controls/correct % LC patients/**correct % total**)	92.3/100/**95.8**
Internal validation threshold	0.50

It has been shown that only utilizing sets of eight m/z as independent variables, it is possible to create a MLP-based model that offers an accuracy above 93% for an 80 sample study. Finally, the VOCs that originate these m/z with discriminatory power (m/z from the models that produced accurate results; 41, 42, 43, 50, 60, 61, 62, 69, 70, 78, 107, 109, 119, 120, 125, 126, 140, 147, 148, 161, 164, and 173) may potentially represent volatile biomarkers to help detect and diagnose LC in a non-invasive and safe manner, as their concentrations may differ when comparing breath samples from LC patients and others from controls. Furthermore, the identity of some these VOCs has been proposed according to the mass of the compounds. They are m/z 61, 107, 147, and 148, which possess masses which coincide with acetic acid, ethylbenzene, 1,2-dichlorobenzene, and glutamic acid, respectively. These molecules, which are present in the body, may represent volatile biomarkers in breath that can aid in LC diagnosis. Nevertheless, these results embody an initial phase that should be backed up by further research to potentially validate the compounds as true LC volatile biomarkers.

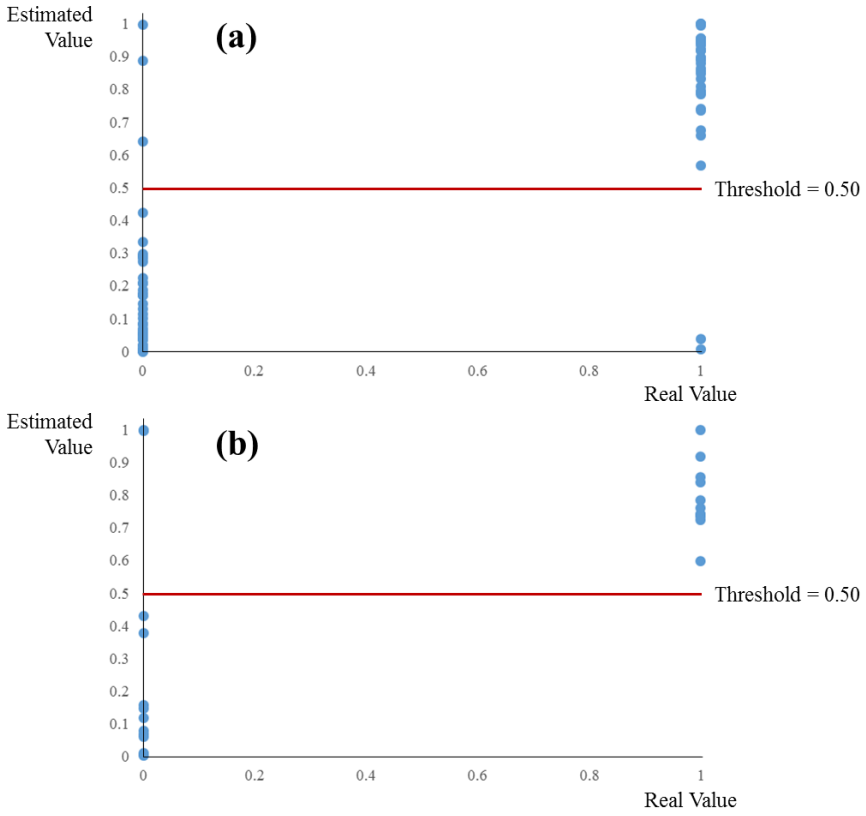

Figure 30. Graphical representation of the results offered by the fully optimized classifier that uses the independent variables selected by relief-F. Blue dots represent samples, and the red line is the optimized threshold. **(a)** Shows the results from the k-fold cross-validation test (93.8% accuracy; 75/80 correct hits) while **(b)** indicates those from the internal validation (95.8% accuracy; 23/24 correct hits).

The second part of the analysis covers the influence of glucose consumption on the exhaled breath samples, and it will be presented next.

3.3.4.3) Distinguishing LC Patients from Controls Considering Glucose Uptake

In this study, the effect that the OGT test has on the breath samples produced by the participants will be assessed. In other words, the evaluation of the role that glucose metabolism plays on the final volatile compounds and breath composition will be carried out. To do so, the first step that has been performed was the subtraction of the post-glucose uptake samples from their corresponding pre-glucose uptake ones (for every sample, which had already been normalized per individual). This way, a single sample for every participant is available, where the final m/z will represent the variation in the amount of VOCs originated by the consumption of glucose (subtracted m/z). Therefore, when attempting to locate the variables with the greatest power to distinguish samples from LC patients and controls, the FS algorithms will determine those m/z that have changed differently between the two groups after consuming glucose. In other words, those VOCs that vary their concentrations in breath differently after consuming glucose when comparing LC patients and controls, will be selected by the FS methods.

After this calculation, 40 samples formed the final database that has been used in this analysis (22 from controls and 18 from LC patients). Once again, 141 independent variables were

available, which in this case were the subtracted m/z. After labeling the samples (zeros were assigned to the controls and ones to the LC samples), the FS calculations were executed just like in the previous study. The results offered by the data filters can be seen in **Table 26**, where the first four variables selected by the FS algorithms are shown (four were selected to not surpass a 1:10 ratio of independent variables/samples for the following modeling tasks, evading possible over-fit models (Torrecilla *et al.*, 2013)).

Table 26. Variables (subtracted m/z) selected by the FS algorithms to distinguish breath samples from controls from LC cases (repeated m/z marked in bold).

FS Algorithm	Subtracted m/z Selected
χ^2	160, 162, 168, and 176
Fisher	**43**, **44**, 131, and **148**
Kruskal-Wallis	41, **43**, **44**, and 61
Relief-F	119, **142**, **148**, and 170
Information gain	108, 132, **142**, and 145

Just like in the previous study, several variables coincide in more than one FS test, which implies that although the algorithms use their own selection criteria, some of them show high relative discriminative power according to different statistical analyses. Each set of four subtracted m/z have been employed as independent variables or inputs for a series of classifying MLPs, which are intended to distinguish samples produced by LC patients and controls. If the results provided by these models are statistically robust, it might imply that those subtracted m/z employed represent volatile compounds in breath that change their concentrations differently after consuming glucose when comparing LC patients with controls. Potentially, this analysis may help locate subtracted m/z that are linked to volatile compounds in breath that might have a modified production due to the Warburg effect. This phenomenon, which was discovered back in the 1950s by Otto Warburg, is associated with cancerous cells and their altered glucose metabolism (Warburg, 1956). These cells typically produce energy through a high glycolysis rate and lactic acid fermentation in the cytosol of the cell, instead of the normal glycolysis and pyruvate oxidation that occurs in the mitochondria and leads to oxidative phosphorylation for a much higher energetic or ATP yield per glucose molecule that non-cancerous cells carry out (Vander Heiden *et al.*, 2009). Therefore, even in the presence of oxygen, tumor cells favor anaerobic glycolysis over oxidative phosphorylation, despite the much lower ATP production (Zheng, 2012).

The five MLP models designed possessed the same parameters and topology in order to reach comparable results, which are gathered in **Table 27** and **Figure 31**, where the parameters and the architecture can be found, respectively.

Table 27. MLP parameters and functions employed in the five classifying models (glucose uptake considered).

MLP Parameters	Selection or Value
Transfer function	Sigmoid
Training function	TrainLM
Lc	0.001
Lcd	0.1
Lci	10

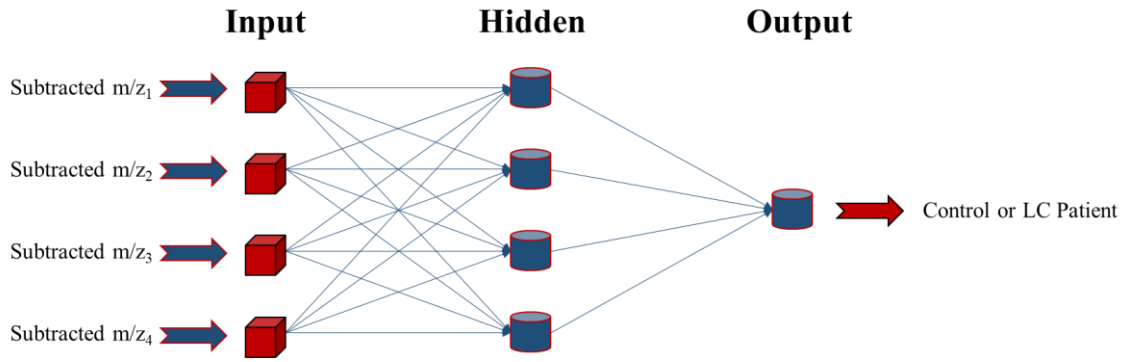

Figure 31. Architecture of the MLPs employed to distinguish breath samples from controls and LC patients, while taking glucose consumption into account. The final architecture is 4-4-1 (input-hidden-output).

In order to assess the statistical performance of the five MLP models, the same process was followed as the previous study. A threshold was optimized for each validation test (k-fold cross-validation (k = 6) (see **Figure 19**, section 2.4.2.3.1) and internal validation (see **Figure 20**, section 2.4.2.3.2)) of the binary classifiers to obtain the accuracy of the non-linear models in terms of correct hits (%). It is worth noting that, once again, the data points were randomly divided into the different datasets to correctly train and validate the applicability of the models, and that this division was analogous for all the MLPs in order to reach results which are as comparable as possible. The statistical performance of every model can be seen in **Table 28**.

In this case, the results show that, in general terms, two of the five models are considerably accurate (those using the variables selected by relief-F algorithm and information gain test), while the remaining three appear to be statistically worse models, especially regarding the internal validation. It is worth noting that the best all-round model, according to both kinds of validations, is the one that employs the variables selected by relief-F (at least 90% correct hits in both validation procedures), which matches with the previous study. Perhaps the non-linear relations in this database, between the clinical status of the participants and the compounds in their breath samples, are located better by this particular FS method when compared to the others (reflected in the more accurate MLPs).

Table 28. Statistical performance of the MLPs, in terms of correct hits (%), obtained through two validation processes (k-fold cross-validation and internal validation). Best results are marked in bold.

MLP	**k-Fold Cross-Validation** (n = 40)				**Internal Validation** (n = 12)			
	Th*	**Control** (n = 22)	**LC** (n = 18)	**Total**	**Th**	**Control** (n = 7)	**LC** (n = 5)	**Total**
χ^2	0.50	63.6%	77.8%	70.0%	0.46	85.7%	40.0%	66.7%
Fisher	0.37	86.4%	**94.4%**	**90.0%**	0.50	85.7%	60.0%	75.0%
Kr.-Wa.	0.50	72.7%	83.3%	77.5%	0.50	57.1%	60.0%	58.3%
Relief-F	0.53	**90.9%**	88.9%	**90.0%**	0.60	**100%**	**80.0%**	**91.7%**
Info gain	0.49	86.4%	**94.4%**	**90.0%**	0.50	85.7%	**80.0%**	83.3%

*__Th__ symbolizes threshold.

Due to the fact that the performance of the model that utilizes the four subtracted m/z selected by relief-F is the best of the five, the parameters and architecture of this MLP have been fully optimized (as explained in section 2.4.2.2). The final results can be seen in **Table 29**, including the statistical results of a k-fold cross-validation (k = 6) and an internal validation.

Table 29. Optimized MLP parameters and architecture of the binary classifier that distinguishes LC patients from healthy controls while considering glucose uptake, as well as the statistical performance of a k-fold cross-validation and an internal validation.

MLP Parameters and Topology	**Optimized Value or Statistical Performance**
Lc	0.001
Lcd	0.001
Lci	2
Inputs	4 (subtracted m/z 119, 142, 148, and 170)
Hidden neurons	4
Outputs	1
K-fold cross-validation performance (correct % controls/correct % LC patients/**correct % total**)	86.4/94.4/**90.0**
K-fold cross-validation threshold	0.39
Internal validation performance (correct % controls/correct % LC patients/**correct % total**)	100/80.0/**91.7**
Internal validation threshold	0.57

The statistical results of both of the validations have been graphically represented and can be seen in **Figure 32**.

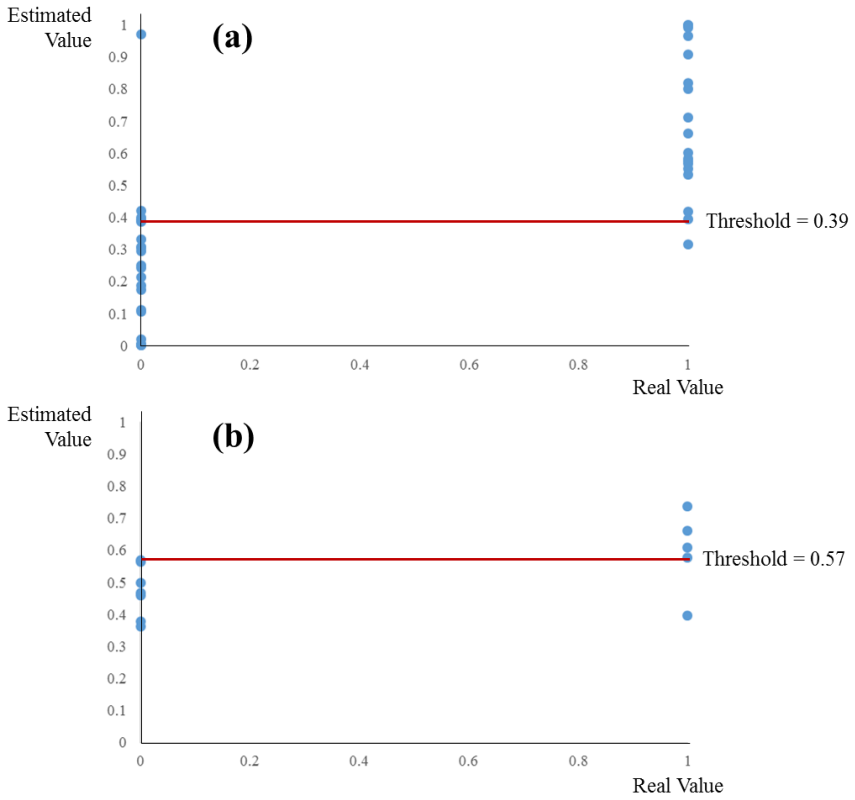

Figure 32. Charts of the results offered by the fully optimized binary classifier that uses the independent variables selected by relief-F. Blue dots symbolize samples, and the red line is the optimized threshold. **(a)** Represents the results from the k-fold cross-validation test (90% accuracy; 36/40 correct hits) whereas **(b)** shows those from the internal validation (91.7% accuracy; 11/12 correct hits).

These results show that accurate neural network models can be designed to distinguish LC patients from control participants, during an OGT test, using data retrieved from the analysis of breath samples through a PTR-MS system consisting of subtracted m/z (to consider the glucose consumption status). MLP models are able to provide compelling results during their meticulous validation, which guarantees the generalization ability and wide applicability range of the algorithmic tools.

The subtracted m/z selected by two of the FS methods have led to significantly accurate MLP models, which solely employ four independent variables each, during a 40 sample analysis. These subtracted m/z with discriminatory power (subtracted m/z from the two models that produced the most accurate results; 108, 119, 132, 142, 145, 148, and 170) may be associated with the Warburg effect, as they represent volatile compounds that possess different concentrations in breath before and after glucose consumption depending on their clinical status (LC vs control) (Vander Heiden *et al.*, 2009). Therefore, these compounds could represent the beginning of the development of a non-invasive LC diagnosing or screening system that relies on an OGT test, a PTR-MS-based breath analysis, and intelligent mathematical modeling.

With these results, the third experiment of the present thesis has been fully presented and discussed. To sum them up, it has been shown that using breath samples processed by a PTR-MS

system from LC patients and controls (high-risk individuals for LC) it is possible to attain reliable and non-invasive diagnosing tools after an appropriate mathematical treatment (FS and MLPs in this case). Two different scenarios have been analyzed, as the data has been collected during an OGT test. The first tool designed used data regardless of the glucose consumption of the individuals (pre- and post-glucose uptake breaths were studied as unrelated samples), while the second one did take it into account (pre- and post-glucose uptake samples from a specific participant were used as a single data point for the models, as they were subtracted). The final outcome of both kinds of MLP models was encouraging, as accurate mathematical tools were reached (never below 90% accuracy in terms of correct participant classification for any of the validation procedures carried out for the best model of each type). Therefore, it has been proven that this is a noteworthy approach which may lead to the design of captivating tools for the biomedical sector, as a safe and non-invasive diagnosing or screening system for LC has been potentially revealed. Some of these results have been presented at a multi-disciplinary meeting in Tel Aviv receiving the 3rd prize (Alkoby *et al.*, 2015), while a full and detailed version of this experiment has recently been accepted by a prestigious scientific journal of the breath analysis field (Feinberg *et al.*, 2016).

Finally, the fourth and last experiment of this research has been reached, and will be covered next. It relies on the use of functionalized GNP-based sensor arrays to identify patients of seven different diseases through their breath samples.

3.4) Non-Invasively Diagnosing Diseases by Combining Gold Nanoparticle Sensor Arrays and Neural Network Modeling to Analyze Breath Samples

This section will cover the results and discussion of the final experiment of this thesis. It is based on a study where cross-reactive sensor arrays based on functionalized gold nanoparticle (GNP) sensors have been used to analyze breath samples from different individuals. A set of 34 different GNP sensors will be employed during this experiment, which will include seven different population studies of seven diseases in order to reach individual mathematical models to distinguish sick patients from healthy controls using breath samples. The different diseases are chronic kidney disease, head and neck cancer, inflammatory bowel disease, multiple sclerosis, Parkinson's disease, preeclampsia, and pulmonary arterial hypertension.

Once the databases were produced, they were statistically analyzed (preliminary mathematical study) and, afterwards, treated with feature selection (FS) algorithms and multilayer perceptrons (MLPs) to reach classifying models (see **Figure 16**, section 2.4). The goals of this experiment can be separated into two. The obvious one is reaching mathematical systems that are able to distinguish samples that come from sick and healthy people, for each of the seven studies, only using data retrieved from breath samples. On the other hand, the FS algorithms are intended to locate those sensing features (or sensors) that have the greatest discriminative power, and, therefore, will be able to determine the specific sensors that are better suited for detecting determined diseases. This would enable the design of smaller and less expensive devices for particular sectors in the health field.

3.4.1) Breath Samples and Population Studies

Every breath sample from the seven different studies was attained as explained in the materials and methods section 2.1, and, during the next subsections, the traits of the volunteering participants of each individual study, which provided their sample after signing a written informed consent, will be presented.

3.4.1.1) Population Study 1 – Chronic Kidney Disease

The breath samples of 109 individuals were taken at the Poria Hospital (Tiberias, Israel) and analyzed to carry out this study. The participants are divided into 27 healthy control subjects and 82 chronic kidney disease (CKD) patients with different severity ranks. All of the patients went through an exhaustive physical examination and both routine blood and urine tests. The results were all available less than 30 days before the breath test. The patients were staged using their estimated glomerular filtration rate, which were determined from a set of parameters which included plasma creatinine levels, age, and gender, and using the equation of the modification of diet in renal disease (Levey *et al.*, 1999). The staging results, as well as additional information regarding the participants can be seen in **Table 30**.

Table 30. Relevant data from the CKD population study.

Data	CKD Patients	Healthy Controls
Amount of participants	82	27
Gender (male/female)	52/30	12/15
Age ± SEM*	65 ± 12	46 ± 2
Smoking status (current or past/never)	24/58	11/16
Staging results (early (1-2)/advanced (3-5))**	27/49	-

*SEM stands for standard error of the mean.
**The staging results of six CKD patients were not conclusive due to the results from the biochemical tests.

It must be noted that none of the patients had been under dialysis or had suffered a kidney transplant prior to the breath test (exclusion criteria).

3.4.1.2) Population Study 2 – Head and Neck Cancer

In this second case, breath samples from 63 people were taken at the Carmel Medical Center (Haifa, Israel). They were divided into 43 head and neck cancer (HNC) patients (different stages) and 20 healthy controls. The controls were matched to the patients in terms of age and lifestyle as best as possible, and they did not undergo any kind of examination (they were not aware of any kind of relevant medical condition). The main characteristics of the individuals can be seen in **Table 31**.

Table 31. Relevant data from the HNC population study.

Data	HNC Patients	Healthy Controls
Amount of participants	43	20
Gender (male/female)	37/6	6/14
Age ± SEM*	62 ± 12	50 ± 12
Smoking status (current or past/never)	25/18	5/15
Benign/Malignant	21/22	-

*SEM stands for standard error of the mean.

Some exclusion criteria were considered during this study and were applied prior to sampling. They are the following: (**a**) having any kind of medical history regarding malignancies or previous oncological treatment, (**b**) being under 18 years old, (**c**) possessing an active infectious disease, (**d**) being under an antibiotic treatment, (**e**) being pregnant, and (**f**) having an active

lactation period. Biopsies were taken from all the patients after collecting the breath samples (this fact did not delay the biopsy or alter the management protocol for any patient).

3.4.1.3) Population Study 3 – Inflammatory Bowel Disease

Regarding the inflammatory bowel disease (IBD) study, 170 total volunteers provided samples at the Rambam Medical Center (Haifa, Israel), of which 123 were sick patients and 47 were healthy controls. Among the sick patients, 97 had IBD (combining 49 cases of Crohn's disease (CD) and 48 cases of ulcerative colitis (UC)) while the remaining 26 had irritable bowel syndrome (IBS). Every participant was evaluated by a gastroenterologist and filled out a physician guided questionnaire. The groups were matched in terms of age, gender, body mass index (BMI), and smoking history as best as possible. These traits can be found in **Table 32**.

Table 32. Relevant data from the IBD population study.

Data	Patients			Healthy Controls
	CD (IBD)	UC (IBD)	IBS	
Amount of participants	49	48	26	47
Gender (male/female)	27/22	27/21	8/18	28/19
Age ± SEM*	38 ± 12	41 ± 16	38 ± 13	41 ± 2
BMI ± SEM*	23.9 ± 1.3	23.8 ± 0.8	23.2 ± 0.9	29.0 ± 1.0
Smoking status (current or past/never)	25/24	21/27	8/18	16/31

*SEM stands for standard error of the mean.

Every volunteer had to be at least 18 years old. The IBD patients that participated in the study were diagnosed by an expert gastroenterologist who employed common standards such as clinical presentation as well as radiologic, endoscopic, and histopathologic data. On the other hand, IBS patients met the Rome Criteria III (Drossman and Dumitrascu, 2006) as they manifested recurrent abdominal pain (or discomfort) at least during three days per month in the previous three months (their symptoms appeared to be unrelated to potential metabolic, inflammatory, or neoplastic processes). Finally, the healthy controls were randomly chosen from an unselected population, and it was verified that they did not present any gastrointestinal symptoms.

3.4.1.4) Population Study 4 – Multiple Sclerosis

In this study, breath samples from a total of 202 volunteers were gathered at the Carmel Medical Center (Haifa, Israel) and analyzed. There were 129 multiple sclerosis (MS) patients (most cases in remission and a small percentage in relapse phase) and 73 healthy controls which

were selected to match the patients in age and gender as best as possible. The main characteristics of the cohort can be found in **Table 33**.

Table 33. Relevant data from the MS population study.

Data	MS Patients	Healthy Controls
Amount of participants	129	73
Gender (male/female)	57/72	28/45
Age ± SEM*	38 ± 10	39 ± 11
Smoking status (current or past/never)	41/88	25/48
Remission/Relapse	112/17	-

*SEM stands for standard error of the mean.

There were a few exclusion criteria during this study which should be mentioned: (**a**) being below 18 years old, (**b**) being pregnant, (**c**) having HIV, hepatitis, or other severe and/or infectious diseases, and (**d**) having any type of autoimmune condition or up to a third degree family member with MS or any other autoimmune disease would exclude healthy control subjects.

3.4.1.5) Population Study 5 – Parkinson's Disease

Breath samples from 97 people were collected for this study at the Carmel Medical Center (Haifa, Israel), of which 60 had Parkinson's disease (PD) (considering as PD cases both idiopathic (44) and atypical Parkinsonism (16)) and 37 were healthy subjects. The patients were diagnosed by an experienced specialist and examined at least two times by a movement disorder expert. All of the patients went through a computerized tomography to exclude other potential diseases such as cancer. The control and sick groups were matched in terms of age and gender, and their main traits can be seen in **Table 34**.

Table 34. Relevant data from the PD population study.

Data	PD Patients		Healthy Controls
	Idiopathic	Atypical	
Amount of participants	44	16	37
Gender (male/female)	23/21	7/9	19/18
Age ± SEM*	65 ± 14	67 ± 8	62 ± 12
Smoking status (current or past/never)	7/37	6/10	9/28

*SEM stands for standard error of the mean.

All participants had to be over 18 years old (exclusion criterion) to be able to enter the population study.

3.4.1.6) Population Study 6 – Preeclampsia

During this study, the breath samples of 71 women were gathered at the Nazareth English Hospital (Nazareth, Israel) to carry out the analysis. It included 24 preeclampsia (PE) patients and 47 controls, of which 26 were healthy pregnant women and 21 were healthy non-pregnant women. All of the pregnant volunteers were past the 24th week of pregnancy, and the ones with PE had been diagnosed accordingly (blood pressure over 140/90 and proteinuria (Hawfield and Freedman, 2009)). The main characteristics of the participants in each group (age-matched) are shown in **Table 35**.

Table 35. Relevant data from the PE population study.

Data	PE Patients	Healthy Controls	
		Pregnant	Non-Pregnant
Amount of participants	24	26	21
Age ± SEM*	30 ± 6	29 ± 4	29 ± 4
Smoking status (current or past/never)	0/24	0/26	0/21

*SEM stands for standard error of the mean.

The healthy pregnant women were free of any kind of pregnancy complications, as well as chronic diseases. Also, the non-pregnant women did not possess any relevant medical history (diseases) or treatments. There were also some exclusion criteria in this study: (**a**) being under 18 years old, (**b**) having a pre-pregnancy body mass index greater than 35, (**c**) possessing any smoking history, and (**d**) having chronic diseases and/or treatments.

3.4.1.7) Population Study 7 – Pulmonary Arterial Hypertension

Finally, in this last study, 45 breath samples were collected at the Antoine-Béclère Hospital (Paris, France). From them, 22 were provided by pulmonary arterial hypertension (PAH) patients (7 heritable cases and 15 idiopathic) and 23 by healthy controls. Heritable PAH was diagnosed if mutations in the genes of the BMP/TGFβ family were detected (Sztrymf *et al.*, 2008) and/or if one or more cases of PAH had been diagnosed in their family (regardless of mutations). On the other hand, the idiopathic PAH cases were recognized as such after ruling out all other possibilities. The main traits of the population study can be found in **Table 36**.

Table 36. Relevant data from the PAH population study.

Data	PAH Patients		Healthy Controls
	Hereditary	Idiopathic	
Amount of participants	7	15	23
Gender (male/female)	2/5	4/11	10/13
Age ± SEM*	48 ± 12	47 ± 12	38 ± 8
Smoking status (current or past/never)	4/3	8/7	10/13

*SEM stands for standard error of the mean.

The main exclusion criterion was that the volunteers had to be older than 18 to be able to participate in the study.

This concludes the description of the different populations that are involved in this experiment. A total of 757 breath samples have been gathered and analyzed to design seven binary classifiers that are intended to identify patients from different diseases, and lead to potential diagnosing devices. To do so, the breath samples were first processed using a functionalized GNP sensor array which contained 34 sensors. This led to a vast amount of data (sensing features) per sample, which will be looked into next.

3.4.2) GNP Sensors and Sensing Features

As mentioned previously in the materials and methods subsection 2.3.2, a set of 34 different molecularly functionalized GNP sensors were prepared and used to create a sensor array to provide information regarding the breath samples gathered in this experiment. These samples were contained in adsorption tubes and, therefore, the trapped VOCs had to be transferred into the chamber where the sensors were. To do so, the samples were put through a thermal desorption process at 250ºC using a TD20 auto-sampling thermal desorption system (Shimadzu Corporation, Japan). The desorbed molecules were provisionally stored in a stainless steel column at 150ºC until the analysis, which was carried out soon after. Meanwhile, the chamber containing the sensor array was maintained under vacuum pressure (at approximately 30 mtorr) until the sample was introduced in it (the extra volume was filled with purified N_2 (99.999%) until atmospheric pressure was reached).

On the other hand, the measurements or resistance readings provided by the GNP sensors in the chamber were attained using a data logger device model 2701 DMM (Keithley Instruments, Inc., Cleveland, OH, USA). Resistance measurements (sensor signals) were gathered following the next sequence: (**a**) five minutes in vacuum (baseline), (**b**) five minutes interacting with the breath sample, and (**c**) five minutes of sensor recovery after evacuating the sample and reaching vacuum conditions once again inside the chamber. When the GNP sensors are exposed to a sample, the interaction between the gaseous molecules (VOCs) and the organic layer that functionalizes each nanoparticle leads to a modified resistance, which reaches baseline values

rapidly after the sample evacuation. From each one of these measurements, four sensing features were extracted. They were the change of resistance originated by the breath sample at the peak (maximum or minimum), at the middle of the measurement, and at the end of the measurement, as well as the area under the curve of the complete measurement (all features were linearly normalized between zero and one). The whole process was controlled through a custom-made LabView software, and an example of the obtained measurements (and sensing features) can be seen in **Figure 33**.

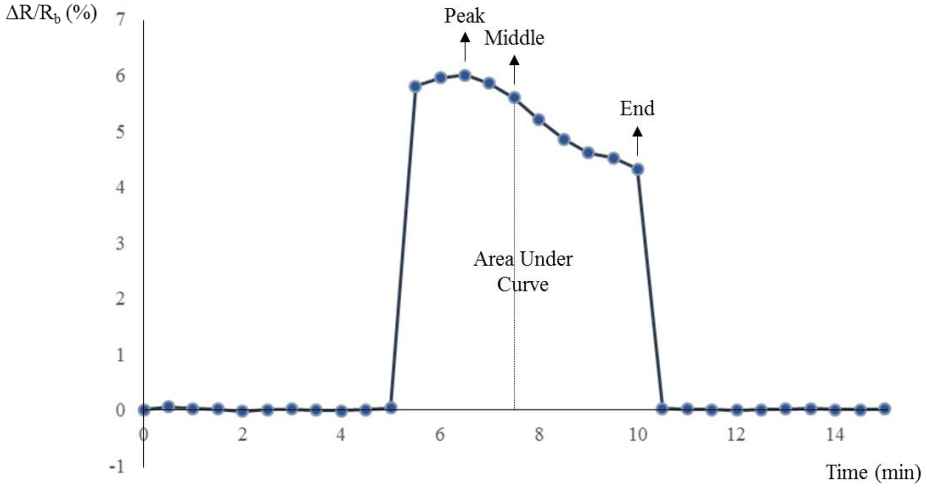

Figure 33. Graphical representation of the typical response provided by a GNP sensor during this study. R_b denotes the baseline resistance in vacuum conditions (first and last five minutes of the measurement), whereas ΔR symbolizes the baseline-corrected steady-state resistance change that occurs when the sensor is exposed to a breath sample (from minute five to ten). The four extracted sensing features can be seen: they were the $\Delta R/R_b$ at the peak, middle, and end of the breath sample exposure, as well as the total area under the curve.

As a precaution during this experiment, in order to monitor the responsiveness of the GNP sensors as well as mitigate the possible response drift, the sensors were calibrated daily employing a fixed gas mixture which contained 11.5 ppm of isopropyl alcohol, 2.8 ppm of trimethylbenzene, and 0.6 ppm of 2-ethylhexanol. All the measurements (raw data) of a particular day were normalized using the responses originated by the calibration gas, which ensured attaining comparable data from different days.

Therefore, now that the data that will be used during this analysis has been described, it is time to determine if a combination of particular sensing features are suited to reach accurate mathematical tools that can classify or locate sick patients from seven different diseases, only using information gathered from their exhaled breath samples.

3.4.3) Mathematical Treatment

In this section, the different steps followed to reach the final binary classifiers will be presented. It is mainly divided into two mathematical procedures based on an initial FS process using the five algorithms previously described in materials and methods section 2.4.1, followed by a non-linear modeling process using MLPs (see **Figure 16**, section 2.4). The goals of this experiment are to reach classifiers that can distinguish healthy controls from sick patients as

accurately as possible for each individual study, as well as locate those GNP sensors from the array that are better suited for specific diseases.

3.4.3.1) Feature Selection

Table 37. Results obtained from the FS process for each one of the seven databases (diseases).

Database	FS Algorithm	Amount of Features	Amount of Sensors Selected
CKD	χ^2	7	3
	Fisher		3
	Kruskal-Wallis		4
	Relief-F		3
	Information gain		3
HNC	χ^2	5	3
	Fisher		4
	Kruskal-Wallis		4
	Relief-F		3
	Information gain		4
IBD	χ^2	10	5
	Fisher		4
	Kruskal-Wallis		4
	Relief-F		4
	Information gain		5
MS	χ^2	19	6
	Fisher		7
	Kruskal-Wallis		6
	Relief-F		7
	Information gain		6
PD	χ^2	7	4
	Fisher		4
	Kruskal-Wallis		3
	Relief-F		3
	Information gain		4
PE	χ^2	5	4
	Fisher		3
	Kruskal-Wallis		3
	Relief-F		3
	Information gain		4
PAH	χ^2	4	3
	Fisher		4
	Kruskal-Wallis		3
	Relief-F		4
	Information gain		4

The first main calculation was based on the use of the five filter FS algorithms that have been described and used in the previous experiments. They were employed to locate those features from the GNP sensors with the highest discriminative power to distinguish breath samples from healthy controls from those originated by patients. Therefore, this was carried out seven times,

one per database, disease, or population study. As 34 GNP sensors have been used, and four sensing features have been extracted per sensor (see **Figure 33**, section 3.4.2), a total of 136 independent variables were available initially in each database before the FS process. All healthy controls were labeled with a zero, while the patients were given ones. The results obtained during this analysis are covered in **Table 37**.

As can be seen, in most cases there are more variables than sensors due to the fact that up to four features have been extracted per sensor (*vide supra*), enabling the algorithms to select multiple features from the same sensor as the ones with the greatest discriminative power in the global database. The amount of features that were selected per study depended on the amount of data points (samples) remaining after withdrawing the statistical outliers. The criterion followed was to stay below a 1:10 ratio of variables/samples in order to avoid potential over-fitting effects in the following modeling phase (Torrecilla *et al.*, 2013) (in no case this ratio was surpassed; there were always at least 10 times more data points than sensing features selected).

3.4.3.2) Multilayer Perceptrons

The next step during the analysis was to use the selected features as independent variables or inputs in a series of MLP models. Therefore, five comparable binary classifiers were designed and trained per database, leading to a total of 35 MLPs. The best of each study, in terms of statistical performance (correct hits (%)), will determine the set of independent variables (or, basically, sensors) with the greatest discriminatory power and, in the end, allow the location of the sensors that are best suited to detect specific diseases through breath analysis.

Within each database, all the non-linear models had to be comparable. In other words, they had to possess analogous parameters and topology. All of this information is gathered in **Table 38**. Regarding the amount of hidden neurons, it was set so the maximum value possible that would lead to a minimum of 2:1 ratio of samples/weights to avoid over-fitting MLPs (Cancilla *et al.*, 2015).

So as to determine the statistical performance of all 35 MLPs, a k-fold cross-validation (k = 6) (see **Figure 19**, section 2.4.2.3.1) was carried out for each one to determine their accuracy and generalization capability (the threshold was set at 0.5 for each classifier). Every MLP from a specific disease used equivalent training and verification datasets (randomly divided) to ensure that the results were comparable. This analysis reveals the best set of sensing features to distinguish healthy controls from patients of a particular disease. In **Table 39**, the results regarding the most accurate model for each database can be seen.

Table 38. MLP parameters and functions employed, as well as architecture of the different binary classifiers.

MLP Parameters and Topology	Selection or Value						
	CKD	**HNC**	**IBD**	**MS**	**PD**	**PE**	**PAH**
Transfer function	Sigmoid						
Training function	TrainLM						
Lc	0.001						
Lcd	0.1						
Lci	10						
Inputs	7	5	10	19	7	5	4
Hidden neurons	4						
Outputs	1						

Table 39. Statistical performance of the most accurate MLPs, in terms of correct hits (%), of each disease according to a k-fold cross-validation (k = 6). The amount of GNP employed and the FS method that led to their identification are also shown.

Database	FS	Amount of GNP Sensors	Healthy	Patients	Total
CKD	Info Gain	3	66.7%	87.5%	**79.1%**
HNC	Relief-F	3	85.0%	93.8%	**90.9%**
IBD	Relief-F	4	83.0%	80.6%	**81.4%**
MS	χ^2	6	71.4%	74.4%	**73.1%**
PD	Relief-F	3	86.5%	83.7%	**85.0%**
PE	Relief-F	3	88.5%	90.9%	**89.6%**
PAH	Relief-F	4	91.3%	90.9%	**91.1%**

As can be seen, five out of the seven best models were attained using the features provided by the Relief-F algorithm. Just like in the previous experiment (PTR-MS), the combination of this particular FS method and neural networks seem to work very efficiently, leading to relatively higher correct classification rates. In the next phase, these best MLPs for each disease will be

fully optimized by calculating the parameters and most suitable topology (as described in section 2.4.2.2). These results are gathered in **Table 40**.

Table 40. Optimized MLP parameters and architecture of the binary classifiers for each of the studied diseases, as well as the statistical performance of a k-fold cross-validation and an internal validation.

MLP Parameters and Topology	Optimized Value/Statistical Performance						
	CKD	**HNC**	**IBD**	**MS**	**PD**	**PE**	**PAH**
Lc	1	0.001	0.001	0.001	1	1	0.500
Lcd	0.001	0.1	0.001	1	0.001	0.001	0.500
Lci	2	10	10	2	2	100	51
Inputs	7	5	10	19	7	5	4
Hidden neurons	3	4	4	4	3	3	3
Outputs				1			
K-fold cross-validation performance (correct % controls/correct % patients/**correct % total**)	81.5 94.0 **89.6**	95.0 90.9 **92.4**	80.8 83.3 **82.2**	69.9 83.6 **78.5**	89.2 81.0 **84.8**	86.1 87.0 **86.4**	91.3 90.9 **91.1**
K-fold cross-validation threshold	0.47	0.48	0.50	0.50	0.49	0.50	0.49
Internal validation performance (correct % controls/correct % patients/**correct % total**)	70.0 92.8 **83.3**	83.3 83.3 **83.3**	78.6 84.2 **81.8**	85.2 84.8 **85.0**	77.8 93.3 **87.5**	81.8 100 **88.9**	85.7 87.5 **86.7**
Internal validation threshold	0.40	0.46	0.53	0.44	0.57	0.48	0.48

In general, the final optimized models provide fairly accurate results in terms of correct classification, proving that the combination of breath analysis, functionalized GNP-based sensing, and intelligent mathematical models can lead to tools that are able to discriminate among healthy people and others which present a disease. In other words, systems that are capable of detecting a wide variety of diseases through breath analysis have been achieved. The accuracy of the different models varies between 78.5% and 92.4% for a k-fold cross-validation and between 81.8% and 88.9% for an internal validation. The worst case is the model trained to detect MS (either in remission or relapse conditions) which correctly identifies 78.5% of the participants. In contrast, 92.4% of the samples are correctly classified in the case of the HNC study, 91.1% for the PAH population, and 89.6% for the CKD group, these three being the MLPs with the best statistical performances regarding the k-fold cross-validation. On the other hand, all models showed performances greater than 81% accuracy for the internal validation, which gives reliability and provides a wide generalization capability to these mathematical tools, as samples that were blind to the MLPs were accurately classified.

The results from this study, which will be submitted soon to a very prestigious scientific journal, have also allowed determining the best sensors from a 34 sensor array which offer

disease-specific signal patterns. These patterns were discovered by the FS algorithms and interpreted by the MLPs, enabling the development of much more cost-effective devices which only contain three to six sensors. These new "specialized" tools could potentially make their way into determined biomedical sectors, acting as non-invasive and safe disease detectors.

With this experiment, it has been proven that different diseases alter the volatile components in breath and that it is possible to take advantage of these changes to design non-invasive tools based on neural networks to aid in their detection. Hopefully, this study will keep encouraging research in this line, as it appears that disease diagnosis can be reached through a simple exhalation. In the next and final subsection of the results and discussion, the statistical results of all the disease detecting systems that have been optimized throughout these experiments will be analyzed and compared, to reach some final conclusions regarding their clinical relevance.

3.5) Analyzing and Comparing the Results of the Disease Detecting Models

In this last subsection, the results that have been attained regarding disease detection during the present thesis will be brought together and analyzed as a group, to summarize the importance behind these findings. As well, they will be compared to other results found in the literature from other research groups that also work with breath analysis to diagnose diseases.

In the first place, the statistical performances of the neural network-based models, specifically, multilayer perceptrons (MLPs), that have been created and trained during the three experiments that involve the use of real breath samples from a total of 1171 volunteers (sections 3.2, 3.3, and 3.4), can be seen graphically represented in **Figure 34**.

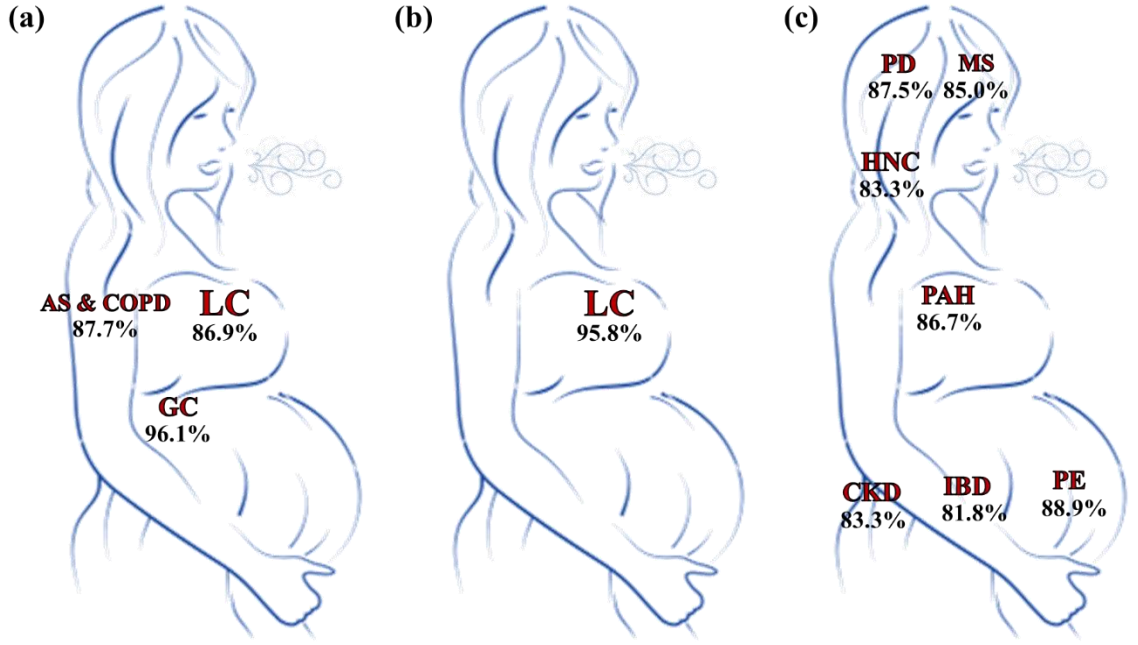

Figure 34. Representation of the total accuracies of the optimized neural network-based disease classifiers for each disease (asthma (AS), chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), gastric cancer (GC), head and neck cancer (HNC), inflammatory bowel disease (IBD), lung cancer (LC), multiple sclerosis (MS), Parkinson's disease (PD), preeclampsia (PE), and pulmonary arterial hypertension (PAH)). **(a)** Results from experiment covered in section 3.2, **(b)** from section 3.3, and **(c)** from section 3.4. The percentages presented symbolize the accuracies given by the models to discriminate between the given disease and a set of comparable healthy controls for an internal validation of the optimized mathematical tools.

As can be seen, all of the MLPs that have been designed and optimized lead to disease detectors that possess accuracies above 81% during an internal validation process (using samples that are not involved in the training process of the models to determine their performance; see **Figure 20**, section 2.4.2.3.2). These results confirm the existence of a true relationship between the clinical status of a person and the composition of his or her breath, as solely information gathered from these exhaled samples has been used to train the mathematical systems. Even diseases not directly related with the respiratory system such as CKD, MS, or PD seem to cause a relevant enough metabolic alteration that it leads to the modification of the composition of breath. These results strengthen the mentioned statement that this non-invasive approach should have its place in the medical field to serve as a reliable and complementary alternative to the current disease diagnosing methods.

Looking more into the results shown in **Figure 34**, it can be noticed that LC patients were evaluated during two different studies. In the first case, shown in **Figure 34-a**, these results correspond to those obtained during the experiment where the cross-reactive silicon nanowire field-effect transistor (SiNW FET) sensors were combined with MLPs to find patterns within the global breath samples to classify the patients (section 3.2). Here, 149 samples were gathered from LC patients and 129 from control volunteers. On the other hand, **Figure 34-b** shows the results from a study where proton transfer reaction-mass spectrometry (PTR-MS) was employed to attain quantitative information about the molecules present in the breath samples of 18 LC patients and 22 healthy controls (two independent samples were gathered per person). The gathered information was then processed and finally inputted into MLPs to reach, once again, classifying or disease detecting models (section 3.3). As can be noticed, there is around a 9% higher correct classification rate that favors the PTR-MS approach (96% versus 87%), which can be explained by the fact that the information is quantitative and specific for individual compounds in the breath, which are directly potential LC biomarkers, and, therefore, possess useful data in terms of disease diagnosis. In contrast, the information which SiNW FET sensors provide cannot be assigned to determined molecules, as the signals they offer are the result of analyzing the bulk of the breath sample. Afterwards, comparable patterns must be found between samples from the same group which will lead to correct classifications. Nevertheless, it is worth mentioning that the sample size of this experiment was much larger than the PTR-MS one, which provides a greater robustness to these models and their results. Additionally, the cross-reactive sensors can be contained in portable devices and are much more adaptable, customizable, and cost-effective than PTR-MS systems.

Finally, some of the results that have been obtained (**Figure 34**) will be compared to others that have been found in bibliographical references where breath analysis was employed as well to classify patients and healthy controls (not all the diseases are shown because they were not all found in the literature in a comparable fashion). In **Table 41**, this comparison can be found in terms of methodologies employed (analytical equipment and mathematical analysis), amount of breath samples analyzed, validation procedures, and statistical performances achieved.

In general terms, most of the comparisons seen in **Table 41** are versus other studies that employed cross-reactive sensor arrays (six out of seven), which is analogous to what was carried out in two of the experiments that have been covered here (sections 3.2 and 3.4). The amount of samples in the CKD, GC, HNC, MS, and PD analysis was larger in the present research (especially for the CKD, MS, and PD databases, where the samples were more than doubled), which enables the creation of more robust models with broader applicability spans. Moreover, the correct classification rate of the models that located CKD, GC, MS, and PD patients was higher when compared to the results found in the literature, despite the use of more samples. This fact reveals the power of the combination of cross-reactive sensors and neural network modeling.

It must be noted as well, that the statistical results which were obtained during this research were gathered from three independent and randomized internal validations per model (see **Figure 20**, section 2.4.2.3.2), which is a strict validating method that employs external data to test the performance of the model (comparable to blind validations). In contrast, five of the seven models that have been found employed different versions of cross-validations (CKD, GC, HNC, MS, and PD), which is perfectly valid, yet subject to potential random associations and over-fitting effects, which are avoided during internal validations (Cancilla *et al.*, 2015; Cohen-Kaminsky *et al.*, 2013).

Table 41. Comparison of the results attained with others found in the literature where breath analysis was carried out to detect diseases.

Disease	Current Research*	In the Literature*	Reference
CKD	Gold nanoparticle (GNP) sensors + MLPs/**109** samples/triple internal validation/**83**%	GNP sensors + support vector machine/**42** samples/cross-validation/**79**%	Marom *et al.*, 2012
GC	SiNW FET sensors + MLPs/**169** samples/triple internal validation/**96**%	GNP sensors + discriminant factor analysis/**130** samples/leave-one-out cross-validation/**90**%	Xu *et al.*, 2013
HNC	GNP sensors + MLPs/**63** samples/triple internal validation/**83**%	GNP sensors + support vector machine/**42** samples/cross-validation/**95**%	Hakim *et al.*, 2011
MS	GNP sensors + MLPs/**202** samples/triple internal validation/**85**%	Polycyclic aromatic hydrocarbons and single-wall carbon nanotube bilayer sensors + discriminant factor analysis/**51** samples/leave-one-out cross-validation/**80**%	Ionescu *et al.*, 2011
PD	GNP sensors + MLPs/**97** samples/triple internal validation/**88**%	Carbon nanotubes and GNP sensors + discriminant factor analysis/**42** samples/leave-one-out cross-validation/**78**%	Tisch *et al.*, 2013
PAH	GNP sensors + MLPs/**45** samples/triple internal validation/**87**%	GNP sensors + discriminant factor analysis/**45** samples/internal validation/**92**%	Cohen-Kaminsky *et al.*, 2013
		Ion-flow tube-mass spectrometry + discriminant factor analysis/**65** samples/internal validation/**83**%	Cikach *et al.*, 2014

* The following is shown: methodology or chemometric tool employed (analytical equipment + mathematical treatment)/**amount of samples**/validation method/**total accuracy** (%).

Another main difference is that five of the seven studies found in the literature revealed the use of linear mathematical modeling tools in the form of discriminant factor analysis (DFA), which is a common linear and supervised pattern recognition approach used for classifying purposes (Tisch *et al.*, 2013). Nonetheless, in occasions, DFA lacks the sufficient power to extract all the information contained in databases, especially when compared to non-linear algorithms such as artificial neural networks. In fact, in four out of the five cases where DFA was used (GC, MS, PD, and second PAH studies), the statistical performance of the MLPs was better, even when significantly larger datasets were involved, proving suitable the use of such non-linear tools for the design of non-invasive disease detectors. On the other hand, the other two studies found in bibliographic references (CKD and HNC) employed support vector machine (SVM) analysis to process the data. This mathematical approach is a supervised learning or pattern recognition method which locates a line that best separates samples from different classes, automatically using the most suitable features or variables to do so. It is also known for excelling when the databases are small (Marom *et al.*, 2012). In this case, when compared to the results given by the MLPs, it can be seen that the accuracy is higher for the CKD analysis, but lower for the HNC when compared to the results given by the SVMs. Nevertheless, it should be noted that in both cases the amount of samples employed was greater during the present research, and, additionally, that the feature selection (FS) process was thoroughly carried out with five different filter-based methods to try and locate the best independent variables to carry out the classifications.

With the end of this subsection, the results and discussion of this thesis conclude. Hopefully, it has been possible to transmit that reliable analytical equipment, such as cross-reactive sensors and PTR-MS, to process breath samples combined with powerful FS algorithms and sophisticated intelligent models, like artificial neural networks, is a worthy approach for the design of non-invasive breath-based disease detecting systems.

4) Conclusion

During this research, a set of four experimental sections involving breath analysis and disease diagnosis have been successfully carried out, revealing that there is information contained in these accessible biological samples that can aid in the detection of many threatening diseases if this data is extracted and analyzed properly. In this section, a list of conclusions that can be gathered from the findings of this research will be covered.

1. There are **clear correlations between the composition of gaseous samples and the signals given by cross-reactive sensors based on functionalized silicon nanowire field effect transistors** (SiNW FETs), as supervised artificial neural networks based on multilayer perceptrons (MLPs) can reliably and accurately identify and quantify volatile organic compounds (VOCs) in artificially prepared gaseous samples that are processed by these sensors (see section 3.1).

2. **Signals from a single functionalized SiNW FET sensor are enough to design accurate binary classifiers based on MLPs to distinguish breath samples from lung cancer** (LC)**, gastric cancer** (GC)**, chronic obstructive pulmonary disease** (COPD)**, and asthma** (AS) **patients, as well as samples from healthy controls**. Specifically, the signals given by a sensor functionalized with 3-aminopropyl-triethoxysilane (S11, **Table 15**, section 3.2.2) led to MLP-based models with statistical performance that ranged from about 87% to distinguish LC samples from controls to over 96% to discriminate LC from GC, GC from COPD&AS, and GC from controls, according to internal validation procedures (see **Figure 20**, section 2.4.2.3.2) (see section 3.2).

3. Algorithms based on **MLPs can successfully aid in the location of the most appropriate sensor from an array to classify breath samples from LC, GC, COPD, and AS patients, and healthy controls**, enabling the design of specific and cost-effective tools for particular purposes. These results can guide the synthesis of future sensors, as potentially viable sensor chemistries have been algorithmically located (see section 3.2).

4. **Proton transfer reaction-mass spectrometry** (PTR-MS) **is suitable to process breath samples from LC patients and high-risk yet healthy controls and originate quantitative data that can be used to classify both groups accurately with MLPs**. This was carried out during the course of an oral glucose tolerance test, enabling a dual study where glucose consumption was firstly not considered, and then it was. The results from both analyses were successful, as accuracies were never below 90% for any of the model validations (see section 3.3).

5. The quantitative nature of **PTR-MS allows the location of potential volatile biomarkers present in breath that could aid in LC diagnosis** (see section 3.3).

6. During the PTR-MS analysis, five **filter-based feature selection** (FS) **algorithms have been employed to locate the VOCs with the greatest discriminative power to classify LC patients and controls**. Several endogenous VOCs have been proposed as **potential LC biomarkers** as a result of the PTR-MS study, and they **are acetic acid, ethylbenzene, 1,2-dichlorobenzene, and glutamic acid**. Furthermore, during

the study that considers the effects of glucose consumption, **biomarkers that are affected by the Warburg effect can be theoretically identified** (see section 3.3).

7. From the five **FS** algorithms employed, specifically one of them, **Relief-F** (see section 2.4.1.4), **located the compounds that led to the development of the MLPs with the highest classification rates** in both cases during the PTR-MS study (considering glucose uptake and irrespective of it). **This fact reveals a powerful tool in the combination of these Relied-F and non-linear MLPs** (see section 3.3).

8. **SiNW FET sensors and PTR-MS have proven to be complementary approaches that can be used to detect LC through breath analysis and MLPs**. The former represents a cost-effective and fast approach that evaluates the bulk of the exhaled breath, and the latter reveals quantitative information of the VOCs, enabling the location of biomarkers, which are the true reason why samples from LC patients can be distinguished from ones obtained from healthy controls (see sections 3.2, 3.3, and 3.5).

9. **Cross-reactive functionalized gold nanoparticle (GNP)-based sensors combined with FS and MLPs can lead to the design of a wide assortment of precise breath-based** disease detectors, which include chronic kidney disease (CKD)**, head and neck cancer** (HNC)**, inflammatory bowel disease** (IBD)**, multiple sclerosis** (MS)**, Parkinson's disease** (PD)**, preeclampsia** (PE)**, and pulmonary arterial hypertension** (PAH), and they were all detected with high accuracies ranging from around 80% for MS to over 90% for HNC and PAH according to a k-fold cross-validation (see **Figure 19**, section 2.4.2.3.1) (see section 3.4).

10. **The signals that were mathematically selected for five out of the seven diseases** (HNC, IBD, PD, PE, and PAH) detected with GNP sensors **were provided by the Relief-F algorithm**, once again showing the power of the combination of this method with MLPs (see section 3.4).

11. The **FS** methods **enabled the reduction of a 34 sensor array to small sets of three to six sensors to accurately detect diseases through breath analysis**. This allows to determine the best chemical synthesis of sensors for particular applications and guide future research in this regard. Furthermore, **it leads to the possible design of much more specialized and less expensive tools** for particular medical sectors (see section 3.4).

The results obtained have proven the main hypothesis of this research, as it has been demonstrated that different diseases people may suffer lead to determined patterns in the composition of their exhaled breath, as only using information from these samples has allowed to design accurate disease detecting devices based on intelligent mathematical models. The main conclusion that can be extracted from this research, which had the invaluable cooperation of a grand total of 1171 volunteers, is that the composition of human breath is comparable to a book which reveals the clinical status of a human being, and that only some of the ways to successfully read it and use this knowledge to save people's lives and greatly improve their quality have been presented. It has been shown that the large-scale production of numerous non-invasive breath-based tools for many biomedical sectors to aid in the detection of a wide variety of diseases in an efficient, reliable, safe, and cost-effective fashion, should be just around the corner. The answer is only a breath away.

5) References

Abalos E, Cuesta C, Grosso AL, Chou D, and Say L. *Global and regional estimates of preeclampsia and eclampsia: a systematic review*. European Journal of Obstetrics & Gynecology and Reproductive Biology, 2013; 170 (1); 1-7.

Aggio R and Probert C. *Future Methods for the Diagnosis of Inflammatory Bowel Disease*. Digestive Diseases, 2014; 32 (4); 463-7.

Alkoby L, Abud-Hawa M, Bar J, Cancilla JC, Shlomi D, Feinberg T, Gai-Mor N, Haick H, Ilouze M, Onn A, Torrecilla JS, and Peled N. *Oral Glucose Tolerance Test as a Breath Challenge Diagnostic Test in Lung Cancer*. The 5th Multidisciplinary Israeli Conference of Lung Cancer, Dan Panorama Hotel, Tel-Aviv Israel, 2015.

Almstrand AC, Josefson M, Bredberg A, Lausmaa J, Siovall P, Larsson P, and Olin AC. *TOF-SIMS analysis of exhaled particles from patients with asthma and healthy controls*. European Respiratory Journal, 2012; 39 (1); 59-66.

Amman A, Costello B, Miekisch W, Schubert J, Buszewski B, Pleil J, Ratcliffe N, and Risby T. *The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva*. Journal of Breath Research, 2014; 8 (3); SI.

Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung BY, and Aggarwal BB. Cancer is a Preventable Disease that Requires Major Lifestyle Changes. Pharmaceutical Research, 2008; 25 (9); 2097-116.

Anonymous. *Biomarkers on a roll*. Nature Biotechnology, 2010; 28 (5); 431.

Aroca-Santos R, Cancilla JC, Matute G, and Torrecilla JS. *Identifying and Quantifying Adulterants in Extra Virgin Olive Oil of the Picual Varietal by Absorption Spectroscopy and Nonlinear Modeling*. Journal of Agricultural and Food Chemistry, 2015; 63; 5646-52.

Assad O, Leshansky AM, Wang B, Stelzner T, Christiansen S, and Haick H. *Spray-Coating Route for Highly Aligned and Large-Scale Arrays of Nanowires*. ACS Nano, 2012; 6 (6); 4702-12.

Bajtarevic A, Ager C, Pienz M, Klieber M, Schwarz K, Ligor M, Ligor T, Filipiak W, Denz H, Fiegl M, *et al. Noninvasive detection of lung cancer by analysis of exhaled breath*. BMC Cancer, 2009; 9; Article Number 348.

Basheer IA and Hajmeer M. *Artificial neural networks: fundamentals, computing, design, and application*. Journal of Microbiological Methods, 2000; 43; 3-31.

Bateman ED, Hurd SS, Barnes PJ, Bousquet J, Drazen JM, FitzGerald M, Gibson P, Ohta K, O'Byrne P, Pedersen SE, *et al*. *Global strategy for asthma management and prevention: GINA executive summary*. European Respiratory Journal, 2008; 31 (1); 143-78.

Befeler AS and Di Bisceglie AM. *Hepatocellular carcinoma: Diagnosis and treatment*. Gastroenterlogy, 2002; 122 (6); 1609-19.

Beneduce L, Castaldi F, Marino M, Tono N, Gatta A, Pontisso P, and Fassina G. *Improvement of liver cancer detection with simultaneous assessment of circulating levels of free alpha-fetoprotein (AFP) and AFP-IgM complexes*. International Journal of Biological Markers, 2004; 19 (2); 155-9.

Bernstein CN, Fried M, Krabshuis JH, Cohen H, Eliakim R, Fedail S, Gearry R, Goh KL, Hamid S, Khan AG, *et al*. *World Gastroenterology Organization Practice Guidelines for the Diagnosis and Management of IBD in 2010*. Inflammatory Bowel Diseases, 2010; 16 (1); 112-24.

Biomarkers Definitions Working Group. *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework*. Clinical Pharmacology & Therapeutics, 2001; 69(3); 89-95.

Bishop JM and Weinberg RA. *Scientific American Molecular Oncology*. Scientific American, Inc., New York, NY (USA) 1996.

Bjermer L, Alving K, Diamant Z, Magnussen H, Pavord I, Piacentini G, Price D, Roche N, Sastre J, Thomas M, *et al*. *Current evidence and future research needs for FeNO measurement in respiratory diseases*. Respiratory Medicine, 2014; 108 (6); 830-41.

Blake RS, Monks PS, and Ellis AM. *Proton-Transfer Reaction Mass Spectrometry*. Chemical Reviews, 2009; 109 (3); 861-96.

Blase X and Serra-Fernández MV. *Preserved Conductance in Covalently Functionalized Silicon Nanowires*. Physical Review Letters, 2008; 100 (4); Article Number 046802.

Block G, Patterson B, and Subar A. *Fruit, vegetables, and cancer prevention - a review of the epidemiologic evidence*. Nutrition and Cancer-An International Journal, 1992; 18 (1); 1-29.

Boots AW, van Berkel JJBN, Dallinga JW, Smolinska A, Wouters EF, and van Schooten FJ. *The versatile use of exhaled volatile organic compounds in human health and disease*. Journal of Breath Research, 2012; 6 (2); Article Number 027108.

Borkenstein RF. *Apparatus for Analyzing a Gas*. US Patent Number 2824789, Indianapolis, IN (USA) 1958.

Brust M, Fink J, Bethell D, Schiffrin DJ, and Kiely C. *Synthesis and reactions of functionalised gold nanoparticles*. Journal of the Chemical Society, Chemical Communications, 1995; 1655-6.

Buszewski B, Kesy M, Ligor T, and Amann A. *Human exhaled air analytics: biomarkers of diseases*. Biomedical Chromatography, 2007; 21; 553-66.

Calverley PMA, Anderson JA, Celli B, Ferguson GT, Jenkins C, Jones PW, Yates JC, Vestbo J, and TORCH Investigators. *Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease*. New England Journal of Medicine, 2007; 356 (8); 775-89.

Cancer Research UK website, 2016-a. Available from <http://www.cancerresearchuk.org/about-cancer/cancers-in-general/cancer-questions/how-many-different-types-of-cancer-are-there>.

Cancer Research UK website, 2016-b. Available from <http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/diagnosis/lung-cancer-tests>.

Cancilla JC, Díaz-Rodríguez P, Izquierdo JG, Bañares L, and Torrecilla JS. *Artificial neural networks applied to fluorescence studies for accurate determination of N-butylpyridinium chloride concentration in aqueous solution*. Sensors and Actuators B: Chemical, 2014-a; 198; 173-9.

Cancilla JC, Díaz-Rodríguez P, Matute G, and Torrecilla JS. *The accurate estimation of physicochemical properties of ternary mixtures containing ionic liquids via artificial neural networks*. Physical Chemistry Chemical Physics, 2015; 17; 4533-7.

Cancilla JC, Wang SC, Díaz-Rodríguez P, Matute G, Cancilla JD, and Torrecilla JS. *Linking Chemical Parameters to Sensory Panel Results through Neural Networks to Distinguish Olive Oil Quality*. Journal of Agricultural and Food Chemistry, 2014-b; 62; 10661-5.

Cao WQ and Duan YX. *Current status of methods and techniques for breath analysis*. Critical Reviews in Analytical Chemistry, 2007; 37 (1); 3-13.

Cappellin L, Loreto F, Aprea E, Romano A, del Pulgar JS, Gasperi F, and Biosioli F. *PTR-MS in Italy: A Multipurpose Sensor with Applications in Environmental, Agri-Food and Health Science*. Sensors, 2013; 13 (9); 11923-55.

Cauchi M, Fowler DP, Walter C, Turner C, Jia WJ, Whitehead RN, Griffiths L, Dawson C, Bai H, Waring RH, *et al*. *Application of gas chromatography mass spectrometry (GC-MS) in conjunction with multivariate classification for the diagnosis of gastrointestinal diseases*. Metabolomics, 2014; 10 (6); 1113-20.

Chan SY and Loscalzo J. *Pathogenic mechanisms of pulmonary arterial hypertension*. Journal of Molecular and Cellular Cardiology, 2008; 44 (1); 14-30.

Chandrashekar G and Sahin F. *A survey on feature selection methods*. Computers & Electrical Engineering, 2014; 40 (1); 16-28.

Chang CPY, Chia RH, Wu TL, Tsao KC, Sun CF, and Wu JT. *Elevated cell-free serum DNA detected in patients with myocardial infarction*. Clinical Chimica Acta, 2003; 327 (1-2); 95-101.

Chen KI, Li BR, and Chen YT. *Silicon Nanowire Field-Effect Transistor-Based Biosensors for Biomedical Diagnosis and Cellular Recording Investigation*. Nano Today, 2011; 6; 131-54.

Chey WD and Wong BCY. *American College of Gastroenterology Guideline on the Management of Helicobacter pylori Infection*. The American Journal of Gastroenterology, 2007; 102; 1808-25

Chin KM and Rubi LJ. *Pulmonary arterial hypertension*. Journal of the American College of Cardiology, 2008; 51 (16); 1527-38.

Cikach FS, Tonelli AR, Barnes J, Paschke K, Newman J, Grove D, Dababneh L, Wang SH, and Dweik RA. *Breath Analysis in Pulmonary Arterial Hypertension*. Chest, 2014; 145 (3); 551-8.

Clini E, Bianchi L, Pagani M, and Ambrosino N. *Endogenous nitric oxide in patients with stable COPD: correlates with severity of disease*. Thorax, 1998; 53 (10); 881-3.

Cohen-Kamisnsky S, Nakhleh M, Perros F, Montani D, Girerd B, Garcia G, Simonneau G, Haick H, and Humbert M. *A Proof of Concept for the Detection and Classification of Pulmonary Arterial Hypertension through Breath Analysis with a Sensor Array*. American Journal of Respiratory and Critical Care Medicine, 2013; 188 (6); 756-9.

Crew KD and Neugut AI. *Epidemiology of gastric cancer*. World Journal of Gastroenterology, 2006; 12 (3); 354-62.

Cui Y, Zhong Z, Wang D, and Lieber CM. *High Performance Silicon Nanowire Field Effect Transistors*. Nano Letters, 2003; 3 (2); 149-52.

Daniel MC and Astruc D. *Gold nanoparticles: Assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology*. Chemical Reviews, 2004; 104 (1); 293-346.

Dawson PH. *Quadrupole Mass Spectrometry and its Applications*. Elsevier Scientific Publishing Company, Amsterdam (the Netherlands) 1976.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, *et al*. *Orange: Data Mining Toolbox in Python*. Journal of Machine Learning Research, 2013; 14; 2349-53.

Demuth H, Beale M, and Hagan M. *Neural Network Toolbox for Use with MATLAB® User's Guide. Version 4.0.6*. Ninth printing Revised for Version 4.0.6 (Release 14SP3), Natick, MA (USA) 2005.

Dharnidharka VR, Kwon C, and Stevens G. *Serum cystatin C is superior to serum creatinine as a marker of kidney function: A meta-analysis*. American Journal of Kidney Diseases, 2002; 2; 221-6.

Dhir CS, Iqbal N, and Lee SY. *Efficient feature selection based on information gain criterion for face recognition*. International Conference on Information Acquisition, 2007; 1-2; 524-8.

Di Francesco F, Fuoco R, Trivella MG, and Ceccarini A. *Breath analysis: trends in techniques and clinical applications*. Microchemical Journal, 2005; 79; 405-10.

Dragonieri S, Annema JT, Schot R, van der Schee MPC, Spanevello A, Carratu P, Resta O, Rabe KF, and Sterk PJ. *An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD*. Lung Cancer, 2009; 64 (2); 166-70.

Drossman DA and Dumitrascu DL. *Rome III: New standard for functional gastrointestinal disorders*. Journal of Gastrointestinal and Liver Diseases: JGLD, 2006; 15 (3); 237-41.

Dryahina K, Spanel P, Pospisilova V, Sovova K, Hrdlicka L, Machkova N, Lukas M, and Smith D. *Quantification of pentane in exhaled breath, a potential biomarker of bowel disease, using selected ion flow tube mass spectrometry*. Rapid Communications in Mass Spectrometry, 2013; 27 (17); 1983-92.

Eisner MD, Anthonisen N, Coultas D, Kuenzli N, Perez-Padilla R, Postma D, Romieu I, Silverman EK, and Balmes JR. *An Official American Thoracic Society Public Policy Statement: Novel Risk Factors and the Global Burden of Chronic Obstructive Pulmonary Disease*. American Journal of Respiratory and Critical Care Medicine, 2010; 182 (5); 693-718.

Farrington SM, Lin-Goerke J, Ling J, Wang Y, Burczak JD, Robbins DJ, and Dunlop MG. *Systematic Analysis of hMSH2 and hMLH1 in Young Colon Cancer Patients and Controls*. American Journal of Human Genetics, 1998; 63; 749-59.

Feinberg T, Alkoby K, Herbig J, Cancilla JC, Torrecilla JS, Gaimor N, Bar J, Ilouze M, Haick H, and Peled N. *Cancerous glucose metabolism in lung cancer – evidence from exhaled breath analysis*. Journal of Breath Research, 2016 (recently accepted).

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, Forman D, and Bray F. *Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012*. European Journal of Cancer, 2013; 49 (6); 1374-403.

Floriano PN, Christodoulides N, Miller CS, Ebersole JL, Spertus J, Rose BG, Kinane DF, Novak MJ, Steinhubl S, Acosta S *et al*. *Use of Saliva-Based Nano-Biochip Tests for Acute Myocardial Infarction at the Point of Care: A Feasibility Study*. Clinical Chemistry, 2009; 55 (8); 1530-8.

Fock KM. *Review article: the epidemiology and prevention of gastric cancer*. Alimentary Pharmacology & Therapeutics, 2014; 40 (3); 250-60.

Frank IE and Friedman JH. *A statistical view of some chemometrics regression tools*. Technometrics, 1993; 35 (2); 109-35.

Galie N, Hoeper MM, Humbert M, Torbicki A, Vachiery JL, Barbera JA, Beghrtti M, Corris P, Gaine S, Gibbs JS, *et al*. *Guidelines for the diagnosis and treatment of pulmonary hypertension*. European Heart Journal, 2009; 30; 2493-537.

Geem ZW and Roper WE. *Energy demand estimation of South Korea using artificial neural network*. Energy Policy, 2009; 37 (10); 4049-54.

Glass CK, Saijo K, Winner B, Marchetto MC, and Gage FH. Mechanisms Underlying Inflammation in Neurodegeneration. Cell, 2010; 140 (6); 918-34.

Global Asthma Report website, 2016. Available from <http://www.globalasthmareport.org/burden/burden.php>.

Gnana-Sheela K and Deepa SN. *Review on Methods to Fix Number of Hidden Neurons in Neural Networks*. Mathematical Problems in Engineering, 2013; Article Number 425740.

Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Blaha MJ, Dai SF, Ford ES, Fox CS, Franco S, *et al*. *Heart Disease and Stroke Statistics-2014 Update. A Report From the American Heart Association*. Circulation, 2014; 129 (3); E28-E292.

Grabowska-Polanowska B, Faber J, Skowron M, Miarka P, Pietrzycka A, Sliwka I, and Amann A. *Detection of potential chronic kidney disease markers in breath using gas chromatography with mass-spectral detection coupled with thermal desorption method*. Journal of Chromatography A, 2013; 1301; 179-89.

Grammaticos PC and Diamantis A. *Useful known and unknown views of the father of modern medicine, Hippocrates and his teacher Democritus*. Hellenic Journal of Nuclear Medicine, 2008; 11 (1); 2-4.

Gruber M, Tisch U, Jeries R, Amal H, Hakim M, Ronen O, Marshak T, Zimmerman D, Israel O, Amiga E *et al*. *Analysis of exhaled breath for diagnosing head and neck squamous cell carcinoma: a feasibility study*. British Journal of Cancer, 2014; 111 (4); 790-8.

Guo L, Rivero D, Dorado J, Rabuñal JR, and Pazos A. *Automatic Epileptic Seizure Detection in EEGs Based on Line Length Feature and Artificial Neural Networks*. The Journal of Neuroscience, 2010; 191 (1); 101-9.

Guo SJ and Wang EK. *Synthesis and electrochemical applications of gold nanoparticles*. Analytica Chimica Acta, 2007; 598 (2); 181-92.

Guyon I and Elisseeff A. *An Introduction to Variable and Feature Selection*. Journal of Machine Learning Research, 2003; 3; 1157-82.

Haick H. *Chemical sensors based on molecularly modified metallic nanoparticles*. Journal of Physics D-Applied Physics, 2007; 40 (23); 7173-86.

Hakim M, Billan S, Tisch U, Peng G, Dvrokind I, Marom O, Abdah-Bortnyak R, Kuten A, and Haick H. *Diagnosis of head-and-neck cancer from exhaled breath*. British Journal of Cancer, 2011; 104 (10); 1649-55.

Halbert RJ, Natoli JL, Gano A, Badamgaray E, Buist AS, and Mannino DM. *Global burden of COPD: systematic review and meta-analysis*. European Respiratory Journal, 2006; 28 (3); 523-32.

Han MK, Agusti A, Calverley PM, Celli BR, Criner G, Curtis JL, Fabbri LM, Goldin JG, Jones PW, MacNee W, *et al*. *Chronic Obstructive Pulmonary Disease Phenotypes The Future of COPD*. American Journal of Respiratory and Critical Care Medicine, 2010; 182 (5); 598-604.

Hanahan D and Weinberg RA. *The hallmarks of cancer*. Cell, 2000; 100 (1); 57-70.

Hanahan D and Weinberg RA. *Hallmarks of Cancer: The Next Generation*. Cell, 2011; 144 (5); 646-74.

Hansel A, Jordan A, Holzinger R, Prazeller P, Vogel W, and Lindinger W. *Proton-transfer reaction mass-spectrometry - online trace gas-analysis at the ppb level*. International Journal of Mass Spectrometry, 1995; 149; 609-19.

Hawfield A and Freedman BI. *Pre-eclampsia: the pivotal role of the placenta in its pathophysiology and markers for early detection*. Therapeutic Advances in Cardiovascular Disease, 2009; 3 (1); 65-73.

Ho AS, Huang X, Cao H, Christman-Skieller C, Bennewith K, Le QT, and Koong AC. *Circulating miR-210 as a novel hypoxia marker in pancreatic cancer*. Translational Oncology, 2010; 3; 109-13.

Hoeper MM, Bogaard HJ, Condliffe R, Frantz R, Khanna D, Kurzyna M, Langleben D, Manes A, Satoh T, Torres F, *et al*. *Definitions and Diagnosis of Pulmonary Hypertension*. Journal of the American College of Cardiology, 2013; 62 (25); D42-D50.

Horowitz G. *Field-Effect Transistors.* Wiley-VCH, Weinheim (Germany) 1998.

Huang YX, Lemberg DA, Day AS, Dixon B, Leach S, Bujanover Y, Jaffe A, and Thomas PS. *Markers of Inflammation in the Breath in Paediatric Inflammatory Bowel Disease*. Journal of Pediatric Gastroenterology and Nutrition, 2014; 59 (4); 505-10.

Hubbard HF, Sobus JR, Pleil JD, Madden MC, and Tabucchi S. *Application of novel method to measure endogenous VOCs in exhaled breath condensate before and after exposure to diesel exhaust*. Biomedical and Life Sciences, 2009; 877 (29); 3652-8.

Humbert M, Sitbon O, Chaouat A, Bertocchi M, Habib G, Gressin V, Yaici A, Weitzenblum E, Cordier JFO, Chabot F, *et al*. *Pulmonary arterial hypertension in France - Results from a national registry*. American Journal of Respiratory and Critical Care Medicine, 2006; 173 (9); 1023-30.

Ionescu R, Broza Y, Shaltieli H, Sadeh D, Zilberman Y, Feng XL, Glass-Marmor L, Lejbkowicz I, Mullen K, Miller A, *et al*. *Detection of Multiple Sclerosis from Exhaled Breath Using Bilayers of Polycyclic Aromatic Hydrocarbons and Single-Wall Carbon Nanotubes*. ACS Chemical Neuroscience, 2011; 2 (12); 687-93.

Ionicon website, 2016-b. Available from <http://www.ionicon.com>.

Ionicon website, 2016-a. Available from <http://www.ionicon.com/information/technology/expert-information>.

Ishikawa S and Schrier RW. *Pathophysiological roles of arginine vasopressin and aquaporin-2 in impaired water excretion*. Clinical Endocrinology, 2003; 58 (1); 1-17.

Jain AK, Mao J, and Mohiuddin KM. *Artificial Neural Networks: A Tutorial*. Computer, 1996; 29 (3); 31-44.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, and Forman D. *Global Cancer Statistics*. CA-A Cancer Journal for Clinicians, 2011; 61 (2); 69-90.

Johnson DH, Marangos PJ, Forbes JT, Hainsworth JD, Vanwelch R, Hande KR, and Greco FA. *Potential utility of serum neuron-specific enolase levels in small cell-carcinoma of the lung*. Cancer Research, 1984; 44 (11); 5409-14.

Jung K, Fleischhacker M, and Rabien A. *Cell-free DNA in the blood as a solid tumor biomarker—A critical appraisal of the literature*. Clinica Chimica Acta, 2010; 411; 1611-24.

Karl T, PrazellerP, Mayr D, Jordan A, Rieder J, Fall R, and Lindinger W. *Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling*. Journal of Applied Physiology, 2001; 91 (2); 762-70.

Kharitonov SA and Barnes PJ. *Biomarkers of some pulmonary diseases in exhaled breath*. Biomarkers, 2002; 7 (1); 1-32.

Kim KH, Jahan SA, and Kabir E. *A review of breath analysis for diagnosis of human health*. Trac-Trends in Analytical Chemistry, 2012; 33; 1-8.

Kingwell E, Marriot JJ, Jette N, Pringsheim T, Makhani N, Morrow SA, Fisk JD, Evans C, Beland SG, Kulaga S, *et al*. *Incidence and prevalence of multiple sclerosis in Europe: a systematic review*. BMC Neurology, 2013; 13; Article Number UNSP 128.

Kinzler KW and Vogelstein B. Lessons from hereditary colorectal cancer. Cell, 1996; 87 (2); 159-70.

Kirkham PA and Barnes PJ. *Oxidative Stress in COPD*. Chest, 2013; 144 (1); 266-73.

Kleint C. *Julius Edgar Lilienfeld: Life and profession*. Progress in Surface Science, 1998; 57 (4); 253-327.

Knoerzer K, Juliano P, Roupas P, and Versteeg C. *Innovative Food Processing Technologies: Advances in Multiphysics Simulation*. Wiley-Blackwell, Oxford (UK) 2011.

Kohonen T. *Self-organization and Associative Memory, 3rd Edition*. Springer, New York, NY (USA) 1989.

Kolarik B, Wargocki P, Skorek-Osikowska A, and Wisthaler A. *The effect of a photocatalytic air purifier on indoor air quality quantified using different measuring methods*. Building and Environment, 2010; 45 (6); 1434-40.

Kosaka N, Iguchi H, and Ochiya T. *Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis*. Cancer Science, 2010; 101 (10); 2087-92.

Kozak KR, Amneus MW, Pusey SM, Su F, Luong MN, Luong SA, Reddy St, and Farias-Eisner R. *Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: Potential use in diagnosis and prognosis*. Proceedings of the National Academy of Sciences of the United States of America, 2003; 100 (21); 12343-8.

Kröse B and van der Smagt P. *An introduction to neural networks*. Eighth edition, the University of Amsterdam (the Netherlands) 1996.

Kruskal WH and Wallis WA. *Use of Ranks in One-Criterion Variance Analysis*. Journal of the American Statistical Association, 1952; 47 (260); 583-621.

Kumar S, Huang JZ, Abbassi-Ghadi N, Spanel P, Smith D, and Hanna GB. *Selected Ion Flow Tube Mass Spectrometry Analysis of Exhaled Breath for Volatile Organic Compound Profiling of Esophago-Gastric Cancer*. Analytical Chemistry, 2013; 85 (12); 6121-8.

Kumar R and Indrayan A. *Receiver Operating Characteristic (ROC) Curve for Medical Researchers*. Indian Pediatrics, 2011; 48 (4); 277-87.

Kynyk JA, Mastronarde JG, and McCallister JW. *Asthma, the sex difference*. Current Opinion in Pulmonary Medicine, 2011; 17 (1); 6-11.

Lau WY and Lai ECH. *Hepatocellular carcinoma: current management and recent advances*. Hepatobiliary & Pancreatic Diseases International, 2008; 7 (3); 237-57.

Lawrence J. *Data Preparation for a Neural Network.* Neural Network Special Report, Miller Freeman Inc., San Francisco, CA (USA) 1992.

Leemans CR, Braakhuis BJM, and Brakenhoff RH. *The molecular biology of head and neck cancer*. Nature Reviews Cancer, 2011; 11 (1); 9-22.

Leunis N, Boumans ML, Kremer B, Din S, Stobberingh E, Kessels AGH, and Kross KW. *Application of an Electronic Nose in the Diagnosis of Head and Neck Cancer*. Laryngoscope, 2014; 124 (6); 1377-81.

Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, and Roth D. *A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation*. Annals of Internal Medicine, 1999; 130 (6); 461-470.

Levey AS and Coresh J. *Chronic kidney disease*. Lancet, 2012; 379 (9811); 165-180.

Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, de Zeeuw D, Hostetter TH, Lameire N, and Eknoyan G. *Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)*. Kidney International, 2005; 67 (6); 2089-100.

Ligor T, Ager C, Schwarz K, Zebrowski W, Amann A, and Buszewski B. *Exhaled breath analysis Comparison of Proton Transfer Reaction-Mass Spectrometry and Gas Chromatography–Mass Spectrometry in Analysis of Breath Samples*. Chemia Analityczna, 2009; 54; 329-38.

Lindinger W, Hansel A, and Jordan A. *On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) - Medical applications, food control and environmental research*. International Journal of Mass Spectrometry, 1998; 173 (3); 191-241.

Ling Y, Johnson MK, Kiely DG, Condliffe R, Elliot CA, Simon J, Gibbs R, Howard LS, Pepke-Zaba J, Sheares KKK, *et al*. *Changing Demographics, Epidemiology, and Survival of Incident Pulmonary Arterial Hypertension Results from the Pulmonary Hypertension Registry of the United Kingdom and Ireland*. American Journal of Respiratory and Critical Care Medicine, 2012; 186 (8); 790-6.

Link H and Huang YM. *Oligoclonal bands in multiple sclerosis cerebrospinal fluid: An update on methodology and clinical usefulness*. Journal of Neuroimmunology, 2006; 180 (1-2); 17-28.

Liu H and Setiono R. *Chi^2: Feature selection and discretization of numeric attributes*. Proceedings of the IEEE 7th International Conference on tools with Artificial Intelligence, 1995; 388-91.

LNBD website, 2016. Available from <http://lnbd.technion.ac.il/>.

Malerba M, Ragnoli B, Buffoli L, Radaeli A, Ricci C, Lanzarotto F, and Lanzini A. *Exhaled nitric oxide as a marker of lung involvement in Crohn's disease*. International Journal of Immunopathology and Pharmacology, 2011; 24 (4); 1119-24.

Malfertheiner P, Megraud F, O'Morain C, Hungin APS, Jones R, Axon A, Graham DY, Tytgat G, Asaka M, Bazzoli F, *et al*. *Current concepts in the management of Helicobacter pylori infection - The Maastricht 2-2000 Consensus Report*. Alimentary Pharmacology & Therapeutics, 2002; 16 (2); 167-80.

Mansoor JK, Schelegle ES, Davis CE, Walby WF, Zhao WX, Aksenov AA, Pasamontes A, Figueroa J, and Allen R. *Analysis of Volatile Compounds in Exhaled Breath Condensate in Patients with Severe Pulmonary Arterial Hypertension*. Plos One, 2014; 9 (4); Article Number e95331.

Marcano-Cedeno A, Quintillana-Dominguez J, and Andina D. *WBCD breast cancer database classification applying artificial metaplasticity neural network*. Expert Systems with Applications, 2011; 38 (8); 9573-9.

Markar SR, Wiggins T, Kumar S, and Hanna GB. *Exhaled Breath Analysis for the Diagnosis and Assessment of Endoluminal Gastrointestinal Diseases*. Journal of Clinical Gastroenterology, 2015; 49 (1); 1-8.

Marom O, Nakhoul F, Tisch U, Shiban A, Abassi Z, and Haick H. *Gold nanoparticle sensors for detecting chronic kidney disease and disease progression*. Nanomedicine, 2012; 7 (5); 639-50.

Martin D (2002, August 17). *Robert F. Borkenstein, 89, Inventor of the Breathalyzer*. New York Times.

Marur S, D'Souza G, Westra WH, and Forastiere AA. *HPV-associated head and neck cancer: a virus-related cancer epidemic*. Lancet Oncology, 2010; 11 (8); 781-9.

Mathworks website, 2016. Available from <http://es.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html>.

Maurya P, Meleady P, Dowling P, and Clynes M. *Proteomic Approaches for Serum Biomarker Discovery in Cancer*. Anticancer Research, 2007; 27; 1247-56.

McCulloch WS and Pitts W. *A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics, 1943; 5; 115-33.

McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, McFarland HF, Paty DW, Polman CH, Reingold SC, *et al*. *Recommended diagnostic criteria for multiple sclerosis: Guidelines from the International Panel on the Diagnosis of Multiple Sclerosis.* Annals of Neurology, 2001; 50 (1); 121-7.

McGuire A, Martin M, Lenz C, and Sollano JA. *Treatment cost of non-small cell lung cancer in three European countries: comparisons across France, Germany, and England using administrative databases*. Journal of Medical Economics, 2015; 18 (7); 525-32.

McLachlan G, Do KA, and Ambroise C. *Analyzing Microarray Gene Expression Data*. John Wiley & Sons Inc., Hoboken, NJ (USA) 2004.

Miekisch W, Schubert JK, and Noeldge-Schomburg GFE. *Diagnostic potential of breath analysis - focus on volatile organic compounds*. Clinica Chimica Acta, 2004; 347 (1-2); 25-39.

Minsky M and Papert S. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA (USA) 1969.

Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, *et al*. *Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review*. Gastroenterology, 2012; 142 (1); 46-54.

Moretti M, Phillips M, Abouzeid A, Cataneo RN, and Greenberg J. *Increased breath markers of oxidative stress in normal pregnancy and in preeclampsia*. American Journal of Obstetrics and Gynecology, 2004; 190 (5); 1184-90.

Mowat C, Cole A, Windsor A, Ahmad T, Arnott I, Driscoll R, Mitton S, Orchard T, Rutter M, Younge L, *et al*. *Guidelines for the management of inflammatory bowel disease in adults*. Gut, 2011; 60 (5); 571-607.

Mukhopadhyay R. *Don't waste your breath*. Analytical Chemistry, 2004; 76 (15); 273A-6A.

Nakhleh M, Broza YY, and Haick H. *Monolayer-capped gold nanoparticles for disease detection from breath*. Nanomedicine, 2014; 9 (13); 1991-2002.

Newman JH, Trembath RC, Morse JA, Grunig E, Loyd JE, Adnot S, Coccolo F, Ventura C, Phillips JA, Knowles JA, *et al*. *Genetic basis of pulmonary arterial hypertension - Current understanding and future directions*. Journal of the American College of Cardiology, 2004; 43 (12); 33S-9S.

Ng EK, Chong WW, Jin H, Lam EK, Shin VY, Yu J, Poon TC, Ng SS, and Sung JJ. *Differential expression of microRNAs in plasma of patients with colorectal cancer: a potential marker for colorectal cancer screening*. Gut, 2009; 58 (10); 1375-81.

Nicholson WJ, Perkel G, and Selikoff IJ. *Occupational exposure to asbestos: population at risk and projected mortality--1980-2030*. American Journal of Industrial Medicine, 1982; 3 (3); 259-311.

Nobili S, Bruno L, Landini I, Napoli C, Bechi P, Tonelli F, Rubio CA, Mini E, and Nesi G. *Genomic and genetic alterations influence the progression of gastric cancer*. World Journal of Gastroenterology, 2011; 17 (3); 290-9.

Novakovic J. *Using Information Gain Attribute Evaluation to Classify Sonar Targets*. 17th Telecommunications forum TELFOR, 2009; 1351-4.

O'Sullivan BP and Freedman SD. *Cystic fibrosis*. Lancet, 2009; 373 (9678); 1891-904.

Pai SI and Westra WH. *Molecular Pathology of Head and Neck Cancer: Implications for Diagnosis, Prognosis, and Treatment*. Annual Review of Pathology Mechanisms of Disease, 2009; 4; 49-70.

Palacios N, Alonso A, Bronnum-Hansen H, and Ascherio A. *Smoking and Increased Risk of Multiple Sclerosis: Parallel Trends in the Sex Ratio Reinforce the Evidence*. Annals of Epidemiology, 2011; 21 (7); 536-42.

Palancar MC, Aragon JM, and Torrecilla JS. *pH-Control System Based on Artificial Neural Networks*. Industrial & Engineering Chemistry Research, 1998; 37 (7); 2729-40.

Palomar J, Torrecilla JS, Ferro VR, and Rodríguez F. *Development of an a Priori Ionic Liquid Design Tool. 2. Ionic Liquid Selection through the Prediction of COSMO-RS Molecular Descriptor by Inverse Neural Network*. Industrial & Engineering Chemistry Research, 2009; 48 (4); 2257-65.

Paone JF, Waalkes TP, Robinsonbaker R, and Shaper JH. *Serum UDP-galactosyl transferase as a potential biomarker for breast-carcinoma*. Journal of Surgical Oncology, 1980; 15 (1); 59-66.

Paredi P, Kharitonov SA, and Barnes PJ. *Elevation of exhaled ethane concentration in asthma*. American Journal of Respiratory and Critical Care Medicine, 2000-a; 162 (4); 1450-4.

Paredi P, Kharitonov SA, Leak D, Ward S, Cramer D, and Barnes PJ. *Exhaled ethane, a marker of lipid peroxidation, is elevated in chronic obstructive pulmonary disease*. American Journal of Respiratory and Critical Care Medicine, 2000-b; 162 (2); 369-73.

Parkin DM, Bray F, Ferlay J, and Pisani P. *Global cancer statistics, 2002*. CA-A Cancer Journal for Clinicians, 2005; 55 (2); 74-108.

Parmigiani G, Berry DA, and Aguilar O. *Determining Carrier Probabilities for Breast Cancer-Susceptibility Genes BRCA1 and BRCA2*. American Journal of Human Genetics, 1998; 62 (1); 145-58.

Paska Y and Haock H. *Controlling properties of field effect transistors by intermolecular cross-linking of molecular dipoles*. Applied Physics Letters, 2009; 95 (23); Article Number 233103.

Patel N, Alkhouri N, Eng K, Cikach F, Mahajan L, Yan C, Grove D, Rome ES, Lopez R, and Dweik RA. *Metabolomic analysis of breath volatile organic compounds reveals unique breath prints in children with inflammatory bowel disease: a pilot study*. Alimentary Pharmacology & Therapeutics, 2014; 40 (5); 498-507.

Pauling L, Robinson AB, Teranish R, and Cary P. *Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography*. Proceedings of the National Academy of Sciences of the United States of America, 1971; 68 (10); 2374-&.

Peacock AJ, Murphy NF, McMurray JJV, Caballero L, and Stewart S. *An epidemiological study of pulmonary arterial hypertension*. European Respiratory Journal, 2007; 30 (1); 104-9.

Peled N, Hakim M, Bunn PA, Miller YE, Kennedy TC, Mattei J, Mitchell JD, Hirsch FR, and Haick H. *Non-invasive Breath Analysis of Pulmonary Nodules*. Journal of Thoracic Oncology, 2012; 7; 1528-33.

Peng G, Tisch U, Adams O, Hakim M, Shehada N, Broza YY, Billan S, Abdah-Bortnyak R, Kuten A, and Haick H. *Diagnosing lung cancer in exhaled breath using gold nanoparticles*. Nature Nanotechnology, 2009; 4 (10); 669-73.

Peng G, Trock E, and Haick H. *Detecting Simulated Patterns of Lung Cancer Biomarkers by Random Network of Single-Walled Carbon Nanotubes Coated with Nonpolymeric Organic Materials*. Nano Letters, 2008; 8 (11); 3631-5.

Pfaffe T, Cooper-White J, Beyerlein P, Kostner K, and Punyadeera C. *Diagnostic Potential of Saliva: Current State and Future Applications*. Clinical Chemistry, 2011; 57 (5); 675-87.

Phillips M. *Breath tests in medicine*. Scientific American, 1992; 267 (1); 74-9.

Phillips M, Altorki N, Austin JHM, Cameron RB, Cataneo R, Greenberg J, Kloss R, Maxfield RA, Munawar M, Pass HI, *et al*. *Prediction of lung cancer using volatile biomarkers in breath*. Cancer Biomarkers: Section A of Disease Markers, 2007; 3 (2); 95-109.

Pijnenburg MWH and De Jongste JC. *Exhaled nitric oxide in childhood asthma: a review*. Clinical and Experimentl Allergy, 2008; 38 (2); 246-59.

Pinnock H, Thomas M, Tsiligianni I, Lisspers K, Ostrem A, Stallberg B, Yusuf O, Ryan D, Buffels J, Cals JWT, *et al*. *The International Primary Care Respiratory Group (IPCRG) Research Needs Statement 2010*. Primary Care Respiratory Journal, 2012; 19; S1-S20.

Pisitkun T, Johnstone R, and Knepper MA. *Discovery of urinary biomarkers*. Molecular & Cellular Proteomics, 2006; 5 (10); 1760-71.

Podolsky DK. *Inflammatory bowel disease*. New England Journal of Medicine, 2002; 347 (6); 417-29.

Poli D, Carbognani P, Corradi M, Goldoni M, Acampa O, Balbi B, Bianchi L, Rusca M, and Mutti A. *Exhaled volatile organic compounds in patients with non-small cell lung cancer: cross sectional and nested short-term follow-up study*. Respiratory Research, 2005; 6; Article Number 71.

Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, Fujihara K, Havrdova E, Hutchinson M, Kappos L, *et al*. *Diagnostic Criteria for Multiple Sclerosis: 2010 Revisions to the McDonald Criteria*. Annals of Neurology, 2011; 69 (2); 292-302.

Ramalingam SS, Owonikoko TK, and Khuri FR. *Lung Cancer: New Biological Insights and Recent Therapeutic Advances*. CA-A Cancer Journal for Clinicians, 2011; 61 (2); 91-112.

Redman CW and Sargent IL. *Latest advances in understanding preeclampsia*. Science, 2005; 308 (5728); 1592-4.

Risby TH and Solga SF. *Current status of clinical breath analysis*. Applied Physics B-Lasers and Optics, 2006; 85 (2-3); 421-6.

Robb KA, Simon AE, Miles A, and Wardle J. *Public perceptions of cancer: a qualitative study of the balance of positive and negative beliefs*. BMJ Open, 2014; 4 (7); Article Number e005434.

Robles AM and Shure D. *Gender issues in pulmonary vascular disease*. Clinics in Chest Medicine, 2004; 25 (2); 373-7.

Röck F, Barsan N, and Weimar U. *Electronic nose: Current status and future trends*. Chemical Reviews, 2008; 108 (2); 705-25.

Rogers TM, Grimsrud ER, Herndon SC, Jayne JT, Kolb CE, Allwine E, Westberg H, Lamb BK, Zavala M, Molina LT, *et al*. *On-road measurements of volatile organic compounds in the Mexico City metropolitan area using proton transfer reaction mass spectrometry*. International Journal of Mass Spectrometry, 2006; 252 (1); 26-37.

Roosta A, Setoodeh P, and Jahanmiri A. *Artificial Neural Network Modeling of Surface Tension for Pure Organic Compounds*. Industrial & Engineering Chemistry Research, 2012; 51 (1); 561-6.

Rosenblatt F. *The perceptron - a probabilistic model for information-storage and organization in the brain*. Psychological Review, 1958; 65 (6); 386-408.

Sánchez JM and Sacks RD. *GC analysis of human breath with a series-coupled column ensemble and a multibed sorption trap*. Analytical Chemistry, 2003; 75 (10); 2231-6.

Sanders AB. *Capnometry in emergency medicine*. Annals of Emergency Medicine, 1989; 18 (12); 1287-90.

Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, *et al*. *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis*. Nature, 2011; 476 (7359); 214-19.

Schapira AHV. *Neurobiology and treatment of Parkinson's disease*. Trends in Pharmacological Sciences, 2009; 30 (1); 41-7.

Schlect NF, Franco EL, Pintos J, and Kowalski LP. *Effect of smoking cessation and tobacco type on the risk of cancers of the upper aero-digestive tract in Brazil*. Epidemiology, 1999; 10 (4); 412-8.

Shehada N, Brönstrup G, Funka K, Christiansen S, Leja M, and Haick H. *Ultrasensitive Silicon Nanowire for Real-World Gas Sensing: Noninvasive Diagnosis of Cancer from Breath Volatolome*. Nano Letters, 2015; 15 (2); 1288-95.

Shehada N, Cancilla JC, Torrecilla JS, Pariente ES, Brönstrup G, Christiansen S, Johnson DW, Leja M, Davies MPA, Liran O, *et al*. *Multiparametric Molecularly Modified Silicon Nanowire Sensors for Multiple Disease Detection*. ACS Nano, 2016 (under revision).

Shirasu M and Touhara K. *The scent of disease: volatile organic compounds of the human body related to disease and disorder*. Journal of Biochemistry, 2011; 150 (3); 257-66.

Sibai B, Dekker G, and Kupferminc M. *Pre-eclampsia*. Lancet, 2005; 365 (9461); 785-99.

Siegel R, Ma J, Zou Z, and Jemal A. *Cancer Statistics, 2014*. CA-A Cancer Journal for Clinicians, 2014; 64; 9-29.

Smith D, Spanel P, Herbig J, and Beauchamp J. *Mass spectrometry for real-time quantitative breath analysis*. Journal of Breath Research, 2014; 8 (2); Article Number 027101.

Smolinska A, Hauschild AC, Fijten RRR, Dallinga JW, Baumbach J, and van Schooten FJ. *Current breathomics-a review on data pre-processing techniques and machine learning in metabolomics breath analysis*. Journal of Breath Research, 2014; 8 (2); Article Number 027105.

Soerjomataram I, Louwman MWJ, Ribot JG, Roukema JA, and Coebergh JWW. *An overview of prognostic factors for long-term survivors of breast cancer*. Breast Cancer Research and Treatment, 2008; 107 (3); 309-30.

Soleymani AR, Saien J, and Bayat H. *Artificial neural networks developed for prediction of dye decolorization efficiency with UV/$K_2S_2O_8$ process*. Chemical Engineering Journal, 2011; 170 (1); 29-35.

Spanel P and Smith D. *Volatile compounds in health and disease*. Current Opinion in Clinical Nutrition and Metabolic Care, 2011; 14 (5); 455-60.

Sze SM. *Semiconductor Devices*. *Physics and Technology*. John Wiley & Sons Inc., New York, NY (USA) 2001.

Sztrymf B, Coulet F, Girerd B, Yaici A, Jais X, Sitbon O, Montani D, Souza R, Simonneau G, Soubrier F, *et al*. *Clinical outcomes of pulmonary arterial hypertension in carriers of BMPR2 mutation*. American Journal of Respiratory and Critical Care Medicine, 2008; 177 (12); 1377-83.

Tansey MG and Goldberg MS. *Neuroinflammation in Parkinson's disease: Its role in neuronal death and implications for therapeutic intervention*. Neurobiology of Disease, 2010; 37 (3); 510-8.

Teranish R, Robinson AB, Cary P, Mon TR, and Pauling L. *Gas-chromatography of volatiles from breath and urine*. Analytical Chemistry, 1972; 44 (1); 18-&.

Tisch U, Billan S, Ilouze M, Phillips M, Peled N, and Haick H. *Volatile Organic Compounds in Exhaled Breath as Biomarkers for the Early Detection and Screening of Lung Cancer*. CML – Lung Cancer, 2012; 5 (4); 107-17.

Tisch U and Haick H. *Arrays of chemisensitive monolayer-capped metallic nanoparticles for diagnostic breath testing*. Reviews in Chemical Engineering, 2010-a. 26; 171-9.

Tisch U and Haick H. *Nanomaterials for cross-reactive sensor arrays*. MRS Bulletin, 2010-b; 35 (10); 797-803.

Tisch U, Schlesinger I, Ionescu R, Nassar M, Axelrod N, Robertman D, Tessler Y, Azar F, Marmur A, Aharon-Peretz J, *et al*. *Detection of Alzheimer's and Parkinson's disease from exhaled breath using nanomaterial-based sensors*. Nanomedicine, 2013; 8 (1); 43-56.

Torrecilla JS, Aragón JM, and Palancar MC. *Optimization of an Artificial Neural Network by Selecting the Training Function. Application to Olive Oil Mills Waste*. Industrial & Engineering Chemistry Research, 2008-b; 47; 7072-80.

Torrecilla JS, Deetlefs M, Seddon K, and Rodríguez F. *Estimation of Ternary Liquid-Liquid Equilibria for Arene/Alkene/Ionic Liquids Mixtures Using Neural Networks*. Physical Chemistry Chemical Physics, 2008-a; 10; 5114-20.

Torrecilla JS, Tortuero C, Cancilla JC, and Díaz-Rodríguez P. *Estimation with neural networks of the water content in imidazolium-based ionic liquids using their experimental density and viscosity values*. Talanta, 2013; 113; 93-8.

Trapp BD and Nave KA. *Multiple sclerosis: An immune or neurodegenerative disorder?* Annual Review of Neuroscience, 2008; 31; 247-69.

Van Berkel JJBN, Dallinga JW, Moller GM, Godschalk RWL, Moonen EJ, Wouters EFM, and van Schooten FJ. *A profile of volatile organic compounds in breath discriminates COPD patients from controls*. Respiratory Medicine, 2010; 104 (4); 557-63.

Vander Heiden MG, Cantley LC, and Thompson CB. *Understanding the Warburg effect: the metabolic requirements of cell proliferation*. Science (New York, NY), 2009; 324 (5930); 1029-33.

Viegi G, Pistelli F, Sherrill DL, Maio S, Baldacci S, and Carrozzi L. *Definition, epidemiology and natural history of COPD*. European Respiratory Journal, 2007; 30 (5); 993-1013.

Vishinkin R and Haick H. *Nanoscale Sensor Technologies for Disease Detection via Volatolomics*. Small, 2015; 11 (46); 6142-64.

Wallis AB, Saftlas AF, Hsia J, and Atrash HK. *Secular trends in the rates of preeclampsia, eclampsia, and gestational hypertension, United States, 1987-2004*. American Journal of Hypertnesion, 2008; 21 (5); 521-6.

Wang B, Cancilla JC, Torrecilla JS, and Haick H. *Artificial Sensing Intelligence with Silicon Nanowires for Ultraselective Detection in the Gas Phase*. Nano Letters, 2014; 14; 933-38.

Wang B and Haick H. *Effect of Chain Length on the Sensing of Volatile Organic Compounds by means of Silicon Nanowires*. ACS Applied Materials and Interfaces, 2013-b; 5 (12); 5748-56.

Wang B and Haick H. *Effect of Functional Groups on the Sensing Properties of Silicon Nanowires toward Volatile Compounds*. ACS Applied Materials and Interfaces, 2013-a; 5 (6); 2289-99.

Wang B, Huynh TP, Wu W, Hayek N, Do TT, Cancilla JC, Torrecilla JS, Nahid MM, Colwell JM, Gazit OM, *et al*. *A Highly Sensitive Diketopyrrolopyrrole-Based Ambipolar Transistor for Selective Detection and Discrimination of Xylene Isomers*. Advanced Materials, 2016; 28; 4012-18.

Wang S, Li D, Song X, Wei Y, and Li H. *A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification*. Expert Systems with Applications, 2011; 38; 8696-702.

Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, *et al*. *Multiple biomarkers for the prediction of first major cardiovascular events and death*. New England Journal of Medicine, 2006; 355 (25); 2631-9.

Warburg O. *On respiratory impairment in cancer cells*. Science, 1956; 124; 269-70.

Warwick G, Kotlyar E, Chow S, Thomas PS, and Yates DH. *Exhaled breath condensate in pulmonary arterial hypertension*. Journal of Breath Research, 2012; 6 (3); Article Number 036006.

Webster GT, de Villiers KA, Egan TJ, Deed S, Tilley L, Tobin MJ, Bambery KR, McNaughton D, and Wood BR. *Discriminating the Intraerythrocytic Lifecycle Stages of the Malaria Parasite Using Synchrotron FT-IR Microspectroscopy and an Artificial Neural Network*. Analytical Chemistry, 2009; 81 (7); 2516-24.

Wess M. Bringing hope and healing to grieving patients with cancer. The Journal of the American Osteopathic Association, 2007; 107 (12, Suppl 7); ES41-7

West D, Dellana S, and Qian JX. *Neural network ensemble strategies for financial decision applications*. Computers & Operations Research, 2005; 32 (10); 2543-59.

Wierzchos K, Cancilla JC, Torrecilla JS, Diaz-Rodriguez P, Davila AF, Ascaso C, Nienow J, McKay CP, and Wierzchos J. *Application of artificial neural networks as a tool for moisture prediction in microbially colonized halite in the Atacama Desert*. Journal of Geophysical Research Biogeosciences, 2015; 120 (6); 1018-26.

Winters WD. *New Techniques for detecting tumor-markers – A prospective*. Cancer Detection and Prevention, 1983; 6 (1-2); 21-31.

Wirdefeldt K, Adami HO, Cole P, Trichopoulos D, and Mandel J. *Epidemiology and etiology of Parkinson's disease: a review of the evidence*. European Journal of Epidemiology, 2011; 26; S1-S58.

Witt K, Inhestern J, Guntinas-Lichius O, and Voss A. *Application of an electronic nose to detect head and neck cancer from exhaled breath*. Biomedical Engineering-Biomedizinische Technik, 2012; 57 (Suppl. 1).

Wroblewski LE, Peek RM, and Wilson KT. *Helicobacter pylori and Gastric Cancer: Factors That Modulate Disease Risk*. Clinical Microbiology Reviews, 2010; 23 (4); 713-39.

Wu B, Chen CC, Kechadi TM, and Sun LY. *A comparative evaluation of filter-based feature selection methods for hyper-spectral band selection*. International Journal of Remote Sensing, 2013; 34 (22); 7974-90.

Xiao Z, Prieto D, Conrads TP, Veenstra TD, and Issaq HJ. *Proteomic patterns: their potential for disease diagnosis*. Molecular and Cellular Endocrinology, 2005; 230; 95-106.

Xu ZQ, Broza YY, Ionescu R, Tisch U, Ding L, Liu H, Song Q, Pan YY, Xiong FX, Gu KS, *et al*. *A nanomaterial-based breath test for distinguishing gastric cancer from benign gastric conditions*. British Journal of Cancer, 2013; 108 (4); 941-50.

Yeretzian C, Jordan A, and Lindinger W. *Analysing the headspace of coffee by proton-transfer-reaction mass-spectrometry*. International Journal of Mass Spectrometry, 2003; 223 (1-3); 115-39.

Zhan XF, Duan JN, and Duan YX. *Recent developments of proton-transfer reaction mass spectrometry (ptr-ms) and its applications in medical research.* Mass Spectrometry Reviews, 2013; 32(2); 143-65.

Zhang K, Li Y, Scarf P, and Ball A. *Feature selection for high-dimensional machinery fault diagnosis data using multiple models and Radial Basis Function networks*. Neurocomputing, 2011; 74 (17); 2941-52.

Zhang L, Xiao H, Karlan S, Zhou H, Gross J, Elashoff D, Akin D, Yan XM, Chia D, Karlan B *et al*. *Discovery and Preclinical Validation of Salivary Transcriptomic and Proteomic Biomarkers for the Non-Invasive Detection of Breast Cancer*. Plos One, 2010; 5 (12); Article Number e15573.

Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, and Liu H. *Advancing feature selection research – ASU feature selection repository*. Technical Report, Arizona State University, AZ (USA) 2011.

Zheng J. *Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review)*. Oncology letters, 2012; 4; 1151-7.

Zhou MG, Liu Y, and Duan YX. *Breath biomarkers in diagnosis of pulmonary diseases*. Clinica Chimica Acta, 2012; 413 (21-2); 1770-80.