Master's Degree in ICT for Smart Society

Preliminary investigation on Microwave Sensing for early detection of Lymphoedema

POLITECNICO DI TORINO

in collaboration with

UPPSALA UNIVERSITY

Supervisors

Prof. Monica VISINTIN

Prof. Guido PAGANA

Candidate

Daniele RUSSO

External Supervisors

Prof. Robin AUGUSTINE

Dr. Mauricio PEREZ

2020/2021

Abstract

In this work we investigate the potential of microwave sensors for the early detection of Lymph Oedema (LO) which could contribute in avoiding complications and reducing costs for the healthcare system. LO is a chronic disease characterized by the over-accumulation of lymphatic fluids in the body, which generally causes the swelling located in one arm or leg, sometimes both arms and/or legs. Although many medical solutions have been explored targeting the possibility of preventing and detecting the illness, especially considering high-risk factors and implying additional expansive tests, like Magnetic Resonance Imaging (MRI) or Computerized Tomography (CT) Scan, the idea of using microwave sensors can represent a valid, efficient and cost-effective alternative. Microwave sensor measurements paired with Machine Learning techniques may provide a feasible solution in the understanding of disease penetration in the body and, possibly, offer a more general comprehension on the structure laying underneath the skin. This thesis analyzes readings performed with a Split Ring Resonator (SRR) on patients undergone surgery, trying to apply clustering and classification methods to discover if any correlation between measurements and patient's metadata is present. In parallel, to interpret this processed information, simulations using CST Studio Suite software are performed emulating the three-layered structure of the human body, composed of skin, fat and muscle. In the simulations same sensor used for the real measurements is present and different scenarios are tested, including changing electro-magnetic parameters of the body materials and addition of Lymph layer in the 3D model. In the end, a deep description on the methodology to be adopted for future data collection will be proposed. This is done to avoid same problems occurred during previous acquisition and to allow better results in the machine learning phase. As a matter of facts, through some experiments performed for this thesis work, some erroneous behaviors have been discovered. Unfortunately, both for this reason and for lack of enough data, machine learning results do not seem to provide any robust result, but in future, possibly with an additional collection of patients' meta-data, this could be feasible.

Summary

In recent years the concern towards cancer-related diseases has grown due to many factors, like the increase of their incidence, the impact on healthcare systems and lack of early and effective diagnosis. To match this concern also interest towards economic, rapid and non-invasive medical solution that could significantly improve healthcare systems has raised. As a matter of facts, the research towards cost-effective solutions capable of enhancing or helping the prevention and diagnosis of these diseases has focused on alternatives of already widely available but costly solutions. For the above-mentioned reasons, the work of this thesis focuses on the analysis of Lymph Oedema (LO) through MicroWave (MW) sensors, collected during one year of follow-up on patients that had undergone a surgery.

LO is a chronic disease characterized by the over-accumulation of lymphatic fluids in the body, which generally causes the swelling located in one arm or leg, sometimes both arms and/or legs, but can affect any part of the body, like genitals, face, neck, chest wall and oral cavity. The capacity of the lymphatic system to transport this protein-rich fluid is exceeded, resulting in a progressive accumulation between the fibro-adipose tissue and the interstitium, i.e. the contiguous fluid-filled space existing between a structural barrier, such as a cell wall or the skin, and internal structures, such as organs.

The idea is to investigate the possibility of detecting skin-related illnesses exploiting a Split Ring Resonator (SRR) antenna connected to a miniVNA. In particular, the objective is to analyze a dataset obtained with the mentioned hardware, gathered by the department of Plastic and Reconstructive surgery at Uppsala University Hospital (Sweden), and detect any pattern that correlates the measurements to the presence and progression of the Lymphoedema. The Vector Network Analyzer (VNA) used is an instrument capable of measuring network parameters, like the Return Loss (RL), which is logarithmically dependent on the S-parameter; these parameters are used to characterize the electrical behaviour of linear electrical networks; in the considered scenario, the human body is seen as the electrical network to be investigated. The microstrip SRR antenna is the sensor used for the measurements: it can be seen as a parallel LC resonator architecture, where currents flow along rings located in its three layered structure. Exploiting the vnaJ software (which provides a Graphical User Interface for real-time readings and allows to store data in a Excel-readable format), three readings of six different

anatomical parts are saved in a dataset; this contains, for each patient, the available measurements performed in different periods with respect to the operation. It must be noted that not all readings are available: for problems related to a too fresh operation (bandage may be present or pain may be too much to put a sensor on the desired area) or to the missed appointment.

Once data were collected, a deep and precise data processing methodology was followed to re-organize and extract valuable information from this original dataset. In the first section, called *Preparation*, all missing file names were added, in such a way to have a correspondence between names of files/folders and the metadata available, i.e. the body position, the date with respect to the operation and if the affected limb was an arm or a leg. It follows the *Standardization* phase, where different readings characterized by different sampling frequencies were standardized considering 500 frequencies in the 2-3 GHz region; this stage was also important because the RL curve was extracted for each measurement, evaluated as the mean between three available readings. In this stage also the standard deviation of both peak amplitude and frequency was calculated, representing the confidence interval to rely on for future steps. Last phase was called *Filtering*, in which possible criteria used to filter out invalid data was first studied and then applied. As a result of this section, the final dataset was extracted, containing 60% of the original data and including patients metadata (body position and date measurement and patient's alias and Body Mass Index) and important information about the curve (bandwidth evaluated at -10 dB and peak frequency and amplitude.

In parallel to the data processing, few experiments were conducted to assess the data and hardware reliability. These were especially needed because during the filtering phase it was noticed a recurrent erroneous behaviour. To check the correct working of the SRR antenna, a comparison was carried out between measurements performed with both the miniVNA and a FieldFox VNA, which is much more reliable but less handy and more expansive, making it not suitable for hospital measurements. The second experiment aimed at detecting any problem in the instruments used: measurements performed with another miniVNA and another copper cable were compared with those obtained by same hardware used in the hospital; in addition, the calibration procedure was investigated, comparing two situation: a calibration performed using the cable attached between the SubMiniature version A (SMA) connector of the antenna and the Device Under Test (DUT) port of miniVNA and a calibration in which the cable is not present (improper calibration). Finally, last experiment was conducted to estimate the effects of temperature on the readings; in this case three scenarios were tested: ambient temperature, where the sensor was left on the desk for a couple of minutes before doing the measurement; warm temperature, in which basically the antenna was kept in contact with the skin for the whole duration of the experiment; cold temperature, where a bag full of snow was used to reduce temperature of the

antenna right before the measurement. The result of these experiments highlighted an erroneous calibration procedure frequently present in the dataset and that the cable used in the hospital was broken and needed to be replaced.

Concurrently to the data acquisition performed with the miniVNA, different simulations were carried out with the aim of obtaining a similar distribution of RL curves. This was done to correlate real data to a 3D model that could be used to better characterize the measurement. To do so, different scenarios were modeled thanks to CST Studio Suite software, providing few modifications to the default structure composed of three layers (skin, fat and muscle), whose ElectroMagnetic (EM) parameters were obtained from the IFAC database. Even though the obtained curves were quite dissimilar to the ones collected from the miniVNA, with a considerable shift to the left of the peak, an unexpected behaviour was noticed when varying the thickness of the skin: besides an almost constant left shift of RL peaks when the width increases, its amplitude grows until 2.25 mm are reached and then decreases again (the considered range is 0.5mm up to 30mm).

In order to find any pattern in the dataset, few Machine Learning (ML) techniques were tested. First of all, a clustering process was used to detect any correlation between RL curves and metadata available. This was done using two different clustering algorithms: KMeans, which is not capable of exploiting categorical features and for this reason was very efficient in differentiating groups of curves in the frequency spectrum, but did not highlight any particular pattern for other parameters; KModes, whose quality of considering categorical features resulted in a better differentiation for date and presence of the disease features, but did not lead to any characteristical curve that could identify the cluster. As a consequence, the Decision Tree (DT) classification was simpler in the first case, considering as important features only the bandwidth and the peak amplitude, and chaotic in the second one, where the output tree was too complex to be read. An additional attempt was done for the detection of the disease: using the same dataset, the binary class to predict was the presence of LO and the Ensamble method was used. Because of the paucity of data and the little reliability for some of them, the outcome was very poor, leading to final considerations reported in the conclusion.

In particular, regarding the conclusions, few crucial suggestions on the methodology to be adopted for a future data collection will be proposed. This is done to avoid same errors and problems occurred during previous acquisition and to allow better results in the machine learning phase. As a matter of facts, through some experiments performed for this thesis work, some erroneous behaviour have been discovered; these could be avoided with some simple but essential precautions. Unfortunately, both for this reason and for lack of enough data, machine learning results do not seem to provide any robust result, but in a future the same technique could be adopted to estimate the extent of more sub-cutaneous diseases or offer any meaningful information about the patient.

Acknowledgements

I am extremely grateful to my parents for their sacrifice, love, support and for educating and preparing me for my future. I am very thankful to all the members of my family for the incredible help they gave me in this very though period. I would like to thank Giorgia in particular for her strength, love and inspiration that allowed me to live this experience at its fullest. I would also like to thank my worldwide friends who supported me and offered the possibility of living a wonderful time, full of joy and fun.

I owe a deep sense of gratitude to my supervisors, Prof. Robin Augustine and Dr. Mauricio Perez, who passionately guided me throughout this project. I thank profusely the whole MMG group, that allowed me to have an amazing experience in Uppsala and supported me as a second family. I would like to express my deep and sincere gratitude to Prof. Guido Pagana for allowing me to live this dream experience, for all the help, care and support given. I would also like to thank prof. Monica Visintin for the assistance in completing this incredibly important goal.

My thanks and appreciations to all the people that stood by my side and to those who did not, allowing me to grow and be ready for everything.

"I have no special talents.
I am only passionately curious."
Albert Einstein

Table of Contents

List of Tables x

List of Figures xi

1 Introduction 1
- 1.1 Outline . . . 2
- 1.2 Objective and goals . . . 3

2 Literature review 5
- 2.1 eHealth . . . 5
- 2.2 Approach and challenges . . . 6
- 2.3 Lymph Oedema . . . 8
- 2.4 Microwave sensors . . . 9
- 2.5 Machine learning in medicine . . . 11
  - 2.5.1 Clustering . . . 12
  - 2.5.2 Classification . . . 13

3 Data collection and simulation 15
- 3.1 Real Measurements . . . 15
  - 3.1.1 Instruments . . . 15
  - 3.1.2 Setup & Output . . . 17
  - 3.1.3 Data Processing . . . 20
  - 3.1.4 Validation Procedure . . . 25
  - 3.1.5 Results . . . 28
- 3.2 Simulated Measurements . . . 45
  - 3.2.1 Software and 3D Model . . . 45
  - 3.2.2 Considered scenarios . . . 46
  - 3.2.3 Data processing . . . 52
  - 3.2.4 Results . . . 53

4 Clustering and classification 65
4.1 Clustering 65
4.2 Classification 66
4.3 Results 67

5 Conclusions & Future works 77

A Extra Plots 81

Bibliography 87

List of Tables

3.1 Sub-folder dates description 19
3.2 Body position w.r.t. Number 19
3.3 Final dataset - categorical 24
3.4 Final dataset - continuous 24
3.5 Interpolated return losses 29
3.6 Valid data distribution 32
3.7 Simulated data - pt.I 48
3.8 Simulated data - pt.II 51

List of Figures

3.1 miniVNA 16
3.2 SRR antenna 16
3.3 miniVNA calibration kit 17
3.4 Validation experiments 26
3.5 Folder disposition 28
3.6 Filtering process 29
3.7 Dataset Characterization 30
3.8 Filtering process - STD 32
3.9 Filtering process - Inconsistent 33
3.10 Filtering process - Corrupted 33
3.11 Filtering process - Inconsistent 34
3.12 Experiment - mVNA vs FFA 35
3.13 Experiment - broken cable 36
3.14 Experiment - good cable 36
3.15 Experiment - STD VS. 37
3.16 Experiment - temperature 38
3.17 Experiment - STD Temperature 38
3.18 Evolution - average 40
3.19 Evolution - ref. vs aff 42
3.20 Comparison - 01 vs. 02 43
3.21 SRR - Simulated 45
3.22 CST 3D model - default layout 47
3.23 CST 3D model - lymph layer pt.I 47
3.24 3D model - lymph layer pt.II 49
3.25 CST 3D model - Blood vessels 49
3.26 CST 3D model - Metamaterial 51
3.27 CST 3D model - Boundaries 52
3.28 Simulation - skin type 54
3.29 Simulation - lymph layer 55
3.30 Simulation - TEST 1 57

3.31 Simulation - TEST 2 . . . 57
3.32 Simulation - TEST 3 . . . 58
3.33 Simulation - TEST 4 . . . 59
3.34 Simulation - Infinite fat . . . 60
3.35 Simulation - Infinite muscle . . . 61
3.36 Simulation - Metamaterial . . . 63
3.37 Simulation - Recap . . . 64

4.1 KMeans - Curves clustered . . . 67
4.2 KMeans - Characterization . . . 68
4.3 KMeans - Feature importance . . . 69
4.4 KMeans - Separate curves . . . 70
4.5 KMeans - Histogram . . . 70
4.6 KMeans - Tree . . . 71
4.7 KModes - Curves clustered . . . 72
4.8 KModes - Characterization . . . 73
4.9 KModes - Feature importance . . . 73
4.10 KModes - Separate curves . . . 75
4.11 KModes - Histogram . . . 75
4.12 KModes - Tree . . . 76

A.1 Evolution - below patella . . . 82
A.2 Evolution - foot . . . 83
A.3 Evolution - shoulder . . . 84
A.4 Evolution - wrist . . . 85

Acronyms

BIA Bioelectrical Impedance Analysis. 1, 9, 79

BMI Body Mass Index. iv, 24, 68, 80

DT Decision Tree. v, 14, 66, 68, 72, 74, 80

DUT Device Under Test. iv, 10, 16, 18, 27, 35, 45, 46, 78

EM ElectroMagnetic. v, 9–11, 53, 78, 79

GUI Graphical User Interface. iii, 17

ICT Information and Communication Technology. 5–8

LO Lymph Oedema. iii, v, 1–3, 8, 9, 12, 20, 41, 44, 69, 77, 78

ML Machine Learning. v, 3, 6, 11, 21, 23, 65, 66, 79, 80

MSI Microwave Sensing and Imaging. 11

MW MicroWave. iii, 2, 4, 9–11, 39, 77

PEC Perfect Electrical Conductor. 50–52, 61

RL Return Loss. iii–v, 2, 18, 19, 21, 23, 25, 28, 29, 39, 46, 47, 50, 52–54, 62, 65, 67, 68, 77–80

SMA SubMiniature version A. iv, 16

SRR Split Ring Resonator. iii, iv, 2, 15, 16, 25, 26, 34, 39, 45, 49, 50, 67, 77–79

VNA Vector Network Analyzer. iii, iv, 2, 15, 17, 26

Chapter 1

Introduction

Lymph Oedema (LO), also known as lymphoedema or lymphatic edema, is a condition that presents localized swelling, generally located in the arms or legs. It is commonly caused by the damage of or removal of the lymph nodes, usually linked to the cancer treatment of the patient. The lymphatic system is blocked, resulting in a prevention of lymph fluid from draining and a consequent swelling. [1]

In order to improve the condition of patients it is important to monitor and comprehend the condition of the affected limb in an unprejudiced and clear way. On the other hand, the method used to evaluate the disease should not be invasive and possibly dangerous. X-rays, for example, could expose patient to additional risk and complications. For this reason, new experimental methods to characterize the tissue layers of the human body are being developed, exploiting different physical principles. Among all, the most common ones are the ultrasound, light and terahertz radiation.[2]

Lymph Oedema has a huge impact on society: it is estimated that more than one in five women who survive breast cancer will develop arm LO [3] and that it affects over 250 million people worldwide [4]. Despite the recent improvements in the treatment of cancer, LO is still representing one of the most important issues for those who already had survived a terrible disease. Besides the enormous psychological impact that it has on patients [5], it is also a very important problem for the healthcare systems, requiring more and more valuable resources. [6]

A volume difference between limbs larger than 10% is usually considered as a clinical approach to detect LO [7], but values smaller than 10% cannot be used as a certain parameter. In order to improve its diagnosis, various methods can be deployed: for example, the Bioelectrical Impedance Analysis (BIA) allows to recognize it directly measuring lymph fluid changes. It provides an easy to operate

and time-efficient tool, suitable for clinical practice. However, it does not give a total reliability and it has mainly been proposed as an integration for clinicians to assess the presence of the LO. [8] Other widely investigated methods for LO detection are based on 3D imaging: Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT) are probably the most common ones. These can demonstrate the alteration in epidermal and subcutaneous tissue, having the first one also the possibility of detecting obstructions caused by LO. However, the costs of these technologies are often an insurmountable problem that limits their adoption as a ordinary procedure for the diagnosis.

The main objective of this project is to characterize the composition of the anatomical district of a subject exploiting MicroWave (MW) measurements with MW technology. To evaluate the evolution of LO in time, measurements were obtained with an Split Ring Resonator (SRR) antenna connected to a MiniVNA and performed on patients before and after a surgery of oedema removal; after 1 month, 6 months and 1 year from the operation, and are still going on. The MiniVNA is a small, handy Vector Network Analyzer (VNA), i.e. an instrument to measure network parameters like the S-parameter (also called reflection parameter). The idea is to focus on the Return Loss (RL) value and characterize it by looking at the minimum value reached, its frequency and the bandwidth found at -10 dB.

1.1 Outline

To have a clear framework of the thesis, its structure will follow divided in chapters:

Chapter 1 - Introduction: after a brief introduction and the description on the topic discussed in this thesis, **Section 1.2** will be focused on the analysis of intended objectives and goals for this work;

Chapter 2 - Literature Review: in the second chapter the main aim will be reviewing documents concerning related work, in order to provide the reader with basic concepts and the state of art achieved. Five sections will be analyzed separately: **Section 2.1**, in which the concept of eHealth will be described and studied with few examples; **Section 2.2**, where the approach of eHealth will be further investigated, highlighting advantages and disadvantages for current state of the art; **Section 2.3** will be dedicated to the description of Lymph Oedema disease, examining major difficulties in its detection and prevention; **Section 2.4** will be a section that provides a brief overview of MW technologies applied to medical health, proposed as a solution for detecting LO, analyzing relative drawbacks; finally, **Section 2.5** will be about

one important aspect of the eHealth, i.e. Machine Learning (ML) techniques applied to problems related to medical field; here few examples will be reported, followed by a detailed description of clustering (**Section 2.5.1**) and classification (**Section 2.5.2**) algorithms used in this thesis.

Chapter 3 - Data collection and simulation: the third chapter is dedicated to the data collection process, distinguished in real measurements, **Section 3.1**, and simulated data, **Section 3.2**. Being the processing of these data the core of the whole project, a detailed description of both procedures will be presented, underlying the important aspects and phases of each of them. Concerning real measurements, deeper investigations will follow the hardware description (**Section 3.1.1**) and setup ((**Section 3.1.2**)), explaining the processing performed on acquired data (**Section 3.1.3**). Because of few unexpected behaviour noticed, (**Section 3.1.4**) will describe few experiments carried out to assess that the system used for the measurements was operating correctly. On the other hand, for the section dedicated to the simulations, after presenting the 3D model used (**Section 3.2.1**), different scenarios considered for each simulation will be described, following a brief explanation on why and how those could be useful (**Section 3.2.2**). As well as seen in previous section, a description of the processing needed for the simulation will be presented (**Section 3.2.3**). Both sections will be concluded with the achieved results, respectively **Section 3.1.5** for real measurements and **Section 3.2.4** for simulations.

Chapter 4 - Clustering and Classification: forth chapter focuses on the analysis of real data collected in previous chapter, review the data characterization performed previously and providing: the configuration of the clustering algorithms used and portraying their differences and the reasons for which they have been chosen (**Section 4.1**); a description of classification algorithms used to extract most important features and to detect the presence of LO (**Section 4.2**); results obtained with both procedures (**Section 4.3**).

Chapter 5 - Conclusions and Future Works: some considerations will be analyzed in this last chapter and conclusions will be drawn, taking into consideration how this project could be possibly improved in future and examining strengths and weaknesses that could provide considerable contribution to medical research.

1.2 Objective and goals

Since the main objective is ambitious, intermediate goals are needed to build a constructive procedure aiming at the precise identification of the affected zone of

the disease.

Data Validation First of all, the validation of measurements at disposal is needed: it will be required to compare different curves and see if any recurrent erroneous pattern is present or not. In this way it will be possible to exclude invalid data and, possibly, improve new measurements by exploiting metadata. To do so, some experiments are performed.

Body Characterization Since different limbs are being measured, beside the main idea to understand if it is possible to find a correspondence between measured data and the width of skin, fat and muscle layers, it would also be useful to look for a correspondence between affected or healthy limb and to understand if it is possible to uniquely characterize the limb and/or the body part.

Evolution Characterization Given that data were collected from patients that had undergone a surgery, the possibility of a using the MW readings to understand the status of the patient could be very useful. The idea will be to look for any correspondence between the output of the measurements and the time elapsed from the operation.

Chapter 2

Literature review

2.1 eHealth

The concept of **eHealth** is of increasing interest for the medicine domain and it is widely analyzed in the literature, bringing out how many applications are already present in numerous medical fields. Current challenges that the health systems are facing nowadays are mainly two: limited health budgets with increasing health costs and the concept of medicine that is rapidly changing, becoming more and more focused on the person necessities. Concerning the first point, it is interesting to notice that the costs have been rising at a faster rate than Gross Domestic Product (GDP) growth, [1] with a decrease in recent years. [2] Regarding the second challenge, a continuous and long-term monitoring for the enhancement of treatment efficiency, induced by several factors (among which it is possible to cite the aging of population and the generalization of chronic diseases), has been highlighted as a fundamental element towards which redirect the research.

eHealth is a concept that refers to the use of Information and Communication Technology (ICT) in healthcare to enhance the health system, providing cost-effective solutions and secure, reliable and efficient use of IT combined with robust communication systems. The Journal of Medical Internet Research [9] defines eHealth as:

[1] Health: spending continues to outpace economic growth in most OECD countries
`http://www.oecd.org/newsroom/healthspendingcontinuestooutpaceeconomicgrowthinmostoecdcountries.htm`

[2] Focus on Health Spending - OECD Health Statistics 2015
`https://www.oecd.org/health/health-systems/Focus-Health-Spending-2015.pdf`

> *"...an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies. In a broader sense, the term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve healthcare locally, regionally, and worldwide by using information and communication technology."*

Thus, besides the previously mentioned traits, this description points out an important aspect related to the eHealth: the state-of-mind approach for a further improvement of the current healthcare system. The enhancement given by ICT aiming for a better prevention, monitoring, diagnosis, treatment and management must be inserted in a framework where the former paradigm is abandoned to make way for a new one.

2.2 Approach and challenges

The main focus for the development of new technologies or the improvement of old ones following the eHealth paradigm is to put patients at the center of the decision: monitoring and prevention will no more be imposed by the healthcare system, but patients themselves will be encouraged to consider the advantages of self-monitoring through cheap and wide available devices [10] [11]; diagnosis will be significantly improved thanks to the deployment of ML techniques supported by doctors experience and vice versa, as their knowledge will be enhanced by the help of those techniques; treatment and follow-up on the patients will be easier thanks to an infrastructure that allows a better communication between patients and doctors.

To allow an instantaneous, efficient and detailed communication between doctors and people is one of the main objective in this environment and it must be noted that patients are not the only ones to be considered, since prevention is the backbone of the healthcare and monitoring on healthy people could save lives and money. The goal is to achieve prevention with non-invasive on-spot monitoring systems for people that are more exposed to risks of developing illnesses, even if they are not affected already. Few examples of eHealth integrated in everyday life follow, highlighting different challenges and critical requirements needed. A recent study [12] has proposed the usage of a real-time electrocardiographic tele-monitoring system to prevent any cardiac-related illness in athletes, addressing specifically sudden cardiac arrest (SCA). This approach suggested to monitor five runners during a marathon, evaluating cardiac parameters in real-time thanks to a device attached

to the chest and connected to cloud servers. While the device allows to acquire electrocardiogram waves from the runner, being positioned on the precordial position over the chest, and transfer timely to the a cloud server via cellular network, servers work as a tele-monitoring system, reading data from a electrocardiogram sensor and sending them through smartphones. Out of the five examined runners, three had reasonable percentage of analyzable data (between 63% and 99%): the inefficiency of measurements was associated with sternal shape, that led to recording instability (no data), and cellular network disturbance. The results seemed promising with a view to reduce the cost associated to the deployment of automated external defibrillators, crucial for saving lives; however, it will need significant improvement to allow a easy-to-connect network. This study proposed to exploit an already available system and apply some little changes; although this could be a feasible solution, one of the critical point in the development of a tele-monitoring system is the creation of an independent server and storage system in order to guarantee privacy of users and tailor the communication setup with the specific needs. Besides the privacy-related issues, availability and collection of data is another key aspect: not reliable, corrupt or unavailable data represent an serious problem for the ICT sector, problem that could only be solved with improved infrastructures and/or more robust devices.

However, an interesting part of this work is the usage of wearable devices to perform in loco monitoring. In fact, wearables are becoming more and more popular among people, both for professional and private usage. The wider accessibility to these accessories provides a great quantity of data to be analyzed and studied, besides the possibility to monitor the health statuses of people remotely and in real time. A remarkable example is given by the Internet of Things platform studied in [13]: the platform is deployed for smart maternal healthcare services, exploiting wearable medical sensors that provide real time feedbacks on pregnant women. The aim of this project is to reduce high-risk conditions during perinatal period, especially for patient presenting risk factors such as diabetes, elderly pregnant women's anaemia, gestational hypertension et cetera. These wearable devices would save huge quantity of time for the healthcare provider, considering that only one fetal hearth rate monitoring takes little less than one hour, and would improve considerably the condition of the maternity, making patients feeling at ease with the idea that everything is fine; besides, it can be used anytime and everywhere, avoiding unnecessary travels for pregnant women. As underlined in the paper too, one of the biggest challenges is the accuracy of the acquired data; among others, additional issues are related to the to the size and complexity in the configuration of available products, to the necessity of tailored and more cost-effective hardware and to the privacy of the collected data. On the other hand, the approach is patient-centered and is focused on improving both the quality of life and the medical assistance.

Another important topic related to ICT infrastructures in the medical field is the enhanced capability to allow patients and doctors rapidly communicate on a secure and reliable channel. As addressed in [14], where an ICT-based approach is used to achieve the empowerment of patients, chronic diseases (the article focuses on cancer, but same concept can be extended for many other chronic illnesses) require an increasing and more significant need for active rehabilitation of the patients. This leads to a raise in the interest of the self-management role of psychological, physical and social aspects of their health. The developed project reported in the paper (called iManageCancer) is a platform containing severe services, including games for increase awareness in kids and decision supporting tools, with the objective of stimulate cancer patients to have a more active role in the management of the illness; the ICT platform can be consequently seen as a system to monitor and improve psycho-emotional status of patients and improve the understanding of the disease, including families significantly in the management process.

2.3 Lymph Oedema

LO is a chronic disease characterized by the over-accumulation of lymphatic fluids in the body, which generally causes the swelling located in one arm or leg, sometimes both arms and/or legs, but can affect any part of the body, usually genitals, face, neck, chest wall and oral cavity. The capacity of the lymphatic system to transport this protein-rich fluid is exceeded, resulting in a progressive accumulation between the fibro-adipose tissue and the interstitium, i.e. the contiguous fluid-filled space existing between a structural barrier, such as a cell wall or the skin, and internal structures, such as organs. [15]

LO can be classified as primary (or genetic) and secondary (or acquired). While the first one is rare and inherited, affecting 1 in 100'000 individuals, the secondary type is the most common cause of the disease and affects approximately 1 in 1'000 Americans; its incidence is mainly studied in oncologic population: in fact, 1 in 5 women surviving a breast cancer developed Lymph Oedema [3]. In another 2017 study [16], 37 % of women treated for gynecological cancer developed LO within 12 months after the treatment. During the early stages, LO may be confused with simple edema and the adopted measures may not be sufficient nor adequate. For this reason it is necessary to uniquely identify it, taking in consideration all the risk factors, from the family medical records, highly influencing the possibility of developing any cancer-related disease [16], to any physical injury, that can induce the lymphatic system to react in an atypical way, especially for severe burns [17].

Generic symptoms can be summarized according to this list: [18]

1. *Edema*: the swelling due to the accumulation of excess fluids in the body;
2. *Hyperkeratosis*: the process in which skin becomes scaly and thicker;
3. *Lymphangioma*: the development of small blisters and bumps on the skin;
4. *Lymphorrhea*: the leakage of lymph fluid from the skin;

Concerning the disease diagnosis, blood, urine and tissue studies are not required: indeed, these tests could be useful to define the underlying causes of lower extremity edema, but will not help in the identification of the illness. In a similar way, although imaging is not needed, it could be useful as a confirmation, assessing the extent of the involvement and which treatment could be more adequate. As repeatedly highlighted in this Section 2.2, one of the fundamental elements for a worldwide diffusion of eHealth relies on the availability of new technologies, affordable and effective at the same time. In particular, regarding the imaging analysis a recurring concept is the absence of valid and reliable alternatives to the ones already implemented: emerging technologies, like 3D Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), ultrasound and BIA, are being used to improve the diagnosis process; even though CT and MRI can be very helpful for the identification of the tissue changes of the patient, given their good sensitivity and specificity, they are very expansive, both in terms of economy and time.

For the aforementioned reasons, a non-invasive, low-cost and accurate method for the identification and characterization of the LO could represent an important breakthrough in the medical field. Whereas MRI and CT have been largely discussed and examined in the literature [19] [20], MicroWave technologies have not been yet investigated extensively as a suitable alternative. The idea of this project is to understand if and how MW measurements could be used to estimate the extent of the LO and to predict any change in the development or remission following the patients' statuses in an year environment.

2.4 Microwave sensors

In recent years, the interest towards MicroWave techniques used in medical applications has considerably grown, also thanks to the fast development and wider availability of semiconductor technology and the discovery of new signal processing methods. In order to proceed, a clarification about the MW domain is needed. A MicroWave is an ElectroMagnetic (EM) wave in the range of 300 MHz and

300 GHz, thus located between infrared and short wave radio wavelengths in the EM spectrum. To analyze and process the interaction between MWs and matter, different kind of sensors are used; these can generically be grouped in resonators, transmission sensors, reflection and radar sensors radiometers, holographic and tomographic sensors and special sensors. [21] The principle on which these sensors are based is given by Equation (2.1) of the relative permittivity of the medium in which the signal propagates:

$$\epsilon_r = \epsilon_r' - j\epsilon_r'' \tag{2.1}$$

The permittivity is characteristical for each material used as medium and if a mixture of different materials is used, it depends on its components, its composition and its structure. In this way, in a mixture, given the permittivity of all materials except one, by evaluating the total permittivity of the mixture it will be possible to estimate the missing one. Since, other parameters, like temperature and density, may influence the permittivity and the mixture may present more than two components, multiparameter measurements are exploited, by considering other factors like resonant frequency, quality factor, insertion loss and phase et cetera.

Referring to the medical field, the **advantages** of MW sensors are listed below:

1. MWs can penetrate all materials except metal, a suitable scenario for measurements on human patients;
2. The capability of seeing a good contrast between water and most other materials makes them an optimal choice for water content measurements;
3. Environmental conditions do not affect significantly the measurement, allowing to achieve a good result independently by the surroundings;
4. MWs do not affect the Device Under Test (DUT) in any way, making MW measurements particularly suited for studying the human body.

Always considering medical applications, the **disadvantages** are also reported:

1. Despite a considerable decrease of prices in last two decades that made this technology more available, to achieve good accuracy level and higher frequencies, electronic components may be expansive;
2. The calibration procedure is crucial and must be performed separately for different materials;
3. Sensors are not adaptable to different applications, lacking of universality and complicating the research related to this topic;
4. Sensor are sensitive to more than one variable, requiring sometimes additional sensors for compensation.

Related to the field of medicine, MW application can be distinguished in:

1. **Medical Treatment**: medical treatments are mostly based on the thermal effect of locally generated MWs aimed at removing malignant tissues and improving the patient's condition; although results achieved with the MW-based techniques seem very promising, this current thesis work will be more focused on exploiting this technology for medical diagnosis and monitoring of patients;

2. **Medical Diagnosis**: the concept of MicroWave medical diagnosis is based on the analysis of a high-frequency signal scattering produced by dielectric difference; the dielectric characterization of bio-tissues allows to estimate if any anomaly is present and, consequently, evaluate the presence of a disease. Furthermore, from the study of MW measurements it is possible to create 2D or 3D images of various tissues, as investigated in the Microwave Sensing and Imaging (MSI) field. [22]

3. **Data Telemetry**: for data telemetry it is intended the procedure of exploiting EM signals to perform a wireless communication between medical and body-implanted devices; in the MW domain, these signals work at high-frequency, allowing the creation of networks with lower risk for the health. As an example, Fat-IBC (Intra-Body Communication) has been investigated in recent years as a possible alternative for the creation of networks connecting devices in the Body Area Network (BAN): fat layer is used as mean of communication instead of the air, guaranteeing more privacy and less interference.

2.5 Machine learning in medicine

The evolution of eHealth goes hand in hand with the improvement and the higher technology's availability. Remarkably, the development and integration of ML tools turned out to be one of the key element in the acknowledgment and characterization of diseases; the great enhancement comes from the objective interpretation of data collected, which would be otherwise impossible for any classification-related process in which many data were available. A better understanding of the human body not only helps in giving an improved picture of possible diseases, but could also predict them by looking at the risk factors whose correlation with the illness was not known before the advent of ML. [23]

Potentiality and effectiveness of ML techniques in the field of healthcare has shown excellent results in several aspects, such as emergency medicine management [24] and diagnosis prediction [25], [26]. However, the main focus in this section will be to address those kind of ML methods that best suit the classification

problem presented this thesis. Since a detailed description of the problem will be discussed later, for the moment a brief overview of the classification problem will be presented. The information collected and processed for this thesis lead to a double-sided problem: on one hand, it is necessary to be classify in clusters different data points, with both continuous and categorical features, to check if any pattern is present and possibly find an explanation to it; on the other hand, the same dataset will be used to train a classifier trying to predict the presence of the LO.

2.5.1 Clustering

In order to highlight any scheme in the dataset, with the idea of using the output of this process for a further classification, a clustering algorithm is needed. The principle is to group data samples in different clusters, taking in consideration that features are both continuous and categorical. To do so, two different well-known algorithms are used in parallel: **KMeans** and **KModes**. It must be noted that for both techniques the number of clusters is an input, i.e. either it is known a priori or different values should be tested to achieve best result.

KMeans: being $\mathbf{X} = \{x_1, ..., x_n\}$ the dataset in an f-dimensional space, $\mathbf{A} = \{a_1, ..., a_c\}$ the centers of each cluster and $\mathbf{z} = [z_{ik}]$ the vector describing if a data point i belongs to a cluster k, where n is the number of data points, f is the number of features and c the number of clusters, the objective function to be minimized is Equation (2.2):

$$J(z, A) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} ||x_i - a_k||^2 \tag{2.2}$$

At each iteration the values of a_k and z_{ik} are updated according to Equation (2.3) and Equation (2.4):

$$a_k = \frac{\sum_{i=1}^{n} z_{ik} x_i}{\sum_{i=1}^{n} z_{ik}} \tag{2.3}$$

$$z_{ik} = \begin{cases} 1, & \text{if } ||x_i - a_k||^2 = \min\limits_{1 \leq k \leq c} ||x_i - a_k||^2 \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

Numerous version have been developed on this basis, to face the recurrent problem of not knowing the number of clusters a priori; however, concerning the work of this thesis, it is reasonable to think that this implementation may be sufficient and, if needed, further investigations on the other versions will be

done. Nevertheless, one huge limit of the KMeans is related to its incapability of dealing with categorical features because, the attempt of minimizing an ordinary least-squares fitting function is not valid for categorical data, leading to a incorrect influence on the outcome. Besides, the concept of cluster center is not suitable for a category. In order to cope with these issues, KModes algorithm is used. [27]

KModes: this clustering algorithm is based on KMeans paradigm, plus, it is capable of clustering categorical data and producing a conceptual description of clusters. The similarity measure, used to estimate to which cluster the element belongs, is unchanged for numerical data. Now the mathematical preliminaries are described.

Being $\mathbf{X} = \{x_1, ..., x_n\}$ the dataset in an f-dimensional space; the number of cluster is c and the cost function is the trace of the within cluster dispersion matrix, as defined in Equation (2.5):

$$E = \sum_{l=1}^{c} \sum_{i=1}^{n} y_{il} d(X_i, Q_l) \tag{2.5}$$

Where $Q_l = [q_{l1}, ..., q_{lf}]$ is the *representative vector* or *prototype* for cluster l, y_{il} is an element of a partition matrix Y_{nxl} and d is a similarity measure, usually defined as the square Euclidean distance. When categorical features are present, the similarity measure is defined according to Equation (2.6):

$$d(X_i, Q_l) = \sum_{j=1}^{f_r} (x_{ij}^r - q_{ij}^r)^2 + \gamma_l \sum_{j=1}^{f_c} \delta(x_{ij}^c, q_{lj}^c) \tag{2.6}$$

$$where : \begin{cases} \delta(p,q) = 0, for p = q \\ \delta(p,q) = 1, for p \neq q \end{cases} \tag{2.7}$$

Here, x_{ij}^r and q_{lj}^r are values of numeric attributes, whereas x_{ij}^c and q_{lj}^c are values of categorical attributes for object i and the prototype of cluster l. f_r and f_c are the numbers of numeric and categorical attributes. γ_l is a weight for categorical attributes of cluster l. [28]

2.5.2 Classification

For the classification procedure, different techniques may be applied according to the type of classification required. In general, in order to test different algorithms and find the optimal one, a grid search that evaluates best hyper-parameters for each of them is always suggested: although it may require some additional time,

the output will be optimized with respect to all variables. In order to test different algorithms and find the one that fits the problem in a better way, **Ensamble Methods** is used: this technique allows to combine the prediction of different estimators, providing more robustness and improving generalizability.

According to the literature on this subject, among many classification algorithms, the**Decision Tree (DT)** classifier is considered the most appropriate for detecting most important features and create a tree that could be eventually used in hospital environments to help doctors with the diagnosis process. In particular, this family of classifiers allows to identify the most important features based on the connection between data and output class; it also provides instruments to categorize new unlabelled data by hand. Each tree is composed of leaves and branches; branches are created with the so-called "*test nodes*"; a branch ends when a "*leaf node*" is found. Leaves will correspond to the class, but classes can be reached with different branches and at different leaves.

Chapter 3

Data collection and simulation

3.1 Real Measurements

The collection of data is an ongoing process in collaboration with the Uppsala Hospital. Patients referred to the department of Plastic and Reconstructive surgery at Uppsala University Hosptial, Sweden, for secondary lymphedema, were examined as part of the study "Surgical treatment of Lymphedema", ethical approval 2016/470 and radiation ethical approval D16/53. This examination was performed at the outpatient clinic at Uppsala University Hospital after the patient was given information and patient left oral and written consent to participation in the study. Generally, the disease affects either arms or legs, usually only one, but few exception may be present; in this experiment both limbs are measured in order to have a reference value.

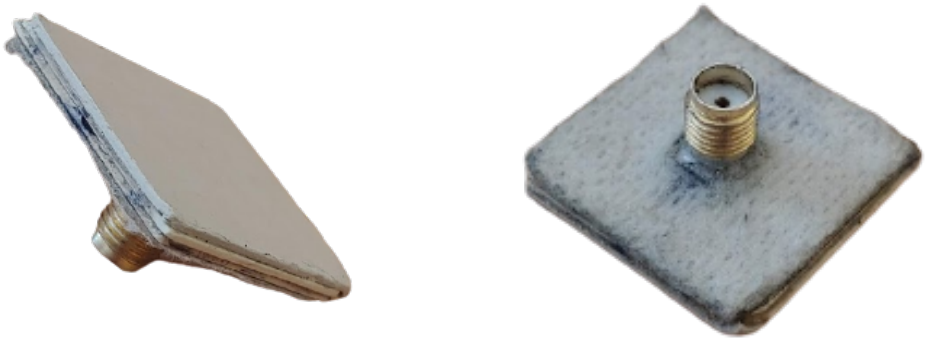
3.1.1 Instruments

In order to perform the measurements a **MiniVNA Tiny** was used. This small and handy object, showed in Figure 3.1, is a VNA, which is an instrument capable of measuring network parameters; with respect to Scalar Network Analyzer (SNA), it can measure both amplitude and phase properties. These parameters, usually linked to the S-parameters, are used to characterize electrical behavior of linear electrical networks.

In order to perform the measurement, a **sensor** was needed. The sensor used consists of a microstrip SRR, composed with three layers, two metal rings separated by one gap layer, in which a magnetic resonance is induced by splits at the rings

Figure 3.1: MiniVNA Tiny hardware

and by the gap between the inner and outer rings. In general, SRR behaves like as parallel LC resonating structure. The induced resonating currents flow along the rings with charges preventing the current from flowing around the ring and the circuit is completed across the small capacitive gap. [29]



(a) Side-view of the SRR **(b)** Top-view of the SRR

Figure 3.2: SRR sensor used for the measurements

To link the SubMiniature version A (SMA) connector (through the DUT port) to the sensor, a **cable** was deployed. The cable used is a male-to-male copper cable that has been changed as consequence of a problem found during latest measurements and confirmed with few experiments; this will be better discussed in section Section 3.1.4. In short, the issues encountered were related either to damage of the cable, due to improper storage and/or excessive bending, and to an recurrent incorrect calibration process.

The measurements were performed using the miniVNA connected to a computer in which the **vnaJ** software was running. This is an application completely written using JAVA programming language that provides a real time Graphical User Interface (GUI) of the measurements. It allows to calibrate the VNA and to store the output measurements as Excel readable files.

3.1.2 Setup & Output

One of the most important steps needed before performing any measurement is the **calibration**. It is a simple procedure used to calibrate the hardware and to validate the results. After the miniVNA is connected and the proper model is chosen among the available ones, different loads are used to ensure that the readings fall within the set limits. Calibration kit used for this operation is reported in Figure 3.3. In particular, this process is used to guarantee that the effects of each instrument (cables, connectors, etc.) are canceled and consequently not considered in the reading.

Figure 3.3: Calibration kit used for the miniVNA

The output of these measurements is an `.xlsx` file in which, for a given range of frequencies, a different parameter is represented. Since the measurement settings changed for different patients, influencing the number of samples and the distances between each value, 500 frequencies are considered between 2 and 3 [GHz] range. A more detailed description will be given in the Section 3.1.3, explaining how the processing was performed.

The parameters extracted from the vnaJ software are:

I **Returnloss (dB)**: $S_{1,1}$ or $S_{2,2}$ parameter, i.e. the proportion of a signal that is reflected as a result of an impedance mismatch;

$$RL(f) = 10 * \log \frac{P_{forward}}{P_{reflected}} \tag{3.1}$$

II **Returnphase (°)**: Reflection coefficient angle;

III **Transmissionloss (dB)**: $S_{2,1}$ or $S_{1,2}$ parameter, i.e. the proportion of a signal that is transmitted as a result of an impedance mismatch;

IV **Transmissionphase (°)**: Transmission coefficient angle;

V **Rs (Ohm)**: Series equivalent resistance of the load;

VI **Xs (Ohm)**: Series equivalent reactance of the load;

VII **|Z| (Ohm)**: Magnitude of the complex impedance;

VIII **Magnitude**;

IX **Rho real**: Real part of reflection coefficient;

X **Rho imag**: Imaginary part of reflection coefficient;

XI **SWR**: Ratio of maximum and minimum voltage amplitudes of the standing wave;

XII **Theta**: Impedance angle;

XIII **GroupDelay (nS)**: Time delay of the amplitude envelopes of the various sinusoidal components of a signal through a DUT;

Concerning the scope of the project, the parameter of interest is the Return Loss. In order to perform comparison with different measurements and to organize the available data an accurate pre-processing was needed, in particular to extract the useful column from the file and to avoid iterating in different folders and sub-folders each time a further processing is needed. Again, a detailed description of this process is present in Section 3.1.3.

The available data were organized in a tree-layered structure: the main folder `"Lymph Oedema Project"` contains a sub-folder for each patient, whose name is a code in the format `MV0xx`, where `xx` is a number referred uniquely to the patient; this format is valid if the affected limb is the arm, otherwise patient's name ends

with `" - leg"`. Inside each patient's folder, other sub-folders may be present depending on if and when the measurements were taken: six names are available, as described by Table 3.1, and refers to the measurement date with respect to the operation of the patient.

Name	Description
`preOP`	Before the operation
`postOP`	After the operation
`1 month`	1 month after the operation
`6 months`	6 months after the operation
`1 year`	1 year after the operation

Table 3.1: Description of the sub-folder names at the date layer

Finally, in each of these sub-folders three measurements for each body part are taken and are distinguished with the ending number of the name of the file. The number, ranging between 1 and 18 (some measurements of free-space were taken, ending with a `"_0"`, but these are not considered) indicates the position of the measurement: for each body part three files were extracted in order to reduce the error, foreseeing a future averaging among them. In total, six body parts are examined, three on the left and three on the right as show in Table 3.2.

Number	Arm	Leg
1 to 3	Left wrist	Left foot
4 to 6	Left elbow	Left below patella
7 to 9	Left shoulder	Left above patella
10 to 12	Right wrist	Right foot
13 to 15	Right elbow	Right below patella
16 to 18	Right shoulder	Right above patella

Table 3.2: Number relationship with respect to the body part position

However, it must be noted that many patients were not recorded in all of these time windows and that some measurements were not taken because, as it happened in few cases, the operation was too fresh and it was preferable to avoid the application of the sensor on the operated limb. These data were then merged into a unique `.xlsx` file, taking all the RL measured so as to have an easier access to data and a faster management.

3.1.3 Data Processing

The processing of data is constituted in three phases: data preparation, standardization and filtering. Each of them is oriented on tackling different problems: primarily the complex disposition of data, secondly the mismatching scales of data, i.e. different frequency sampling, and the presence of outliers and/or invalid data.

The tools used to perform the processing are essentially **Python** and **Excel**: while the first one has been used to perform the core work, i.e. read and manage data, the second is used to cross-check that the processing was performing correctly.

Preparation

This phase was focused on handling different folders and reading all files in it, uniquely associating them to patient, date and body part. To do so, a Python script has been written, allowing to export the data into an Excel file; consequently, Excel has been used to confirm that the operation was performed as intended.

Given the heterogeneity of the dataset and the presence of metadata in the folders, a fundamental step done in this stage was the re-organization of the folders' and files' names. To iterate all the files maintaining only useful information, a uniform folder nomenclature was applied according to the following rules:

1. If the LO affected a leg, `" - leg"` was added to the name of the patient's folder; it must be noted that the metadata present in the original folders were not sufficient to acknowledge this information and that it was necessary to retrieve it directly from the hospital;

2. The folder containing the measurement can only have names reported in Table 3.1, relative to the date from the operation; if a different name is used, data are not considered;

3. All measurements need to have the same format name:
   `"VNA_YYMMDD_hhmmss_X.xls"`
   where: `YYMMDD` is the day format, from year to day; `hhmmss` is the hour format, from hour to seconds; `X`, is the already introduced number linked to the body part of the patient.

Finally, information about available measurements and affected limb was collected (both from the hospital and from metadata contained in the folders) and merged into an Excel file named `"Patient_status.xlsx"`.

Standardization

The main focus of this stage was merging all data into unique files, maintaining as much information as possible. From the previous stage it has been noticed that the frequency sampling considered was different in some measurements and that the main parameter is the RL evaluated in Decibel. In particular, the important information about the RL curve were found between 2 and 3 GHz frequency range. For this reason, it has been decided to standardize using 500 samples with a step of 2 MHz: if more values were found in same 2 MHz range, the mean value of them was used; otherwise, if any value could not be found, a blank space was left.

As already mentioned, to perform all these operations Python was used; in particular, Excel files were read using **pandas** library, which allowed to exploit DataFrame structures. This library also provides very useful methods that allowed to manage data easily: performing operations for each row, appending new columns to the DataFrame or evaluating statistical parameters, such as mean or standard deviation, are examples of operations performed in this phase. Among the available methods, one in particular has been used in the Python script: `interpolate()`, which allows to fulfill missing values (either blank spaces or NaNs); the inputs needed for this operation are the type of interpolation and, eventually, its order. The `"polynomial"` method of order 3 has been chosen, noticing that only negligible differences were found if many and many data were missing (behaviour that occurred rarely, especially for corrupted data that will be removed later on).

Basically, when a file was read, a temporary DataFrame was used to save into a different column the RL readings; when all three readings linked to the body part were present, the mean value for each frequency was evaluated and saved into another common DataFrame's column where its name had the following format:

`"Patient_name+position+date"`

In this way it was possible to compare the collected values and the ones composing the original files. When all the folders were iterated, the DataFrame was saved in a file named `"Returnlosses.xlsx"`. Besides the evaluation of the mean, also the standard deviation was extracted in the same way, i.e. for each frequency sample this statistical information was estimated between three readings and saved in another common DataFrame; in this case the name chosen was `"StDev-Returnlosses.xlsx"`. Finally, to have some parameters linked to data reliability, which could be possibly implemented in the ML part, the standard deviation of the peak amplitude and frequency was evaluated `"f&A_std-Returnlosses.xlsx"`; these two parameters could be seen as a reliability box where the peak could be found: the bigger the box, the less reliable its position.

Filtering

From the confidence-box and the cross-checking procedure performed, it was possible to notice that some peaks were very unreliable and that a filtering process was especially needed. To do so, the previous Python script was enhanced with a filtering function: before averaging the three readings and saving the result into one DataFrame, the difference for each pair of curves was evaluated; the parameter used for this estimation is the absolute value of the difference of areas, normalized with respect to the area of the mean, as reported in Equation (3.2) by parameter $D_{(i,j)}$:

$$D_{(i,j)} = \frac{A_i - A_j}{A_{mean(i,j)}}[\%] \tag{3.2}$$

where: $(i, j) \subset [a, b, c]$, being $[a, b, c]$ the three readings relative to same body part; A_i is the area of measurement i; $D_{(i,j)}$ is the percentage difference of the two readings (i, j).

Besides this parameter, the filtering process takes into consideration the amount of zeros in the averaged curve $f_{(i,j)}[n]$ (mean between the couple (i, j)), a behaviour that could be seen as a sort of cutting off the graph; few examples of this behaviour will be shown in the Section 3.1.5. Finally, for each body part measurement $f[n]$, a different scenario is determined:

1. If the following condition is met

$$\mathbf{D_{(i,j)} > 5}[\%], \qquad \forall(\mathbf{i}, \mathbf{j}) \subset [\mathbf{a}, \mathbf{b}, \mathbf{c}] \tag{3.3}$$

   measurement $f[n]$ is discarded: this option has been evaluated following a trial and error procedure and considering that most of the percentages were below the 5% threshold; it allowed to distinguish between inconsistent and valid measurements.

2. If the following condition is met

$$\mathbf{1.3 \cdot D_{(i,j)} < D_{(i,k)}} \quad \& \quad \mathbf{1.3 \cdot D_{(i,j)} < D_{(k,j)}}, \qquad \forall(\mathbf{i}, \mathbf{j}) \subset [\mathbf{a}, \mathbf{b}, \mathbf{c}] \tag{3.4}$$

   reading k is discarded: this is due to the higher difference between the curves in which reading k is present with respect to the curve only composed by the pair (i, j); in this condition the threshold chosen is that the difference between (i, k) and (k, j) must be smaller than the one of (i, j) increased by 30%. This condition leads to measurement $f[n]$ being equal to $f_{(i,j)}[n]$.

3. If the following condition is met

$$\mathbf{50} \leq \sum_{\mathbf{n=0}}^{\mathbf{500}} \mathbf{f[n]} = \mathbf{0} \leq \mathbf{125} \quad \& \quad \mathbf{min(f[n])} \geq \mathbf{-15}, \qquad \forall \mathbf{f[n]} \tag{3.5}$$

measurement $f[n]$ **is discarded**: in this case the curve considered, $f[n]$, is the average of the three readings; if it has a number of data points equal to zero in between 10% and 25% of the total, respectively 50 and 125 out of 500, but it does not have a clear peak visible, i.e. the peak value smaller than -15 dB, it is considered to be corrupted. However, if the amount of zeros is in that range but the peak is clear, data is kept and could be further modified by the previous conditions 3.3 and 3.4.

4. If none of the above conditions are met **measurement** $f[n]$ **is kept** and it will be evaluated as the mean of the three readings $f_m[n]$.

From this procedure it was possible to obtain the final database, in which corrupted or inconsistent data were filtered out. These measurements and their relative statistics were saved in `.xlsx` files as done before, but with different names: `"Returnlosses-filtered.xlsx"`, `"StDev-Returnlosses-filtered.xlsx"` and `"f&A_std-Returnlosses-filtered.xlsx"`. The idea of creating new files instead of overwriting the previous ones was due to have the possibility of comparing before and after filtering data, as shown in Section 3.1.5.

To evaluate if any characteristical behaviour could be linked to the the body position and the date in which the measurement was taken, other additional datasets were extracted: `"XXXXX-mean_values.xlsx"` is the generic name of the file containing RL curves, where `XXXXX` is the name of the body part, e.g. "foot" or "elbow". The information on being a left/right measurement was incorporated in the dataset by distinguishing between affected and reference limbs. Each of these datasets contains fifteen columns, named after the possible date names shown in Table 3.1; three columns are present for each date: one refers to the reference limb (where column name is ending with `" - ref"`); another is related to affected limb (ending with `" - aff"`); last one is the mean evaluated not taking into consideration if it was affected or not (without any termination, just the date).

For the Machine Learning classification part it was needed to extract information from the RL curves and merge it with available metadata. In particular, the information related to the curve was shrank into three parameters, used to characterize each curve: position and amplitude of the resonant peak and bandwidth evaluated at -10 dB.

#	POS	DATE	AFFECTED	POS_NAME	NAME
1	foot	1 year	TRUE	Left foot	MV001
2	below patella	1 year	TRUE	Left below patella	MV001
...	...	...	...	...	...
216	shoulder	postop	TRUE	Left shoulder	MV021
217	elbow	postop	FALSE	Right elbow	MV021
218	shoulder	postop	FALSE	Right shoulder	MV021

Table 3.3: Example of the final dataset containing categorical features and metadata

#	BMI	MIN	MIN IDX	BW[1]	AMP_STD	FRQ_STD
1	20	-17,4392	2.53E+09	64000000	0,4196	9018499
2	20	-12,7318	2.51E+09	52000000	0,5587	10583005
...	...	...	...	...	...	...
216	29	-35.2324	2.43E+09	92000000	2,1554	0
217	29	-29.4759	2.44E+09	1.1E+08	1,5707	0
218	29	-3.55654	2.55E+09	0	0,0252	65817930

Table 3.4: Example of the final dataset containing continuous features

The final dataset is shown in Table 3.3 and Table 3.4[1] where each column is referring to a different feature that can either be categorical, binary or continuous. As mentioned before, metadata were merged with information related to the curves in order to have a completely descriptive summary of the information about the readings. It follows an explanation of the features:

1. **POS**: *categorical*; position of the measurement, without the left/right information. Since only three position are studied but the body portion can either be upper or lower, six values are available;
2. **BMI**: *continuous*; Body Mass Index is the information related to patient's weight status. It must be noted that different values can be found for same patient since it may have changed weight from one measurement to another;
3. **DATE**: *categorical*; as largely described before, this categorical feature describes the temporal distance between measurement and operation. Five values are available;

[1] **BW** short form is used to fit in the page, the name used in the file is **BANDWIDTH**

4. **MIN**: *continuous*; the minimum value found on the RL curve, ranging between 0 and -50 [dB] (most of the values are found between -5 and -30 [dB]);
5. **MIN IDX**: *continuous*; this is the frequency at which the minimum value is found, in other words where the peak is located, ranging between 2 and 3 [GHz] (most of the values are found between 2.3 and 2.6 [GHz]);
6. **BANDWIDTH**: *continuous*; this is the frequency distance found at -10 [dB], expressed as a positive number ranging between 0 and 0.7 [GHz] (0 is used if the minimum value is not reaching -10 [dB]);
7. **AMP_STD**: *continuous*; standard deviation of the peak's amplitude evaluated between the three (or two, depending on the filtering process) readings;
8. **FRQ_STD**: *continuous*; standard deviation of the peak's frequency evaluated between the three (or two, depending on the filtering process) readings; together with the previous parameter it is deployed to give an evaluation on the reliability of the data point;
9. **AFFECTED**: *binary*; feature related to the status of the measured limb; can either be True or False;
10. **NAME & POS_NAME**: *metadata*; these columns are used to correlate the measurement to the relative RL curve found in another file; this information is not used for the data analysis, but for checking that the correlation is correct;

3.1.4 Validation Procedure

One of the most important step after the processing phase was the data validation, aimed at understanding problems affecting the available data, possibly to avoid or correct them in future, in case more measurements will be collected. As already discussed, one very unexpected behavior was found in the shape of some curves, presenting sometimes multiple and/or not-clear peaks. Very ambiguous and recurring behaviours are listed below:

- the curve was not reaching values lower than -10 dB;
- the curve looked like a cut-off graph, resulting in zeros because no positive values are allowed;
- the curve was oscillating producing many peaks and local minima, instead of a unique minimum point;

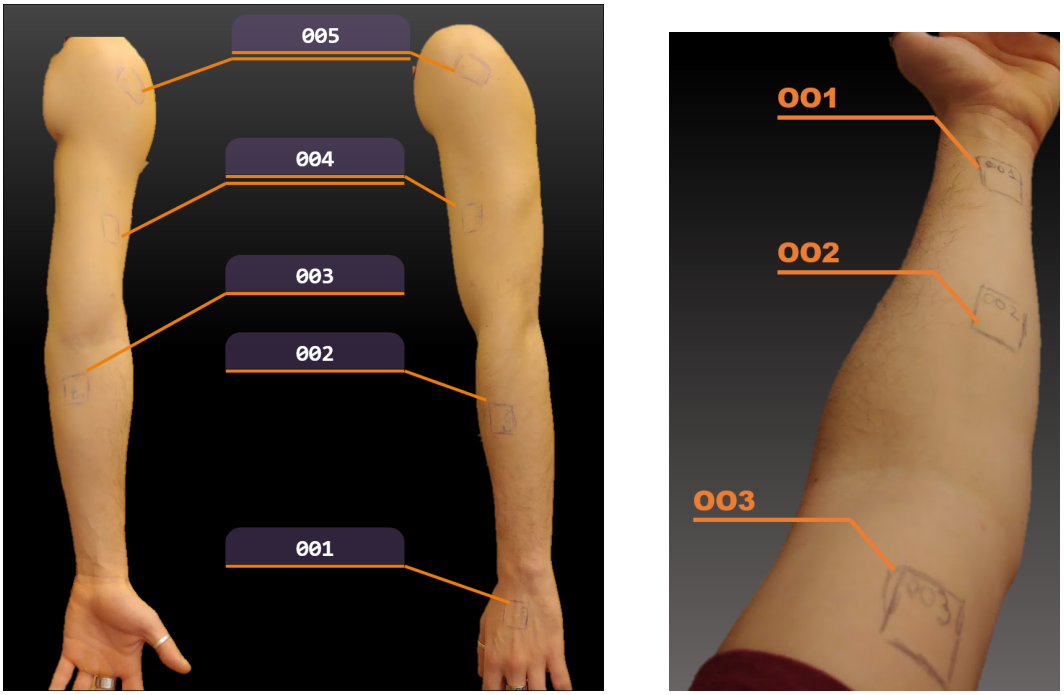
The investigation started from checking if the instruments were properly working: SRR antenna, copper cable and the miniVNA were tested in different ways.

SRR antenna

To test the SRR antenna, the miniVNA measurements were compared to the ones obtained from the FieldFox Vector Network Analyzer. Five different spots on the left arm plus one in the air were selected and, for each of them, three different readings were taken and then averaged. The spots chosen are shown in Figure 3.4a, where they are distinguished by a number. Readings are saved in the format:

`"XNNN_Y.Z`

where: `X` can either be "A" or "B", respectively being miniVNA or FieldFox measurement; `NNN` is a three digit number, referring to the position in which the measurement was taken; `Y` is an integer number ranging from 1 to 3, referring to one of the three readings per measurement; `Z` is the data format, being `".xlsx"` and `".csv"` the format files extracted respectively from the miniVNA and the FieldFox.



(a) Spots SRR antenna test **(b)** Spots temperature test

Figure 3.4: Spots chosen for the readings used for testing the SRR antenna (a) and the effects of temperature (b)

Copper cable, miniVNA and calibration procedure

Besides the antenna, also the cable and the miniVNA used for the measurements in the hospital needed to be checked. In order to test them, another miniVNA and another cable were used to perform the same measurement on the same spot of the skin, where three readings were taken for each scenario. By performing this test, since the calibration was included in this process, two different techniques were also tested: performing the calibration with and without the cable attached to the DUT port. It must be noted that both this last method and the second cable used are known to output wrong measurements and the attempt was done with the purpose of evaluating if same erroneous behaviour could be found in the dataset. The names related to the tests are in the format:

`"X_AAAA_BBBB.xlsx"`

where: `X` refers to the miniVNA used, being `A` and `B` respectively for the one used in hospital and the another available miniVNA; `AAAA` refers to the cable condition and can be `GOOD` or `BROK` (standing for broken); `BBBB` is the calibration technique, where `with` means that the calibration was performed with the cable connecting the DUT port to the calibration kit, while `wout` refers to the situation in which the calibration kit was directly connected to the DUT port.

Temperature influence

Another element that was considered to be affecting negatively the dataset was the temperature. In particular, to assess whether it affected the readings or not few tests were performed investigating three spots on the arm, as shown in Figure 3.4b. The methodology followed in this part consisted in leaving the sensor on the desk for a couple of minutes to reach ambient temperature each time a reading was extracted for all the mentioned spots. Later on, to increase the temperature and simulate effects of a prolonged contact with skin, the readings were taken only after the sensor was kept in contact with skin for at least one minute. To simulate the effect of cold, some snow was collected from outside and placed in a plastic container; the sensor was then carefully laid on the external surface of the plastic bag (which was cleaned and dried to avoid presence of water interfere with the test) and kept there for a couple of minutes after every measurement. To prove that the temperature was effectively different, a laser thermometer was deployed: the minimum temperature reached was around 10°C, highly different from the 20°C and 24°C reached respectively with ambient and warm temperatures. The name format of each readings is the following:

`"X_NNN_Y.xlsx"`

where: `X` refers to the temperature of the sensor, being `C`, `A` and `W` used for referring to "cold", "ambient" and "warm"; `NNN` is the number of the spot location; `Y` can either be `a`, `b` or `c` and is used to differentiate the three readings. 7

3.1.5 Results

Dataset characterization

Data are collected exporting the readings of a miniVNA, done thanks to vnaJ software. These readings are re-organized in folders with a specific hierarchical order, as largely described in Section 3.1.3.

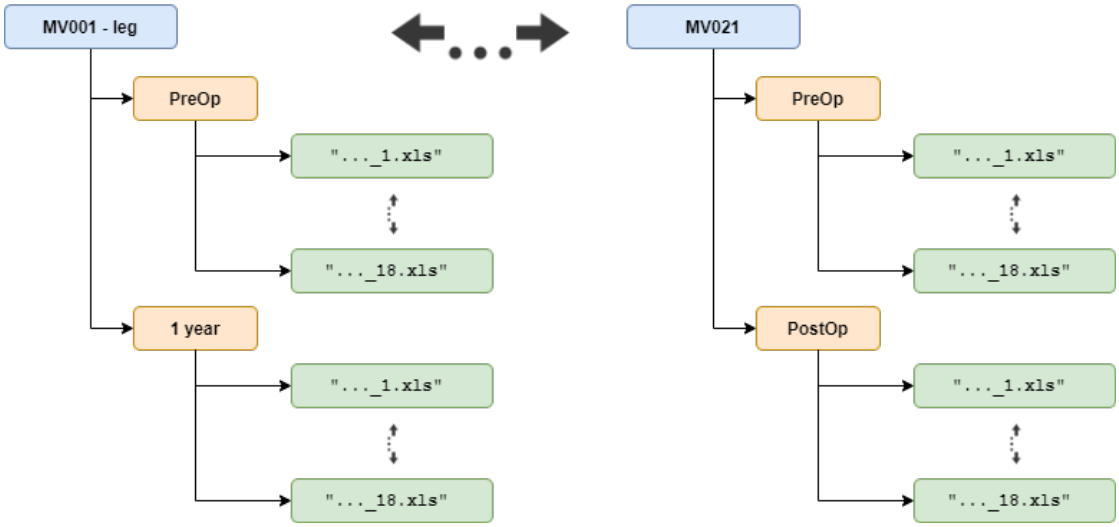


Figure 3.5: Folder disposition

The diagram found in Figure 3.5 allows to understand how the folders and the relative files are organized. On the top layer the names of the patient is found: in each sub-folder, indicating the date with respect to the operation, the measurements are found. Each file ends with a characteristic number that marks the position of the reading.

Data are then processed in such a way to extract needed information: the frequency range of interest is located between 2 and 3 [GHz], considering 500 samples with a step of 2 [MHz]. Among the parameters available from the reading, the one chosen for the work is the RL. Furthermore, a filtering process has been applied so as to remove misleading measurements. In addition to the filtering process, since holes may have been left by the sampling, interpolation has been performed.

As a result of this filtering and interpolation process, a file containing all the valid RL samples is created. An example of how it is composed is reported in Table 3.5: the table is composed of 218 columns $\times$ 500 rows, where each column is

	MV001 - leg +Left foot +1 year	MV001 - leg +Left below patella +1 year	...	MV021 +Right shoulder +postop
2E+09	0	-0.03278	...	0
2E+09	0	-0.03068	...	-4.6E-07
...	...	...	...	...
3E+09	-0.07763	-0.37332	...	-1.23639
3E+09	-0.21433	-0.45111	...	-1.32082

Table 3.5: Example of the `Interpolated-Returnlosses.xlsx` file

related to a different RL reading, while rows refer to a different frequency. The column name allows to understand where the measurement is located in the folders in a human-readable way, while the row index indicates the frequency. Histograms characterizing the dataset are reported in Figure 3.7.

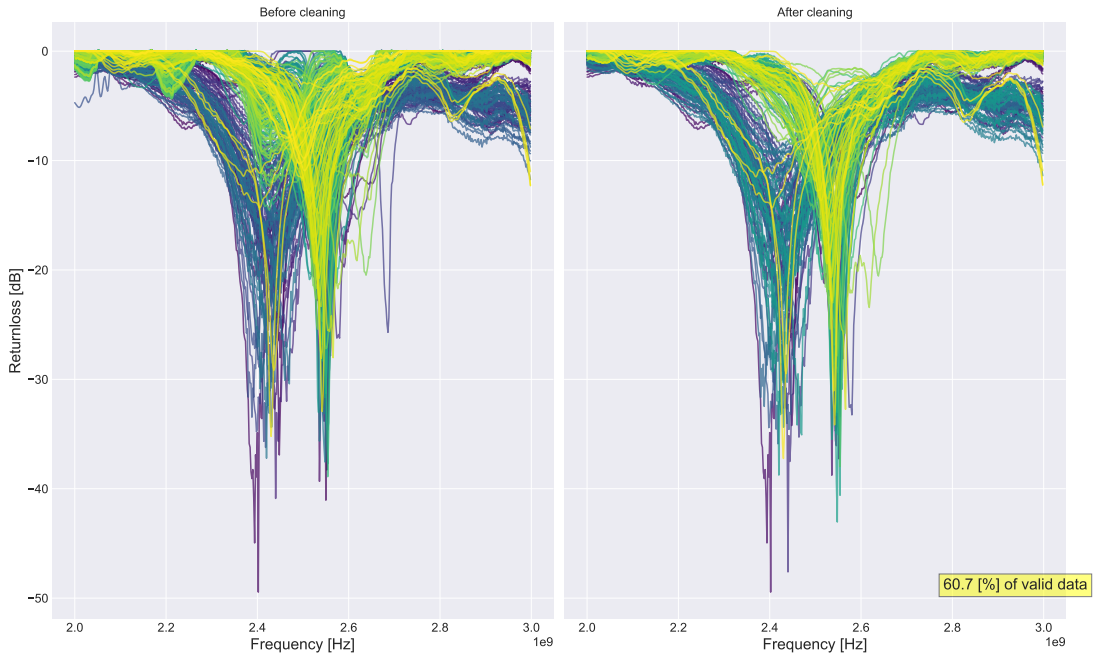
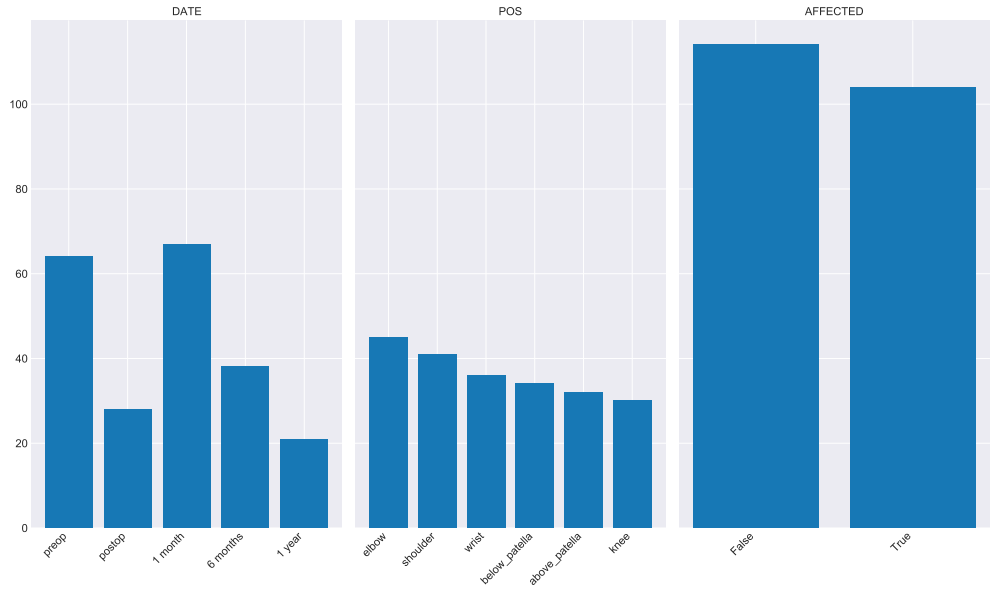
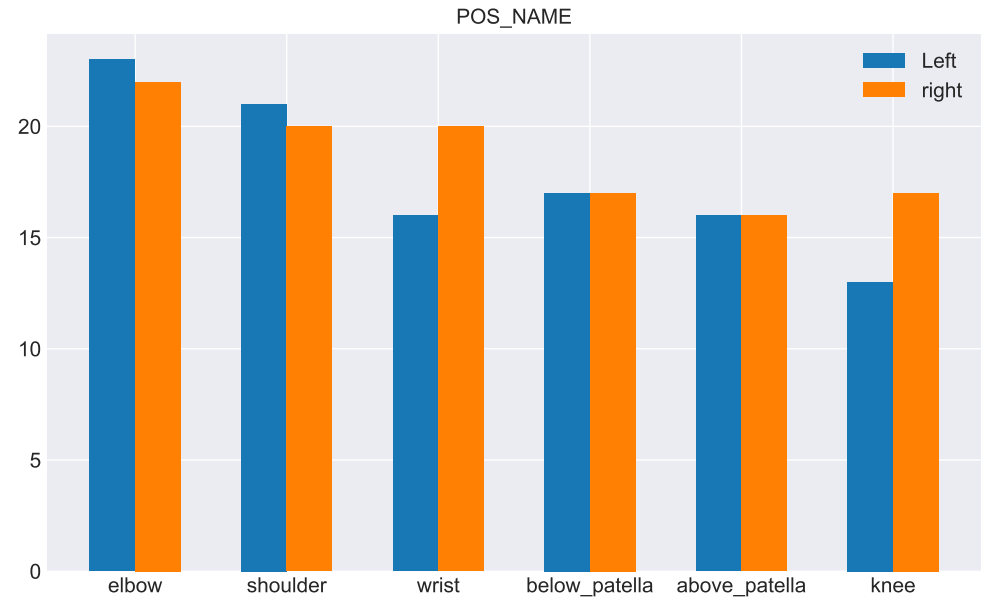


Figure 3.6: Result of the filtering process

The filtering process allowed to preserve 60.7 % of the total data: Figure 3.6 shows the cloud of curves before and after the filtering process. It can be noticed that the region 2.4 and 2.6 GHz and between 0 and -10 dB is clearer, as a consequence of the fact that most eliminated curves are located in that region.



(a) Histogram characterizing the dataset, considering body position, date and presence of disease



(b) Histogram highlighting differences for body parts, comparing affected and reference limbs

Figure 3.7: Histograms to characterize the dataset, looking at *DATE*, *POS* and *AFFECTED* categories (a) and *POS* where left and right are distinguished (b).

Filtering process

To better understand the reason for which a filtering process was performed and how data were considered valid, in this sub-section a description of filtering result is reported. One first fundamental step was the necessity of discovering the reliability of three readings, how much they were alike and if any problems occurred: to do so, the standard deviation for each frequency of the three readings was evaluated. The result of this process is a blue curve representing the mean value of the three readings and two red curves representing upper and lower bound estimated as the mean plus/minus the standard deviation evaluated at each frequency. An example is reported in Figure 3.8, where measurement related to the left above patella after 1 year from the operation of patient MV006 is shown. It is also interesting to notice that on the right of the figure the uncertainty region is shown: this represents the reliability of this peak (the bigger it is, the less reliable it is); its height is determined by the standard deviation evaluated for the three peak amplitudes, the width is calculated as the standard deviation between the peak location of the three readings. As it can be noticed by this figure, the uncertainty region is quite extended, especially for the location in frequency. In order to take into consideration only reliable measurements for performing the averaging, results of the filtering mentioned in Section 3.1.3 are reported here.

In figure Figure 3.9 all three readings used for evaluating mean and standard deviation of Figure 3.8 are shown: in this case, the measurement is discarded because the three curves are very dissimilar. Another criterion used to discard all readings related to one measurement is being corrupted, i.e. having too many zeros and missing a clear peak on the curve; an example is reported in Figure 3.10. As already mentioned, there is a possibility that two out of three readings are valid: in this case the idea is to retain the two valid readings, considering the third one as an outlier. This happens in Figure 3.11, where it is possible to notice that one reading, the number 15 denoted by a green curve, has a different peak; moreover, on the bottom right corner, the area-ratio evaluated between each pair of curves is shown: since the value estimated between curves 13 and 14 is smaller than the ones evaluated with reading 15, curve 15 is dropped but not the measurement, which will be now composed of only two readings.

To summarize the effects of the filtering process Table 3.6 is reported: each row represent the way different measurements are classified. Besides the previously described "Inconsistent", "Corrupted" and "1-removed" labels, "Clear peaks" refers to those measurements that were considered to be valid even if the number of zeros is in between 50 and 125, but the peak has an amplitude smaller than -15 dB; nevertheless, these data may be further filtered.

Status	Total amount	Discarded
Inconsistent	7	yes
Corrupted	134	yes
Clear peak	32	no
1-removed	193	no
Approved	25	no

Table 3.6: Distribution of curves, describing if they are discarded or not

.

Figure 3.8: Standard deviation of measurement related to the left above patella after 1 year from the operation of patient MV006.

Figure 3.9: Three readings related to the left above patella after 1 year from the operation of patient MV006; measurement discarded.

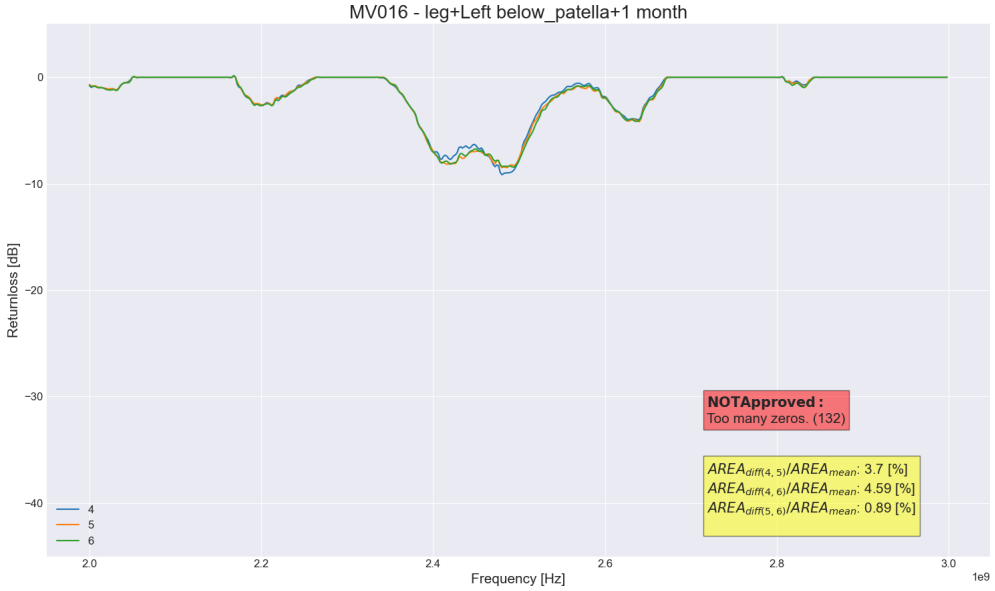


Figure 3.10: Three readings related to the left below patella after 1 month from the operation of patient MV016; measurement discarded.

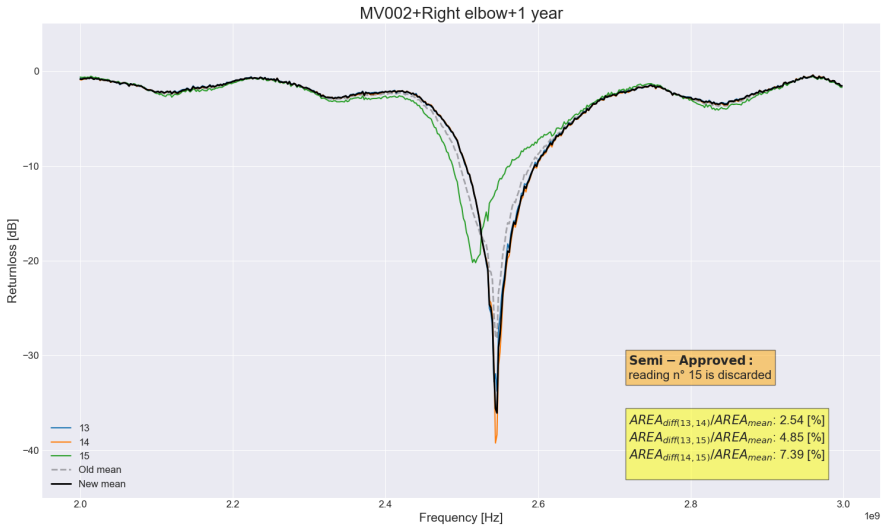


Figure 3.11: Three readings related to the right elbow after 1 year from the operation of patient MV002; 1 reading discarded.

Experiments

As largely described in the data validation sub-section, three experiments were performed. The first one, responsible for testing the SRR antenna, in which the FieldFox and miniVNA readings were compared, shows no major differences: the peak is located essentially in the same region, indicating that the sensor is operating well and it is not affected by any problem. In Figure 3.12 main discrepancies can be detected for spot 001; smaller amplitude gaps are present for 002, 004 and 005; the problem could either be associated with the physical instrument or a slightly different positioning of the same on the skin. In any case, since differences between the two hardwares are negligible, it is reasonable to consider that the SRR antenna do not affect negatively measurements.

For the second experiment the cable used for latest measurements was tested; in parallel, the calibration process and another miniVNA were tested to detect if any other problem could be highlighted. In Figure 3.13 it is possible to see how measurements are notably irregular: each of them has a different peak location and amplitude, considering also that each trend is inconsistent with the ones seen before. It is also possible to notice another erroneous behaviour met before: both the green and blue curves appear to be cut off, having many points equal to zero; as mentioned earlier, this is a quite frequent behaviour happening in the dataset and, possibly, by simply substituting the cable this should not be present anymore.

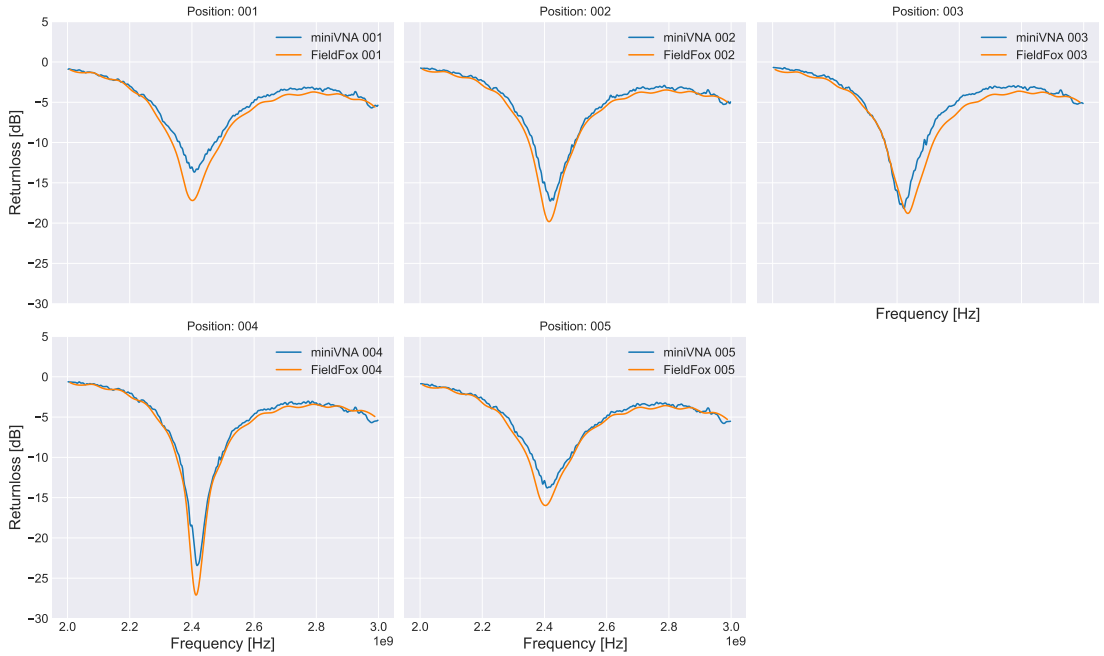


Figure 3.12: Experimental measurements taken with both the miniVNA and the FieldFox Analyzer, for each of the spot considered.

As a matter of facts, in the second part of the experiment a new cable was used: in this case, shown in Figure 3.14, from blue and green curves, the ones associated with a correct calibration procedure, it is possible to see the expected behaviour, i.e. a clear peak with no major differences between the two miniVNA used. On the other hand, from the orange and red curves, it is possible to notice another very common behaviour already met in the dataset: the oscillation of the curves (practically the same for the two miniVNA) can now be linked with an improper calibration, performed using the calibration kit directly attached at the DUT port, instead of connecting the copper cable between the two (clearly this is valid only if a cable is being deployed for the measurements).

Interestingly, by looking at the standard deviation, evaluated as explained before, i.e. between the three readings related to the same spot, it is possible to note that if the cable is broken the uncertainty is significant; the standard deviation is very small if the cable is in good condition, either if the calibration process is correct or not. This is reported in Figure 3.15.

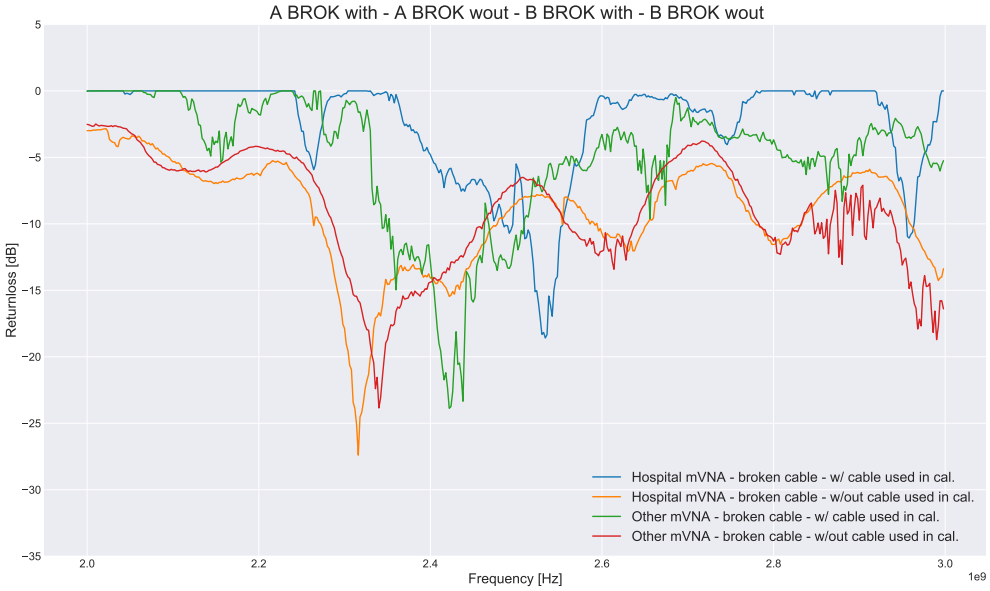


Figure 3.13: Experimental measurements taken with both two different miniVNA and performing two types of calibration, testing same cable used for latest measurements.

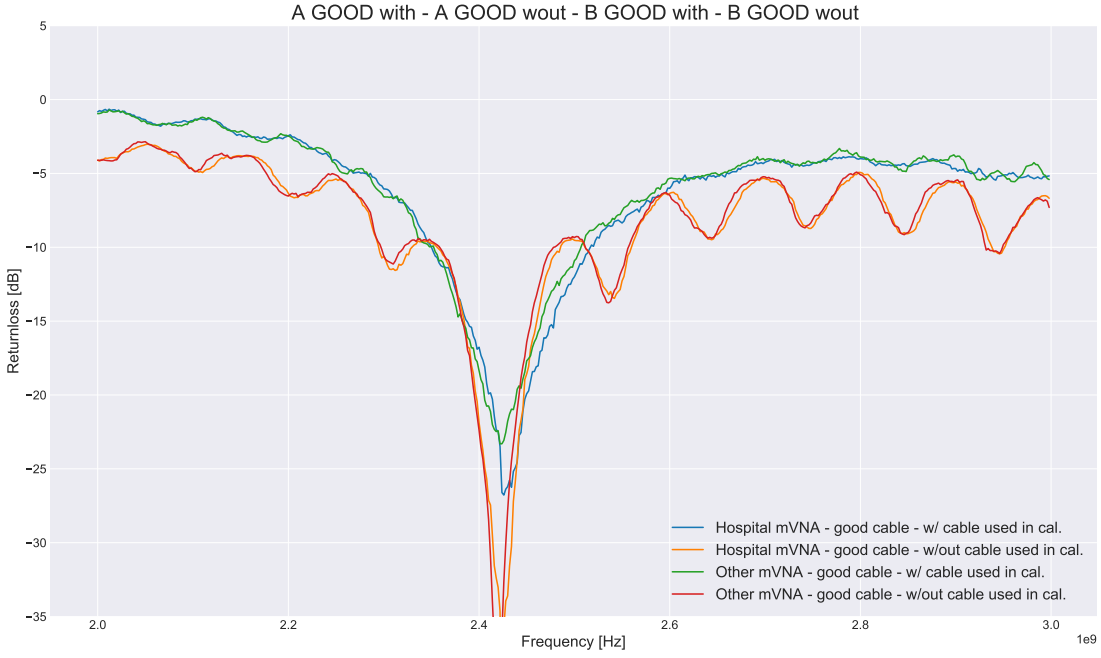
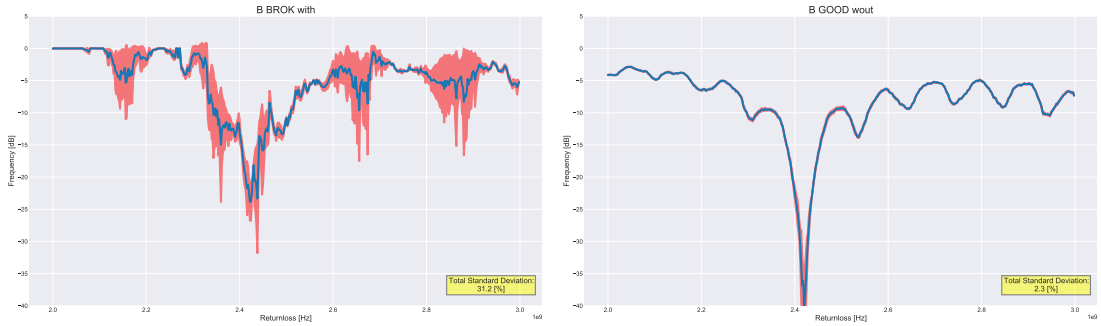


Figure 3.14: Experimental measurements taken with both two different miniVNA and performing two types of calibration, testing the calibration procedure.



(a) Broken cable and correct calibration **(b)** Good cable and incorrect calibration

Figure 3.15: Mean and standard deviation between three readings performed with a broken cable and with a correct calibration procedure (a) and with an incorrect calibration but with a functioning cable (b).

The last experiment was performed to test the influence of temperature on sensor and, indirectly, on measurements. The fear was that, if sensors were left in contact with skin for a prolonged period and changed their temperature, especially in winter when ambient temperature is low, this could have affected the reading, leading to a different curve. Results of this experiment are shown in Figure 3.16, where three different curves for each spot are represented: in here it is possible to acknowledge that the peaks are rather similar, especially for the first two spots (001 and 002); concerning 003 spot, warm temperature measurement is moderately different, but this could be related to a minor misplacement of the sensor (with respect to the desired position) or to a problem occurred during the reading, like the sensor that was not perfectly attached to the skin: these problems may have been present in the dataset as well and must be taken in consideration for a future data collection.

However, to remark that the effect of temperature is marginal, Figure 3.17 shows the standard deviation evaluated on the 003 spot, the one having bigger differences: as it can be noticed by the larger amount of red color, corresponding to a bigger uncertainty region, this measurement should not be taken into account and it is possible to consider minimal the effect of the temperature.

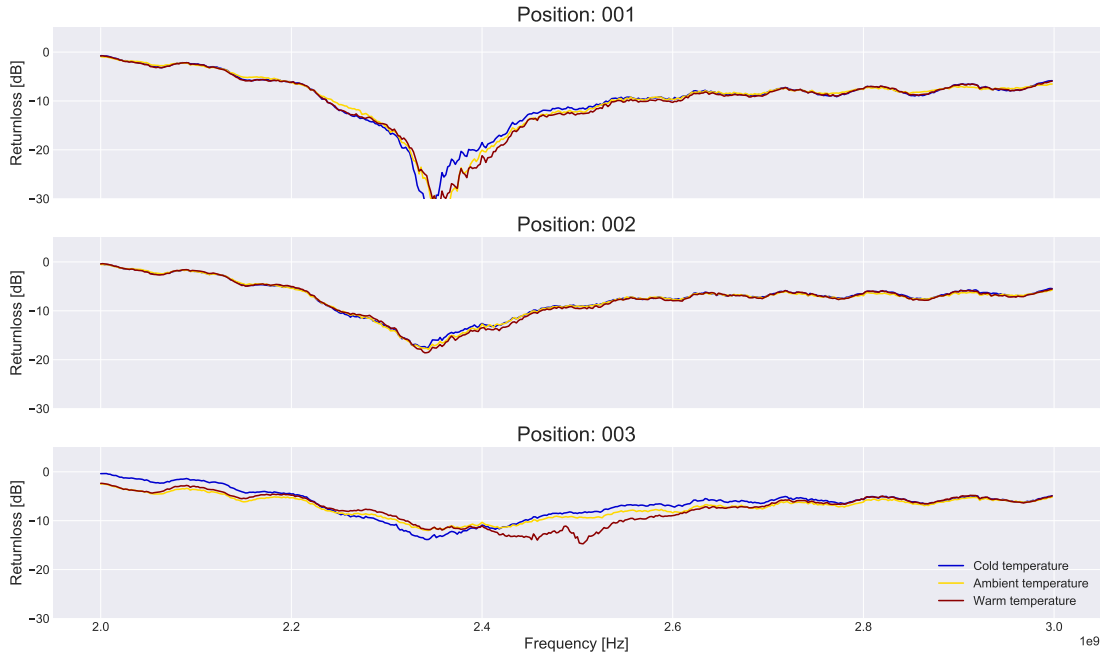
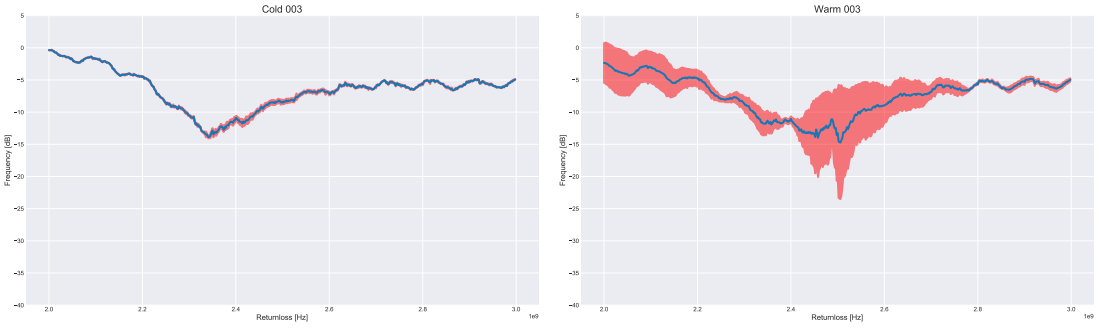


Figure 3.16: Experimental measurements to test the effects of temperature on the sensor.



(a) Cold temperature **(b)** Warm temperature

Figure 3.17: Mean and standard deviation between three readings performed with a sensor having cold (a) and warm (b) temperature on 003 spot.

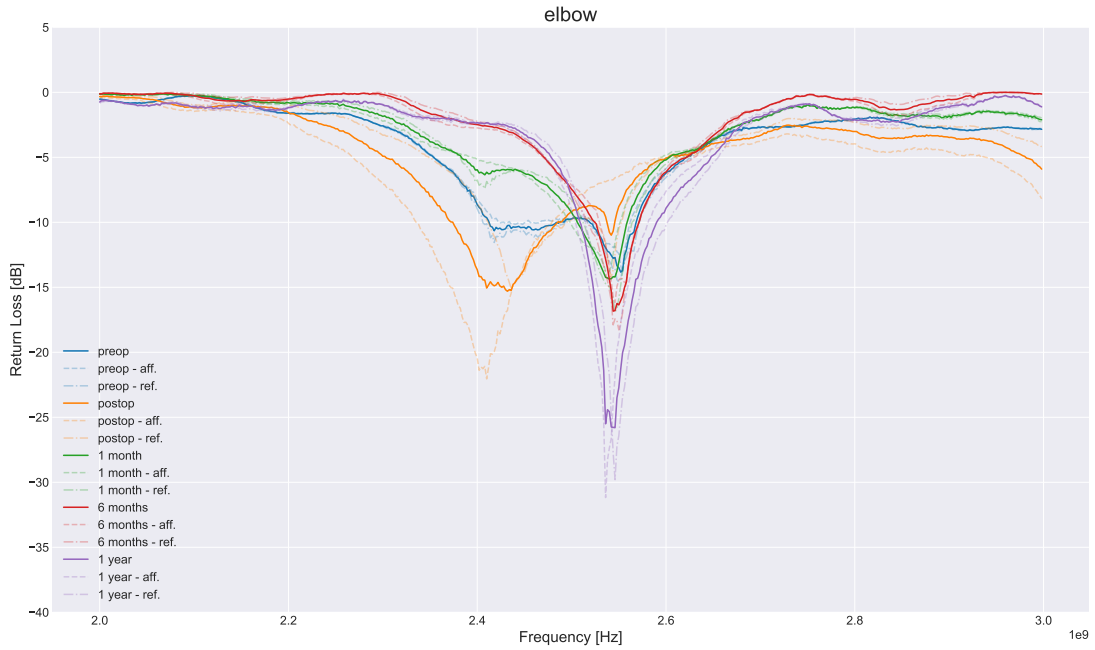
Pattern detection

As already mentioned, one major aim of this work was the detection of any pattern in the available dataset, attempting to correlate information of the measurement, such as presence of disease and location of the measure, with the RL curve. To do so, different comparison were done manually considering several factors, beginning with the evolution in time: for each body part of the measurement, an average of all the measurements for each date is evaluated, distinguishing between affected and reference. In the first place, to comprehend if any patter can be detected, for the same body part all the curves present in Figure 3.18 and of evolution through time can be seen, looking at the mean evaluated with respect to the date (not considering if it is affected or not).

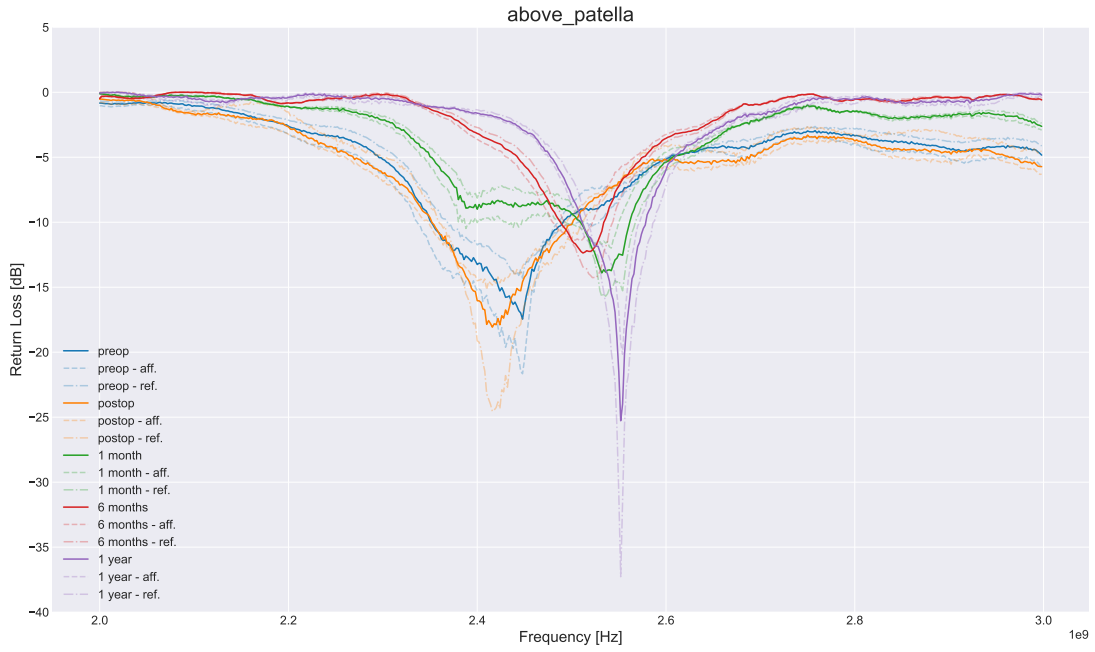
As a matter of facts, in Figure 3.18a it is reported an example in which affected elbow measurements are shown: concerning pre-operation peaks, it seems that they are quite wide-spread between 2.4 and 2.6 GHz, resulting in two local minima; the measurements after the operation highlight a clearer peak around the 2.4 GHz region; however, as the time passes by, a radical shift to the right is noted (*1 month*, *6 months* and *1 year* peaks are located after 2.5 GHz), with a constant increase in the amplitude. In order to confirm this thesis, another body part is considered.

To perform the operation previously mentioned, measurements located above the patella are investigated in a similar way: in here it can be noticed that the ones close to the operation have similar peaks, while the others are showing a similar shift to the right (higher frequencies); nevertheless, in this case a progressive increase in the amplitude is not present (*1 month* peak is higher than *6 months* peak) and the location of the peaks is more widespread than before.

These considerations could be linked to the physical phenomena happening in the body of the patients: assuming that enough data were available for the averaging process, from the operation, an increase in the accumulation of lymph fluids could have happened; progressively, this will lead to a structural change in the layered composition of the limb. In particular, the presence of a water-like fluid in between the skin and the fat layers may result in a drastic change of the readings: having this fluid a permittivity more similar to the one of the muscle, the MW travelling from the SRR sensor will have more difficulties on reaching fat and muscle layers resulting in a shift to the right of the RL peak. To further investigate this behaviour, the average evaluated on reference measurements is compared with the one evaluated on affected ones, considering the two cases discussed before. On top of this, further investigation are carried out by simulating different scenarios and will have a section dadicated to it.



(a) Elbow measurements



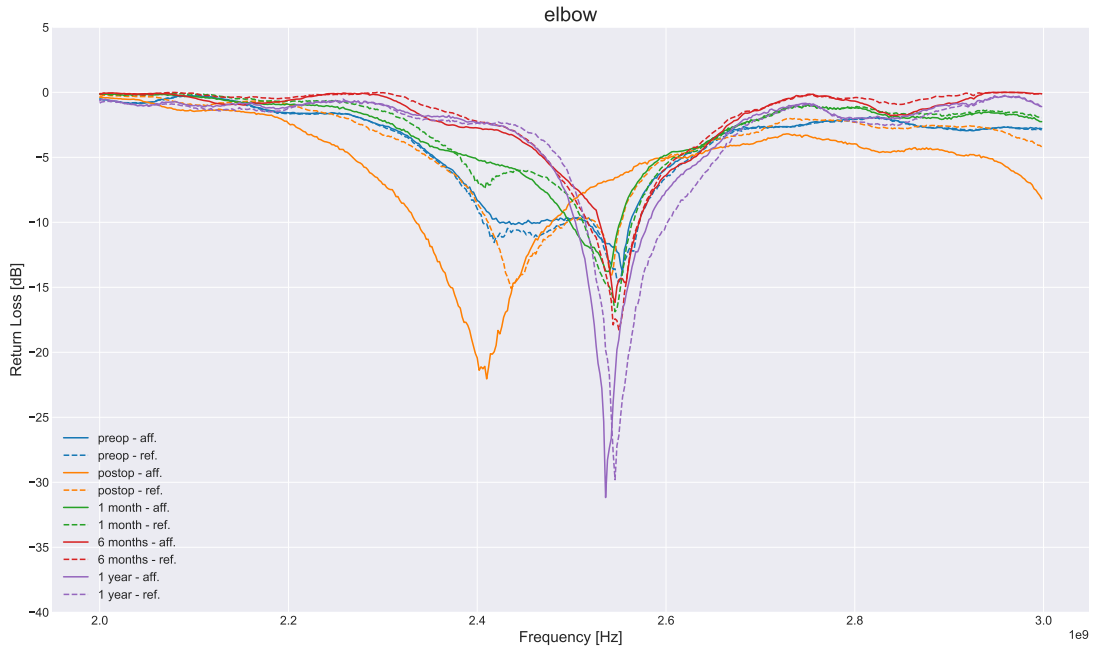
(b) Above patella measurements

Figure 3.18: (a) elbow and (b) above patella evolution in time, considering the average evaluated independently if is a reference or affected limb.

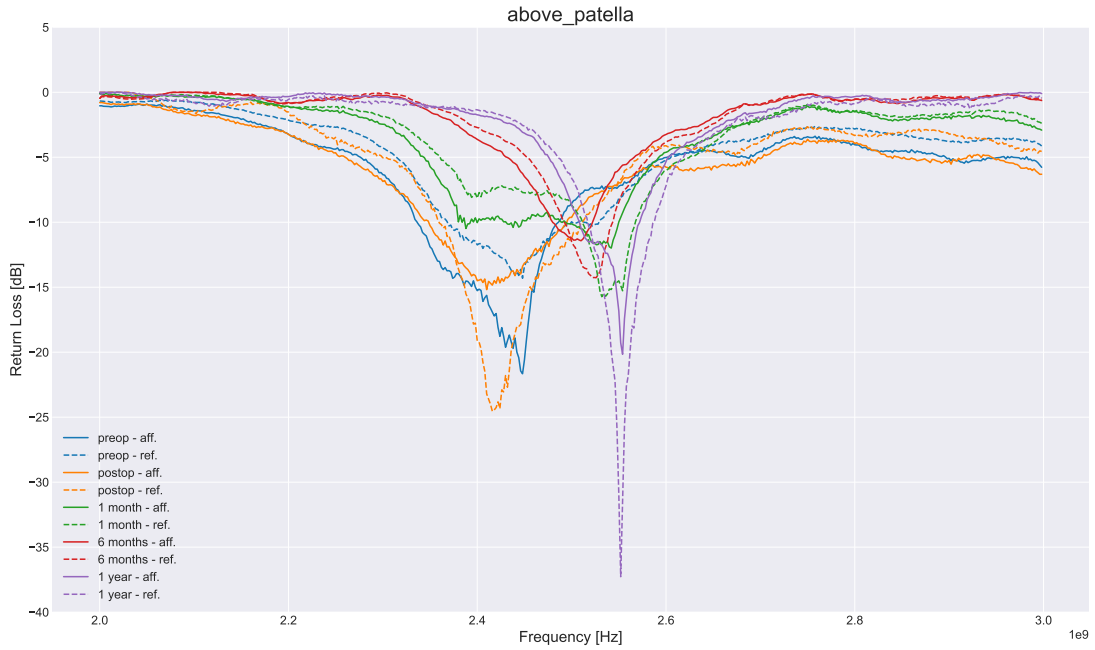
In Figure 3.19, by comparing Figure 3.19a and Figure 3.19b it is possible to focus on smaller but significant differences. Clearly, these two images are used as examples, further considerations may be done by comparing other images of body parts; here the main focus is to highlight generalized behaviours. As partly cited before, we can see that the peak related to the period before the operation is located around 2.5 GHz for elbow measurements, both for affected and reference limbs; instead, if Figure 3.19b is considered, both *preop* peaks related to average of reference and affected measurements are located near 2.4 GHz, only distinguished by their amplitude. Concerning other dates, the peaks can be located around the 2.47 region, some more some less peaked. If other images are taken in consideration, reported in Appendix A, the clear pattern of two distinguished peaks related to the date (near and far away from operation) is not anymore present and, in general, a less peaked behaviour is found. In conclusion a clear pattern could not be detected; thus, to deeply investigate on the nature of each curve, each patients' measurements are analyzed separately.

To estimate the effects of LO on the measurements another attempt is done comparing, for each patient, reference and affected limb, always discerning dates. As it can be noticed from Figure 3.20a, *preop* and *1 year* readings are shown (the only available and valid ones for patient MV001), where the peak for affected and not-affected limb seems separated, either in amplitude or in frequency. However, an oscillatory behaviour is detected for *preop* measurements and this, as discussed in Section 3.1.4, was linked with an erroneous data collection procedure, possibly leading to a peak whose position is not correct, either in amplitude or in frequency. Nevertheless, it does not seem that the information related to the disease contributes to the characterization of it: concerning *preop* measurements, for the foot and the above patella positions the peak is located on the same frequency but affected reading has higher amplitude; for below patella position the behaviour is reversed, therefore affected peak is less marked and is slightly moved to the left; by looking at *1 year* measurements the peaks point out a shift to the right, but again a characteristical pattern could not be detected.

Taking as example patient MV002, Figure 3.20b does not highlight any major difference between affected and reference curves, but a clear difference could be noticed if the date is considered: while *preop* is generally located close to 2.4 GHz frequency, both *6 months* and *1 year* measurements have peaks beyond 2.5 GHz and they seem to have same resonant frequency. Remarkably, there is a 5 dB gap between affected and reference measurements for the wrist after 1 year from the operation: although this could be traced back to an improper calibration (the oscillatory behaviour found for all *1 year* curves), the local minima are similar.

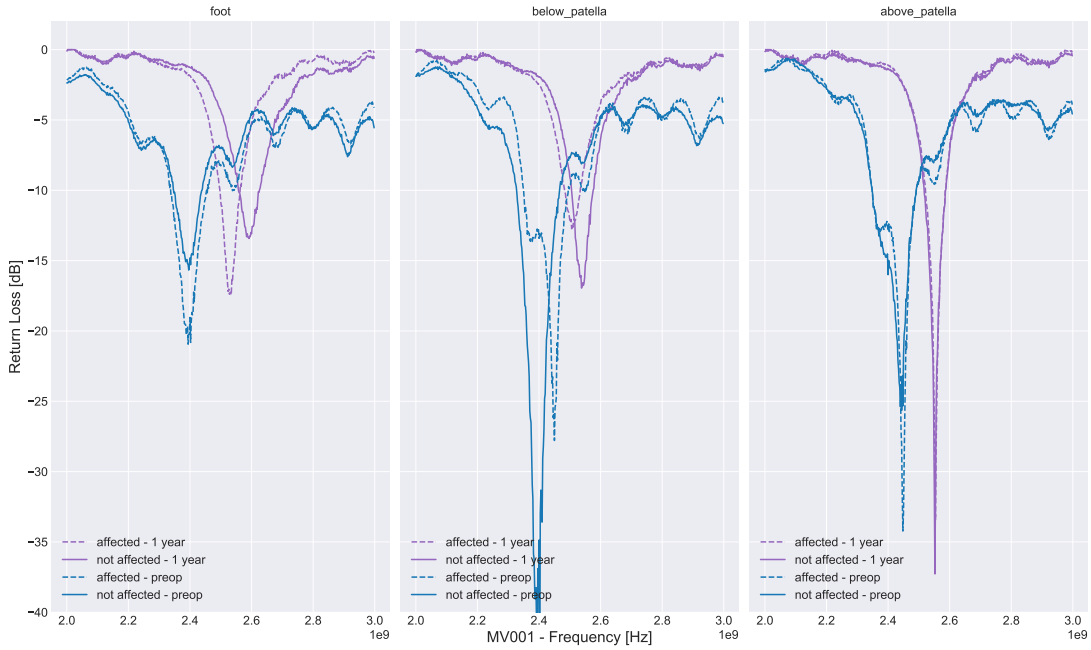


(a) Elbow measurements

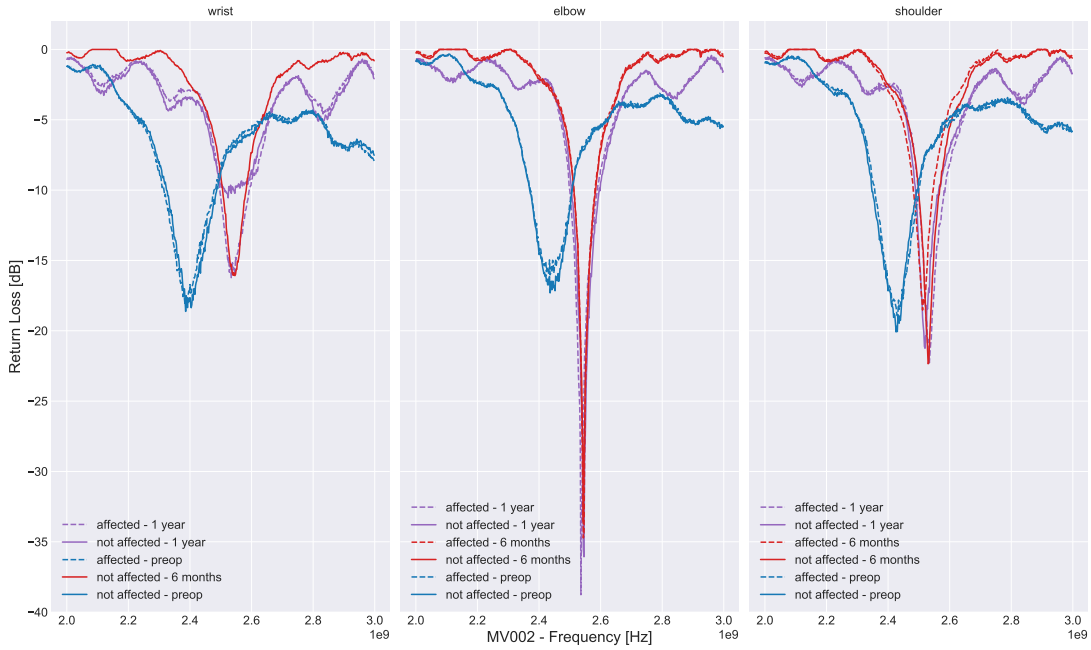


(b) Above patella measurements

Figure 3.19: (a) elbow and (b) above patella evolution in time, considering the average evaluated separately with reference and affected limbs.



(a) MV001 measurements



(b) MV002 measurements

Figure 3.20: MV001 (a) and MV002 (b) patients' measurements are considered to compare behaviour of different body parts.

In such a way to summarize all the results achieved and considerations drawn, a brief list describing them will follow:

I All the measurements near the operation seem to be located around the 2.4 GHz region (especially *preop* and for some patients *postop*);

II The amplitude of the peak seems to vary a lot and a characteristical behaviour to be associated with could not be detected;

III Standard deviation, evaluated on the three readings related to the same spot, may be a useful tool to detect any deceiving behaviour of the measurement (including the possibility of discovering any problem with the hardware);

IV The expected output is a peaked curve, having only a global minimum, approximately located in the region 2.4-2.6 GHz (concerning patients examined);

V If local minima are present, it may be due to an erroneous calibration procedure, leading to partly unreliable data (further investigation may be needed to extract valuable information from these curves, until now no data processing was capable of improving the obtained output);

VI The peaks, even for the same patient, may be very different and a deeper investigation on the effect of the type of skin may be needed (part of this analysis was performed with simulations, as described in Section 3.2.2);

VII A huge quantity of data was discarded (almost 40 %) and some more data were kept even if the output was not much reliable: a precise modus operandi is needed for future data collection, also to evaluate accurately if the output is the result of an inaccurate data collection or some unexpected behaviour is being detected by the measurement;

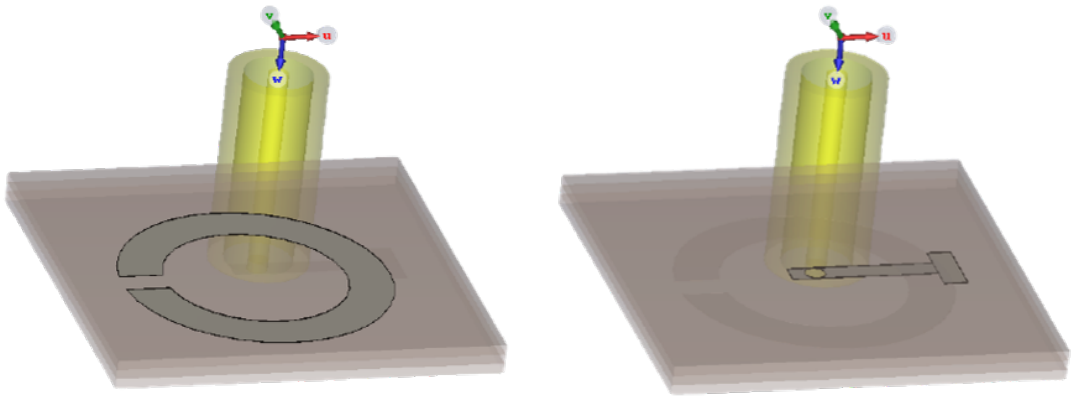
VIII A more careful management of the hardware is needed, especially for the copper cable whose integrity may be compromised by an excessive bending (even if they are rolled up on themselves very gently);

IX More metadata may be really useful to improve the characterization of the curve, both to evaluate the evolution of LO and to understand what the output of the measurement is telling;

3.2 Simulated Measurements

In parallel to the miniVNA data acquisition, **CST Studio Suite** software was exploited in order to extract simulated measurements of the $S_{1,1}$ parameter with respect to different layer compositions. The 3D model of the SRR antenna (the same used for the measurements on the patients), whose internal structure is reported in Figure 3.21, was positioned over different DUTs representing human body stratified models, composed by skin, fat and muscle layers.

3.2.1 Software and 3D Model



(a) Bottom substrate **(b)** Top substrate

Figure 3.21: 3D representation of the SRR antenna, highlighting internal bottom (a) and top (b) substrates.

The dielectric parameters of each material were obtained from the **IFAC database** of the *parametric model for the calculation of the dielectric properties of body tissues* developed by C. Gabriel and collegues[2]. In order to insert this information in the CST software, ϵ'' needs to be evaluated: IFAC dataset contains ϵ' (called *relative permittivity*) and $tan(\delta)$ (called *loss tangent*); according to Equation (3.6), ϵ'' can be easily obtained as the multiplication between these two values.

$$\epsilon'' = \epsilon' * \tan\delta \tag{3.6}$$

[2] `http://niremf.ifac.cnr.it/tissprop/htmlclie/htmlclie.php`

For the simulated data, the folders were created during the acquisition phase, thus, a preparation phase was not needed. It is important to notice that many data were simulated, more than the ones reported in Section 3.2.2 containing additional scenarios; however, since those data were not enough informative or information could be found in other simulations, they were discarded and are not taken in consideration for the next phases.

3.2.2 Considered scenarios

The main objective with these data was to obtain a population of curves similar to the one found with the real measurements, with the intention of correlating the known information about the layer disposition and thickness to the measured data. To do so, a folder containing all these simulated data was created: the output file from the CST software is a `".txt"` file with two columns, one for the RL curve data points and one for the frequency at which the point is located. The general format name of each file is composed of two elements: `"PARAM_VALUE.txt"` and contains information about the 3D model from which the curve was extracted. `PARAM` is the parameter that was considered, `VALUE` is the thickness expressed in millimeters. For more complex scenarios other formats were used. The 3D structure of the DUT representing the human body is a cuboid whose dimensions are:

$$\text{height} = 100\ [\text{mm}];\ \text{length} = 100\ [\text{mm}];\ \text{width} = \textit{variable}$$

In the main folder three types of simulations were performed to check the effects of three different types of skin: *dry*, *wet* and *mid*, i.e. the average between these two; it should be noted that the IFAC database contains information about the dry and wet skin tissues, while the third type was evaluated based on the first two. The disposition of the default three layered 3D model is shown in Figure 3.22.

In the two folders named `fat15` and `fat20`, two scenarios in which a layer of lymph was present are analyzed: the name of the folder refers to the thickness of the fat, respectively 15 and 20 mm; the two dispositions, addressed as `lymph1` and `lymph2`, imply that the layer of lymph is located either between the skin and the fat, Figure 3.23a, or between the fat and the muscle, Figure 3.23b; four thickness values are tested for each scenario, with sixteen simulation in total considering also the variation of the fat thickness.

Inside the folder named `AMPLITUDE`, different scenarios were investigated aiming to reach high values of amplitude in the 2.45-2.6 GHz region. This is done because measured data were characterized by peaks in this region, but simulations were not presenting the same behaviour: as a matter of facts, from the previous step it was possible to discover that decreasing the skin width would lead to a shift of the

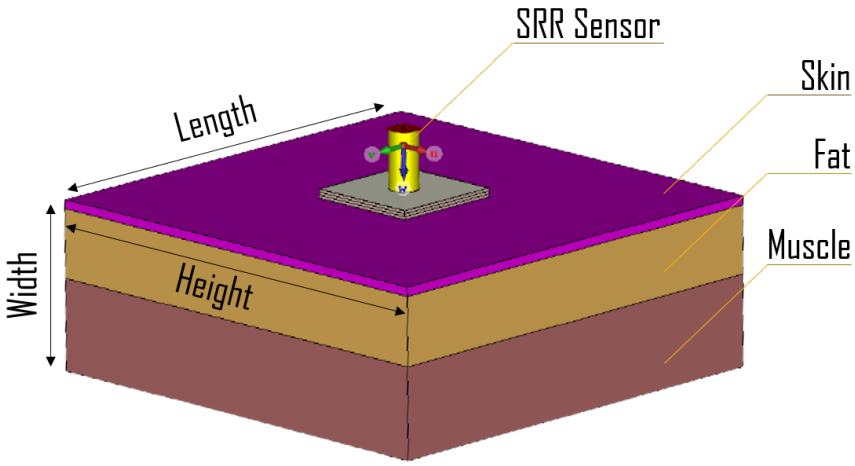
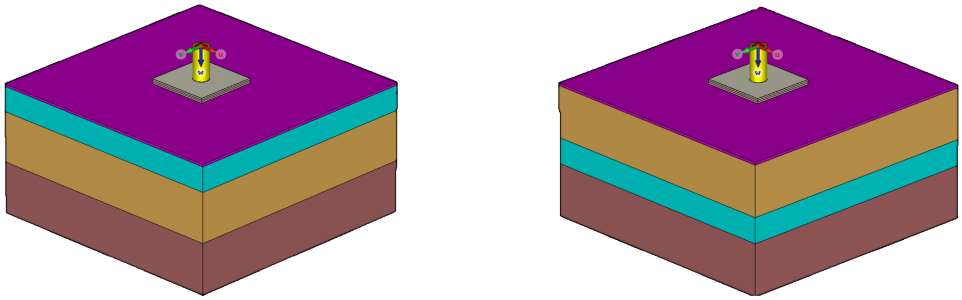


Figure 3.22: Default three layered structure used in the CST environment for the simulations; top layer is skin, middle one is fat and bottom is muscle tissue.



(a) "Lymph1" model **(b)** "Lymph2" model

Figure 3.23: CST Studio Suite 3D model containing lymph layer in two different layouts: (a) with the lymph in between skin and fat layers and (b) with the lymph in between fat and muscle layers.

resonant peak towards right and a decrease of the absolute value of the RL. The main idea was to perform a targeted search, looking at the possibility of moving the peak towards right and increasing its depth. As shown in here, different file names were used to distinguish the scenario considered:

1. `FatWidth_X.txt`: Change the fat thickness (X) while keeping other parameters constant.

   Skin width = 1 [mm]; **Muscle** width = 20 [mm]

2. `Fat-Y%_X.txt`: Change the fat thickness (X) while varying the loss tangent of the fat tissue (Y), increasing or decreasing it by 10, 25 or 75% with respect to the original value.
   Skin width = 1 [mm]; **Muscle** width = 20 [mm]

3. `LymphY_X.txt`: Evaluate effects of the presence of a lymph layer in different scenarios (each one associated with the number Y) by changing its thickness (X) and its position. For $Y = 1$ and $Y = 2$, the same layout adopted before is used (Figure 3.23), but in this case the effect of the lymph layer variation is investigated more thoroughly.
   Skin width = 1 [mm]; **Fat** and **Muscle** width = 20 [mm]

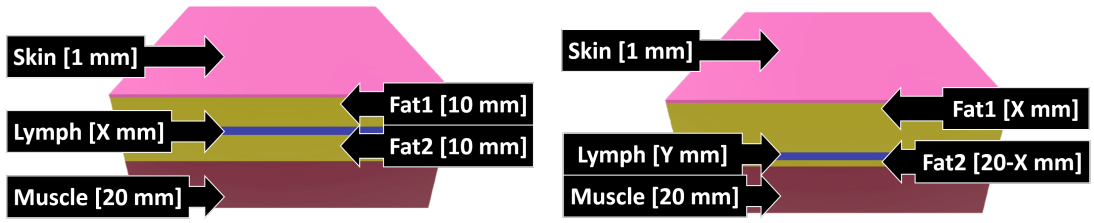
   For $Y = 3$, Figure 3.24a, the lymph layer is positioned in between two layers of fat of 10 mm thick.
   Skin width = 1 [mm]; **Muscle** width = 20 [mm]

4. `Lymph4-Y_X.txt`: Evaluate effects of the presence of a lymph layer varying its thickness (Y) between two fat layers of dimensions X (between skin and lymph) and 20-X (between lymph and muscle), Figure 3.24b.
   Skin width = 1 [mm]; Total **fat** width = 20 [mm]; **Muscle** width = 20 [mm]

5. `BV-Y_X.txt`: Evaluate effects of the presence of blood vessels (whose number is determined by Y), at the center of the fat layer, by varying their radius (X). In the case of three blood vessels (Y=3), also the distance between one blood vessel and the other (expressed by W) is considered, Figure 3.25.
   Skin width = 1 [mm]; **Fat** and **Muscle** width = 20 [mm]; **W** = [5; 10; 15]

File name	**X range [mm]**	**Y range**
`FatWidth_X.txt`	[2.5; 5; 7.5; 10; 15]	\
`Fat-Y%_X.txt`	[5; 10; 15; 20]	[-25; -10; +10; +25] [%]
`LymphY_X.txt`	[0.5; 1; 2.5; 5; 10]	[1; 2; 3] [-]
`Lymph4-Y_X.txt`	[2.5; 5; 7.5; 10; 12.5; 15]	[1; 5] [mm]
`BV-Y_X.txt`	[1; 2; 3; 4; 5]	[1; 3.W] [-]

Table 3.7: File names of the simulated data contained in the folder `AMPLITUDE`



(a) "Lymph3" model **(b)** "Lymph4" model

Figure 3.24: A replica of the 3D model containing lymph layer in between two fat layers with the fat thicknesses fixed (a) or with the fat thicknesses depending on parameter X (b).

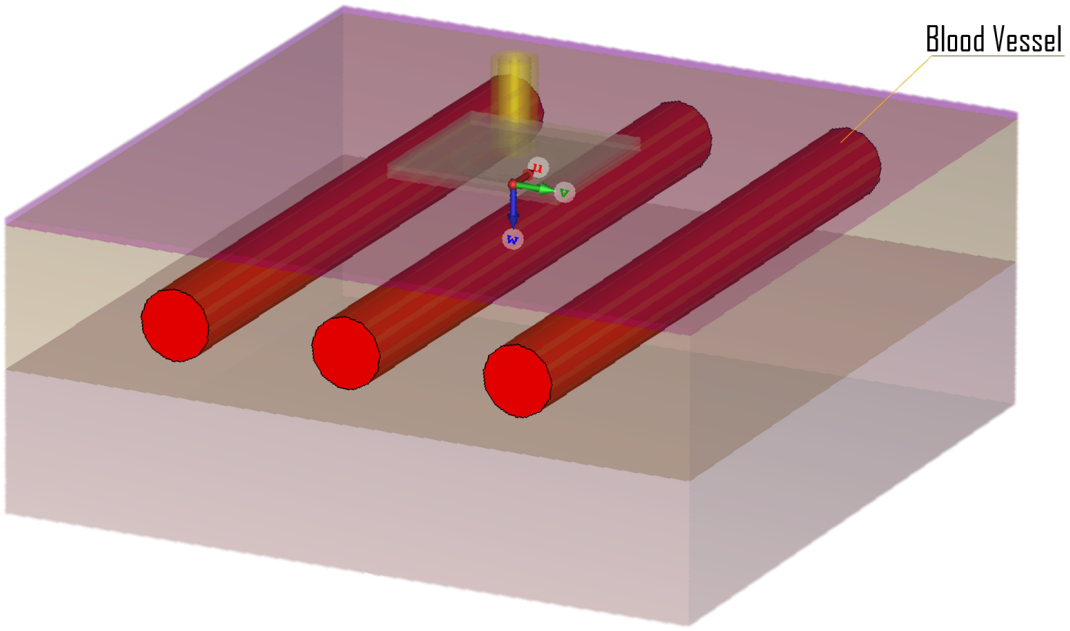


Figure 3.25: CST Studio Suite 3D model containing blood vessels in the fat layer.

Another folder, named `EXPERIMENT`, was created containing experimental simulations performed due to an unexpected behaviour found with the previous simulated data. As it will be better described later in Section 3.2.3, it has been noticed that, by modifying the thickness of the skin layer, it was possible to obtain a huge variation on the resonant peak position and amplitude, larger with respect to the one observed with previously collected data. This phenomenon is certainly linked to the fact that skin tissue is the one in direct contact with the SRR antenna, resulting

in a high impact on the measured RL curve. To deeply investigate this behaviour, different simulations were performed varying the thickness of one material at the time and changing the environment in which the 3D model was inserted; Table 3.8 shows the range of values considered and a description of each scenario follows:

1. `FAT_X.txt`: Skin thickness (`X`) varying with an infinite fat layer on background;
2. `PEC_X.txt`: Skin thickness (`X`) varying with a Perfect Electrical Conductor (PEC) layer on background and without any fat or muscle layer in between;
3. `Fat-pec_X.txt`: Skin thickness (`X`) varying with PEC layer on background and with 20mm of fat layer in between;
4. `MUS_X.txt`: Skin thickness (`X`) varying with muscle layer on background and with 20mm of fat layer in between;
5. `META1_X.txt`: Skin thickness (`X`) varying with a metamaterial, i.e. a material that does not exist yet and with properties that are physically not allowed, with thickness $t_m = t_{skin}$;
6. `META2_X.txt`: Skin thickness (`X`) varying with a metamaterial with thickness $t_m = 1/2 * t_{skin}$;
7. `Fat-only_X.txt`: Fat thickness (`X`) varying with a PEC layer on background;
8. `Pec-2.25_X.txt`: Fat thickness (`X`) varying in between a skin layer 2.25mm thick and a PEC layer on background;
9. `Mus-2.25_X.txt`: Fat thickness (`X`) varying in between a skin layer 2.25mm thick and an infinite muscle layer on background (these simulations were done in order to better understand the effects of the fat tissue alone);

Simulations described in point 5 and 6 were performed with the idea that, if such metamaterial will ever be invented, the effects of the skin could be cancelled out and the measurements could further be improved. This metamaterial has been interposed between the SRR sensor and the skin, as shown in Figure 3.26.

Another important aspect considered at this stage is the boundary of the simulation: until now, all the curves were extracted using an *open (add)* background, meaning that all the simulations were performed considering an air gap between the 3D model and the effective boundary in any direction. In order to analyze the effects of this parameter, different scenarios described previously were tested; it should be keep in mind that only bottom one was modified (the one facing the opposite direction with respect to the sensor) and this was done for few reasons:

1. changing the side boundaries into "infinite", would correspond to extend infinitely the material and, consequently, to see it as an infinite flat surface; this was avoided because other simulations were performed changing the size of the surface and no major differences were found.

2. using a PEC background in any direction but the bottom one would not be useful since it would not represent any real scenario.

3. if the PEC boundary is used on the bottom layer it could tell how much of the signal has been absorbed by the 3D model by comparing the simulation with the one in which air is used as a background;

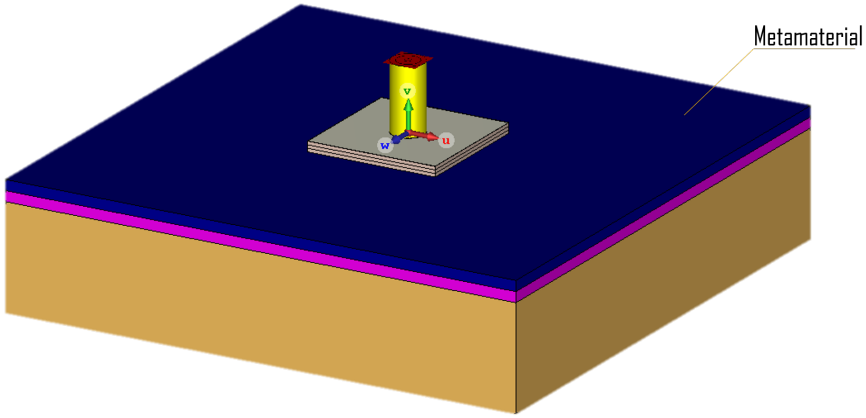


Figure 3.26: CST Studio Suite 3D model where a meta-material is interposed between the sensor and the skin.

File name	X range [mm]
`FAT_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`PEC_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`Fat-pec_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`MUS_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`META1_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`META2_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`Fat-only_X.txt`	from **0.25** to **5** step of **0.25**; from **7.5** to **30** step of **2.5**
`Pec-2.25_X.txt`	[**5; 20; 35; 50; 65; 80; 95; 110; 125; 140; 150**]
`Mus-2.25_X.txt`	[**5; 20; 35; 50; 65; 80; 95; 110; 125; 140; 150**]

Table 3.8: File names of the simulated data contained in the folder `EXPERIMENT`

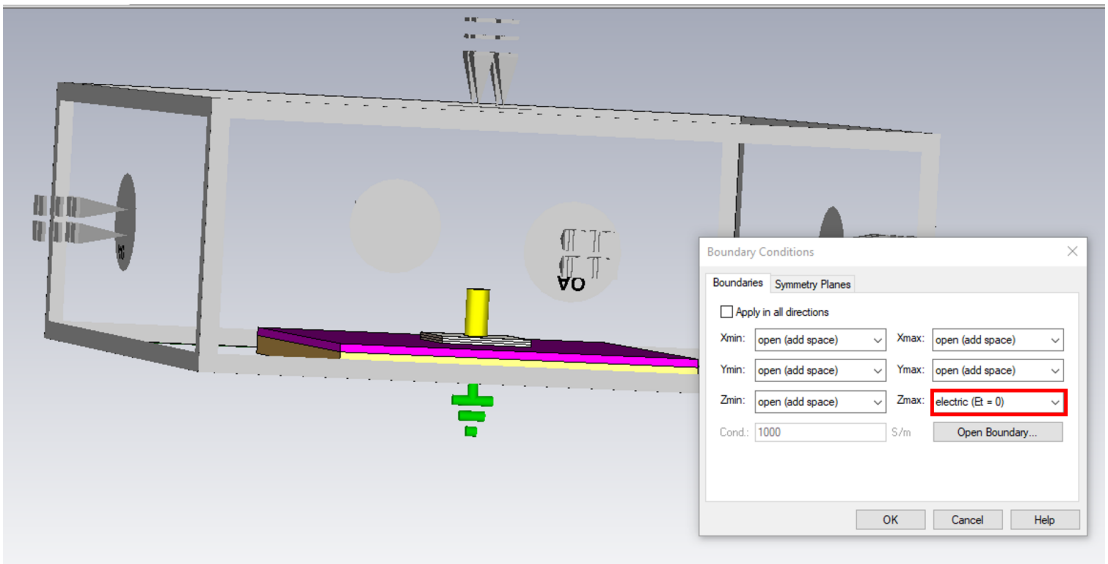


Figure 3.27: CST Studio Suite 3D model, showing how the boundaries are selected; here the PEC boundary is used below the fat layer.

Figure 3.27 shows the way boundaries are selected, highlighting how the *open (add)* background considers an empty space. It is also possible to notice that the PEC boundary corresponds to the option in which $E_t = 0$, i.e. the electrical field in that direction is zero.

3.2.3 Data processing

Standardization

The standardization procedure was particularly suitable for merging different types of data: the simulations and the measured data had different frequency sampling too and with this phase it was possible to compare the curves in a easier way and save them in a unique format. Same considerations done for real measurements are valid for the standardization process on the simulated data.

Filtering

No filtering process was needed for the outputs of simulations. However, it should be considered that interpolation connecting all the missing frequency samples was required: this can be considered the reason for which the RL curve of simulated data looks smoother than the other data.

Validation

Considering simulated data, the main concern was to understand the impact of each layer with the objective of clarifying if it was possible to extract any valuable information from the measurement itself. This phase was needed because an unexpected behaviour was found during the attempt of obtaining higher amplitude for peaks located in the 2.45-2.6 GHz. In fact, in order to create a similar dataset to the one of measured data, many factors were taken into account, as largely described in Section 3.2, like the skin and fat thicknesses, the type of skin, the fat loss factor ϵ'', the presence of a lymph layer and so on. Nevertheless, most of these attempts were not providing the expected outcome, i.e. a shift towards higher frequency was obtained only coupled with a reduction in the amplitude; in particular, during the analysis on the effects of the skin type (dry, wet or average between the two), it was noticed that increasing the skin width, the frequency changes were noticeable.

This analysis led to a deeper investigation on the effects of the skin tissue alone: the main issue for this project was to understand if this information was essential, since it was not available, or how it could be bypassed. If the measurements highly depend on the material in direct contact with the antenna, it will be required to cancel its effect on the RL curve in order to characterize hidden layers and estimate the presence and/or amount of lymph collected.

3.2.4 Results

The first result presented for the simulations is obtained by comparing different type of skin (*wet*, *dry* and *mid*): here the graph is composed of three different curves each one associated to a different type of skin and all the layers have fixed width. As it can be seen from Figure 3.28, the three curves are very similar, but they are not overlapping: peaks have a distance of few MHz and almost no difference in amplitude (approximately 1 dB). If the skin is more hydrated the peak is slightly shifted to the left and, as expected, by averaging EM parameters the resulting curve falls in between the other two. Since a bigger difference among these three was not found, the *mid* type skin was used for other simulations.

The second scenario considered is the presence of a lymph layer: at the beginning two different simulations were tested, namely `lymph1` and `lymph2`, reported in Figure 3.29 where different lymph thicknesses are considered having a fat layer width equal to 15 mm and 20 mm. The skin thickness is kept constant and equal to 1 mm. Concerning the first scenario, from Figure 3.29a it is possible to notice that the effect of the lymph is quite important; on the other side, curves that have same lymph thickness but different fat width are completely overlapping (the name found in the legend is `fatX Y`, where `X` and `Y` are respectively the thickness of fat

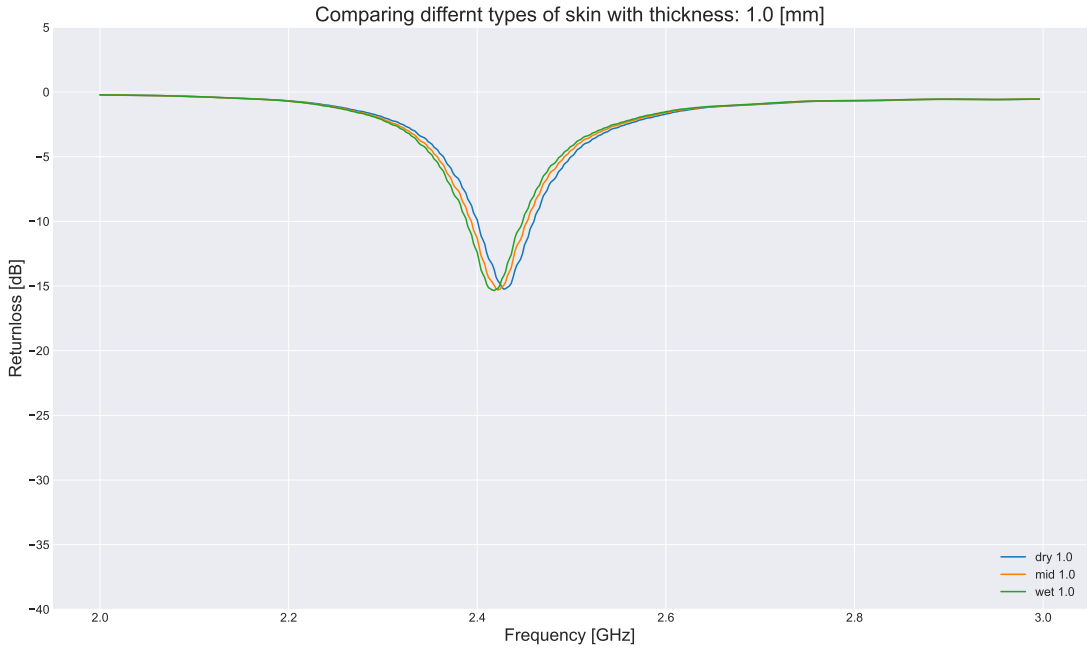
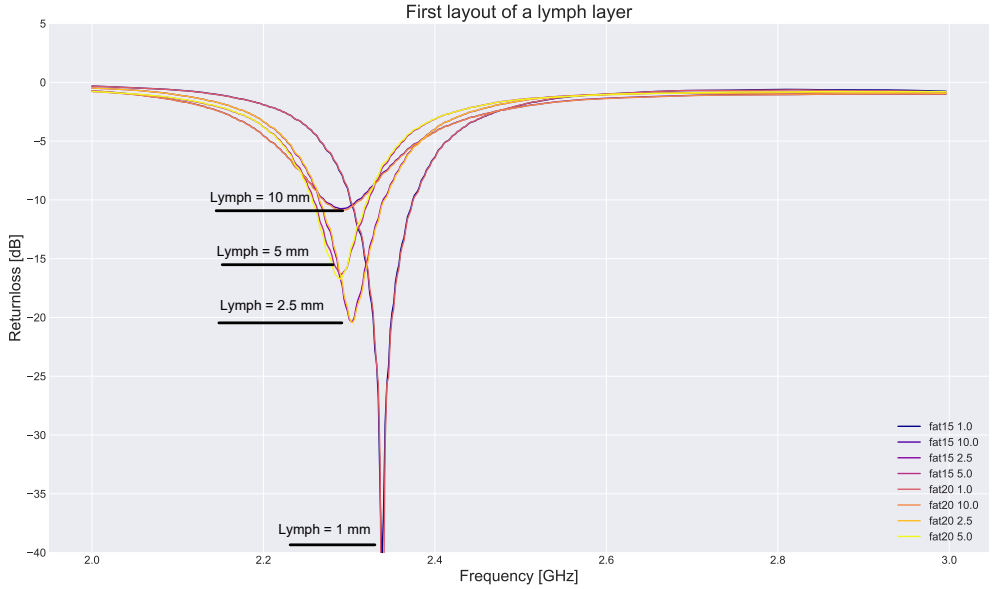


Figure 3.28: Different skin types are compared having a thickness of 1 mm.

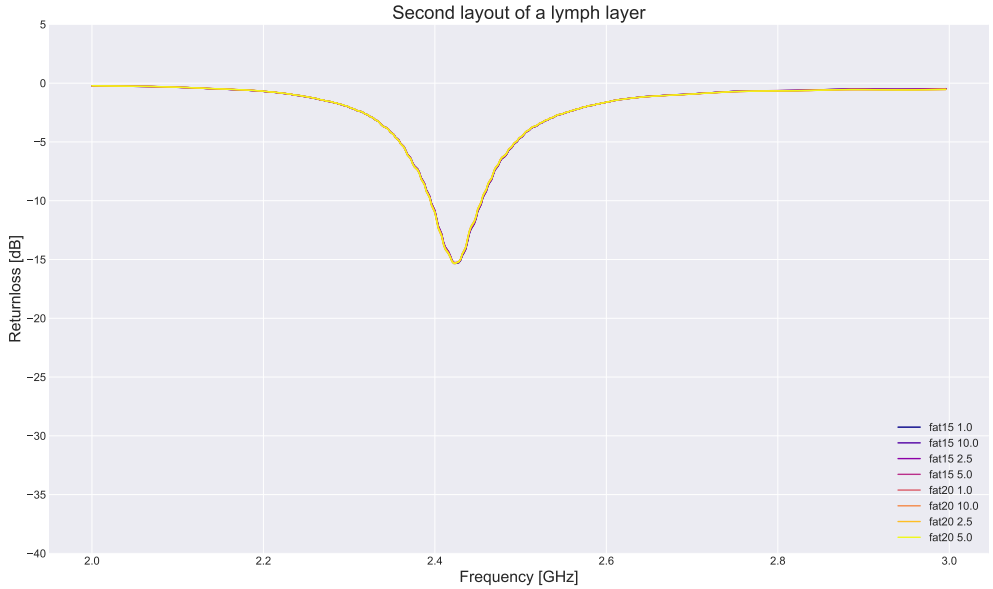
and lymph layers). In here, the peaked behaviour is maintained as the thickness of the lymph increases from 1 mm to 5 mm; it is not present when the lymph layer is equal to 10 mm, having the RL curve barley reaching -10 dB. However, it is interesting to point out the fact that all these peaks are located in the region 2.3 - 2.4 GHz, far away from the real measurements. On the opposite, no effect can be noticed in the second scenario, Figure 3.29b, where all the curves overlap: this may be due to the fact that the newly added material has a similar composition to the muscle and the signal is not capable to perceive it presence.

As previously discussed, the second part of the simulations was focused on the possibility of increasing the depth of RL curves by achieving a shift to the right (the 2.4-2.5 GHz region covered by real measurements). To summarize, four different tests were performed using as a reference a model with 1 mm of *mid*-type skin:

Test#1 Evaluate effects of the fat thickness: by maintaining constant all the parameters apart from the fat layer, in Figure 3.30 different curves are represented. Generally, all peaks are located near the reference one with very small differences (both in amplitude and frequency), except for the 2.5 mm layer: this may be traced back to the fact that this fat layer is not strong enough to reflect or absorb the signal, resulting in a predominant effect given by the muscle (the peak is moved to the left, as happened with the lymph).



(a) "Lymph1" Return Loss curves



(b) "Lymph2" Return Loss curves

Figure 3.29: Simulations with different lymph layer widths located between skin and fat (a) and between fat and muscle (b).

Test#2 Evaluate effects of loss tangent in fat layer: the parameter analyzed here is the loss tangent in the fat layer. The remarkable behaviour discovered here is an increase of the peak amplitude (since it is negative it is a reduction of its value), with a small shift to the left when the $\tan(\delta)$ is increased by 75%; it is reasonable to think that the material is denatured and probably it is not anymore representing the human fat. These considerations are derived from Figure 3.31. Smaller variations of the loss tangent (both increase and decrease) are not providing any noteworthy differences with respect to the reference curve; the only intriguing fact is that a decrease in the loss tangent produces smaller changes compared to an increase of the same amount.

Test#3 Evaluate effects of lymph layer presence: besides the previously mentioned experiments `lymph1` and `lymph2`, whose results were replicated, two additional scenarios were tested. For `lymph3`, where the lymph is located between two 10 mm fat layers, only a small shift of few dB was achieved for the peak, but no movement in frequency. In `lymph4-Y` (where `Y` stands for the lymph thickness) the lymph layer is basically shifted by increasing the top fat layer while decreasing the bottom one, maintaining a total of 20 mm of fat; here, again, no major differences could be remarked with respect to the reference curve, as shown in Figure 3.32. The only strange behaviour is seen when the top fat layer is equal to 2.5 mm (red curve), where a small movement to the left is obtained.

Test#4 Evaluate effects of blood vessels presence: to test if the blood vessel may have an effect on the measurements, their presence was simulated having different size. In both scenarios, with either one or three vessels, there are no major changes with respect to the reference simulation, as it can be seen in Figure 3.33.

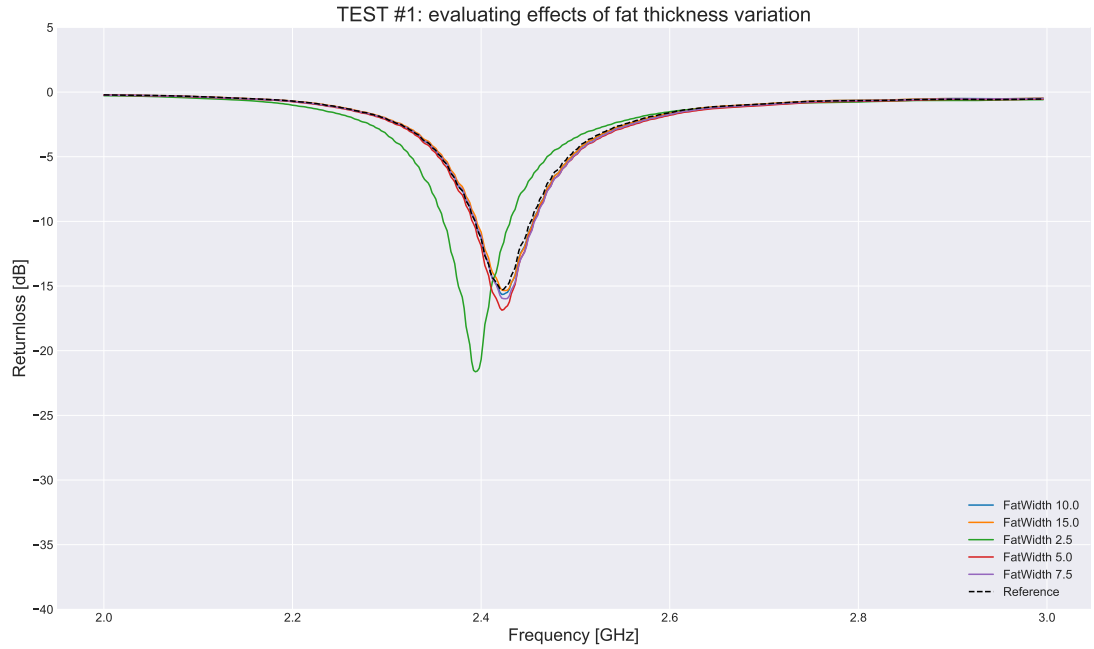


Figure 3.30: TEST #1: Different fat widths not varying other parameters.

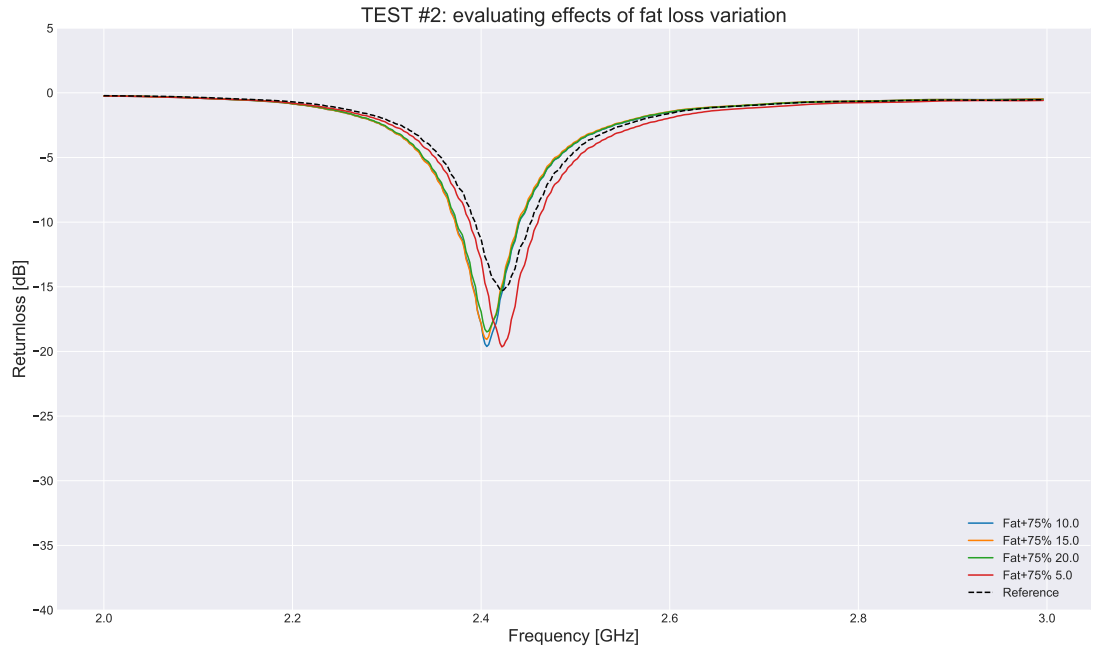
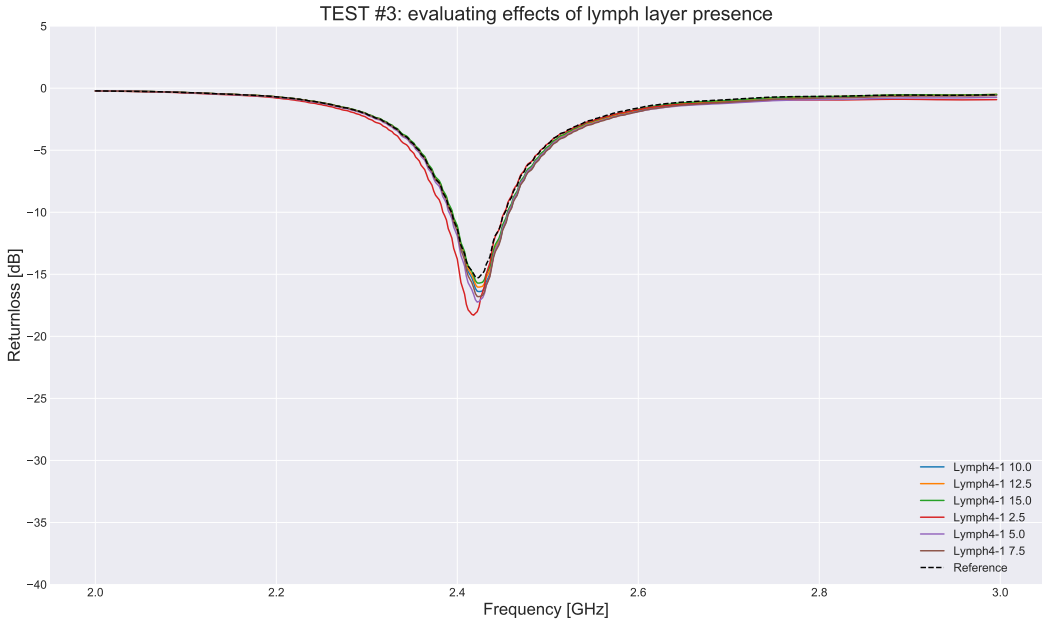
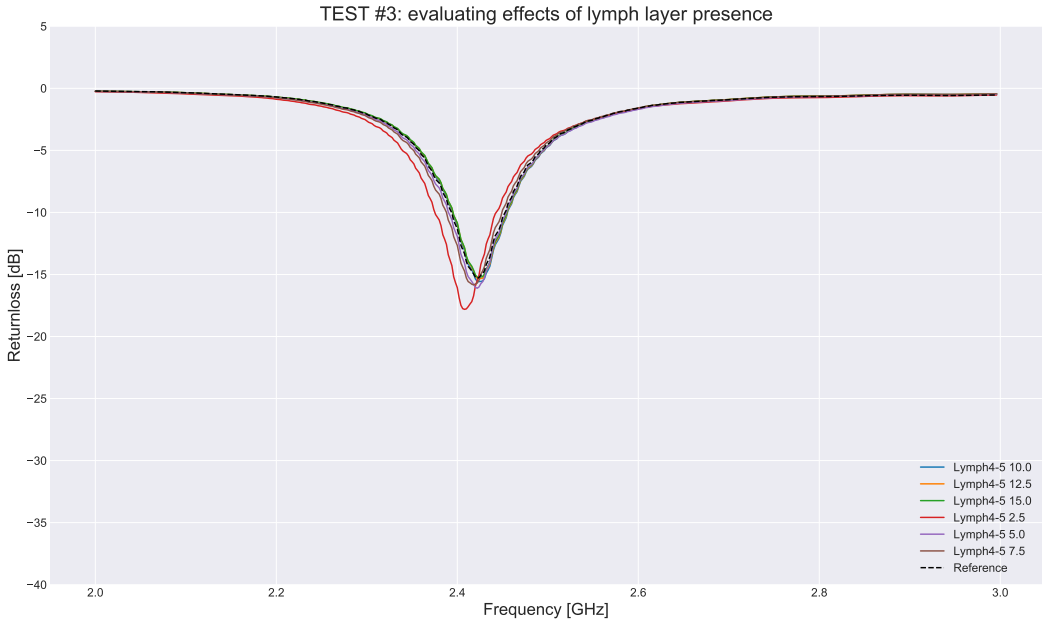


Figure 3.31: TEST #2: Different fat widths with an increase of the fat loss tangent by 75%.

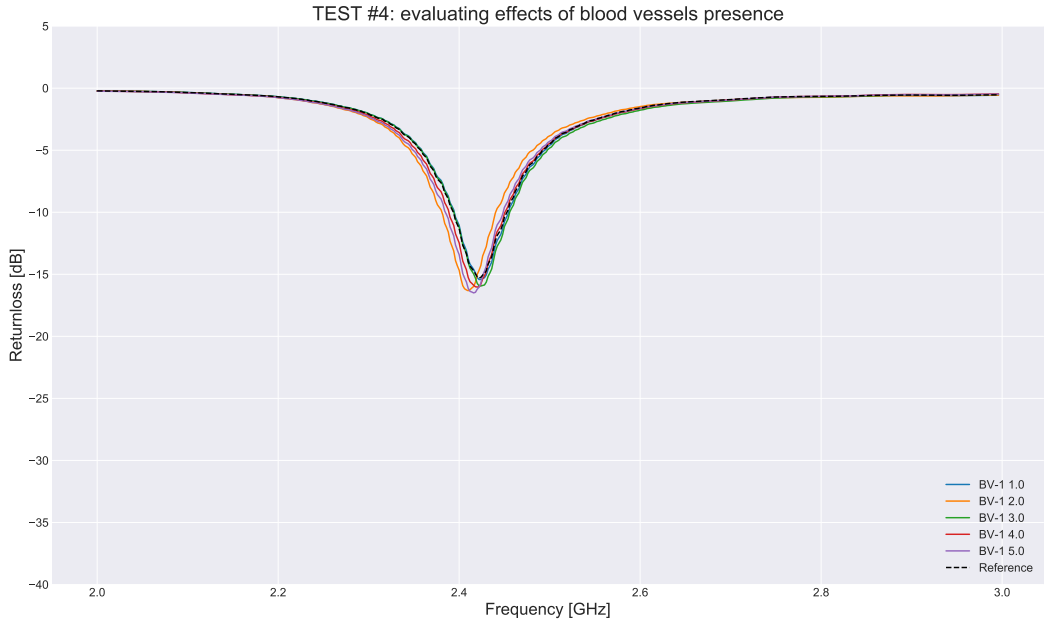


(a) "Lymph4-1" model: 1mm of lymph layer

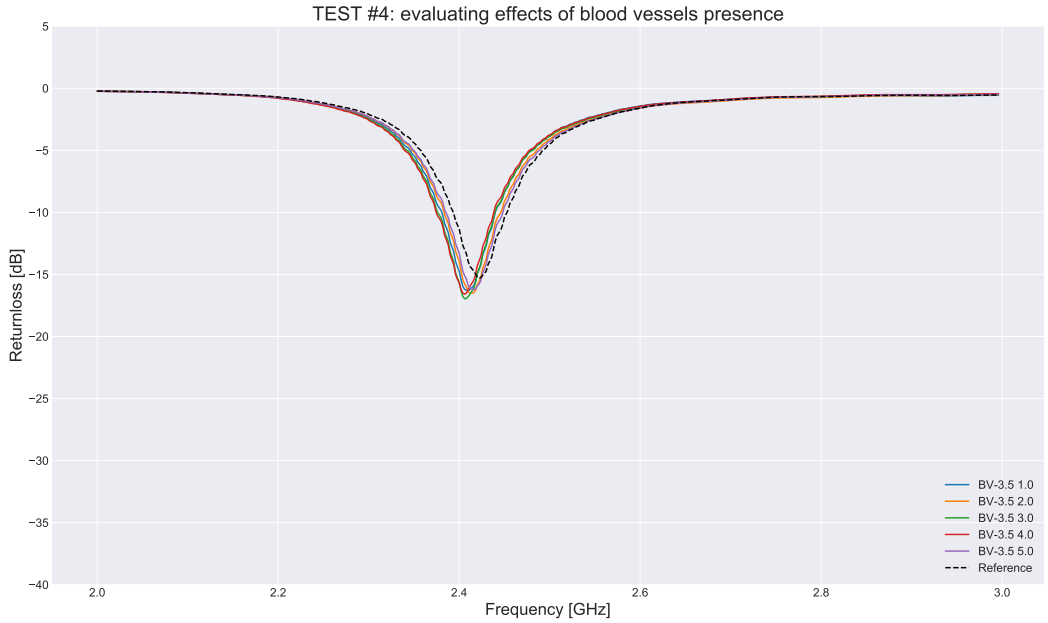


(b) "Lymph4-5" model: 5mm of lymph layer

Figure 3.32: TEST #3: Lymph layer shifted by varying the fat thickness, with a lymph width equal to 1 mm (a) or equal to 5 mm (b) .



(a) "BV-1" model: only one blood vessel



(b) "BV-3.5" model: three blood vessels

Figure 3.33: TEST #4: Blood vessel with different radius; the number of vessels is either one (a) or thrree (b), where the distance between them is 5mm.

The third and final phase of the simulation was focused on the analysis on the effects of skin thickness. In particular, this was performed consequently to the analysis performed on the type of skin, where different widths were tested as well; noticing that the resonant peak moved to the right as the skin width decreased, a deeper investigation was performed with more thicknesses and a step of 0.25 mm. At this stage also different boundaries were tested to better understand the effects of the background.

The first scenario under study is the skin thickness variation with an infinite layer of fat as background, i.e. the boundary adjacent to the fat layer is infinite. As it can be seen from Figure 3.34, the peaks are moving from frequencies around 2.5 GHz towards 2.3 GHz when the thickness of the skin is increased from 0.25 mm to 5 mm. Another characteristical behaviour is the fact that, coupled with a shift to the left, there is an increase in the amplitude until a certain width is reached (2.25 mm); subsequently, the peak amplitude decreases.

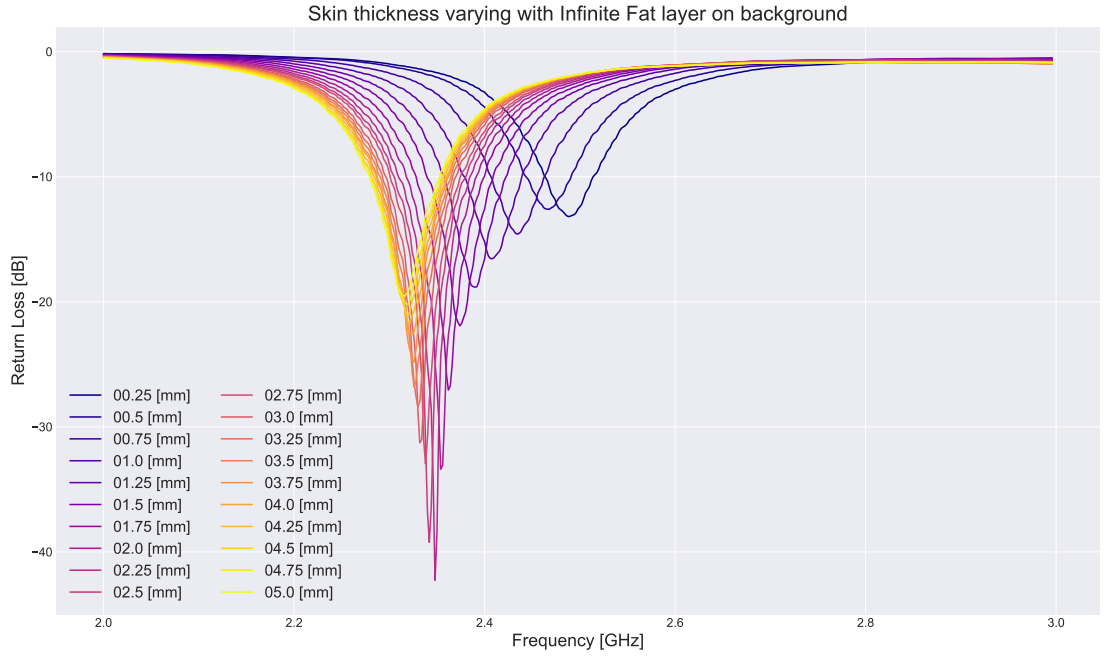


Figure 3.34: Skin thickness varying with an infinite fat layer as background; range is 0.25 - 5 [mm].

Another test has been performed using a 20 mm fat layer and a PEC boundary: results are reported in Appendix A but the difference is quite unnoticeable. Instead, to evaluate the effect of a muscle layer located on the background (also in this case it has infinite width), another simulation was performed and results are reported in Figure 3.35: also here the curves seems to be almost identical, if the same range is considered; indeed, here an additional range for skin widths is studied, from 5 mm to 30 mm, with a step of 2.5 mm. The result is not a continuation of the trend highlighted previously: for the new analyzed range, the peak frequencies fluctuate around 2.3 GHz and the peak amplitude around -12 dB. The major difference between these simulations is the minimum peak amplitude, reached in both cases at 2.25 mm of skin but having a value of -50 dB in latter case against the -42 dB obtained before.

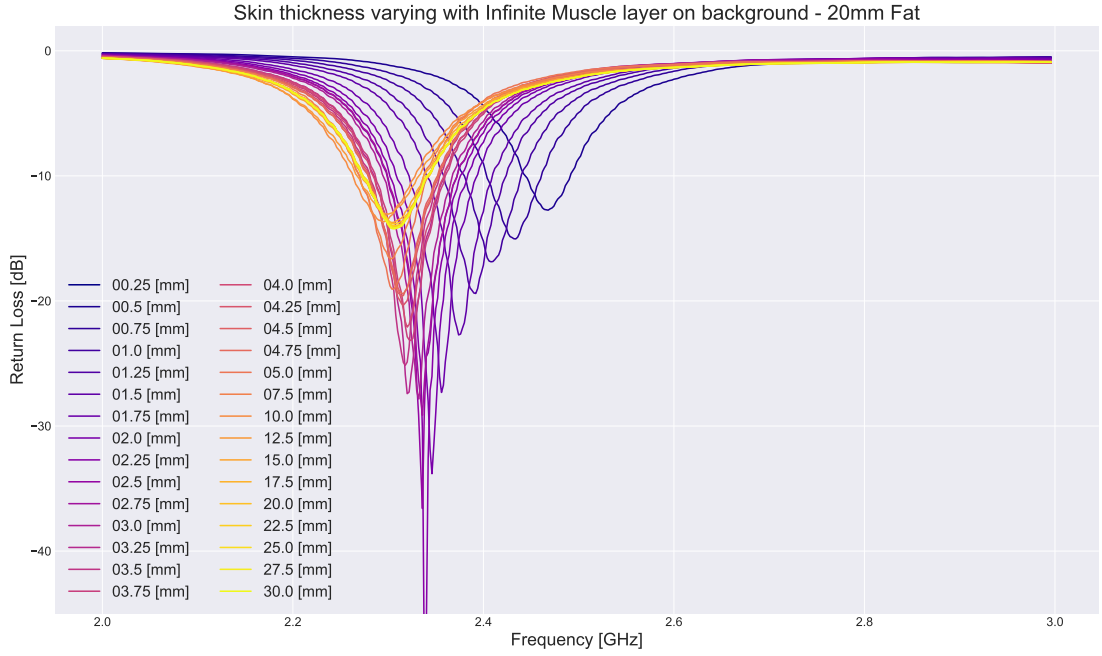
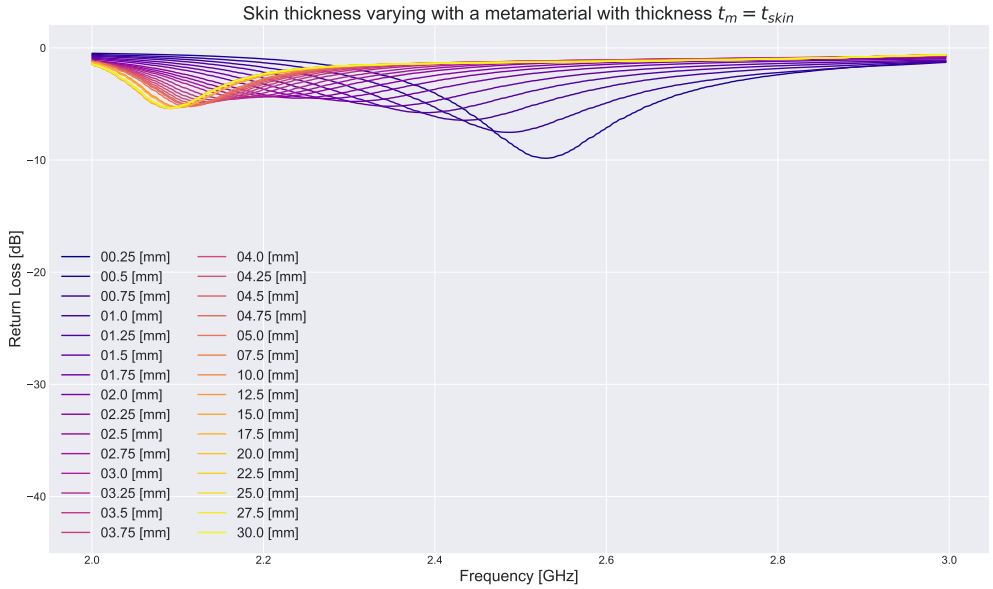


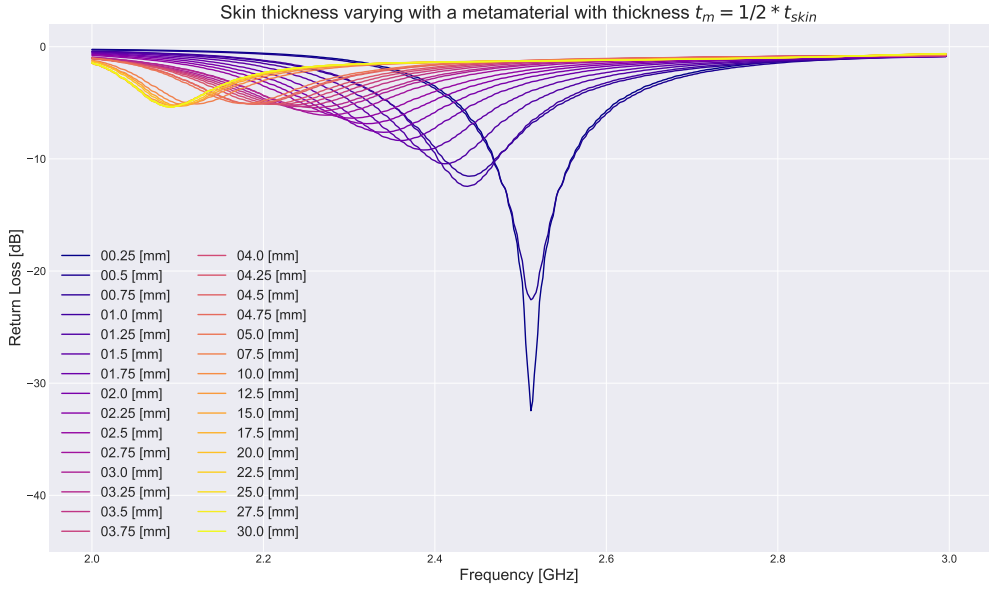
Figure 3.35: Skin thickness varying with an infinite muscle layer as background, having 20mm of fat between skin and muscle; range is 0.25 - 30 [mm].

As already presented in Section 3.2.2, another scenario investigated was the presence of a metamaterial in between the sensor and the skin. Two different simulations were performed: the first one having a width of this material equal to the one of the skin, Figure 3.36a, and the second one having thickness equal to half of the skin tissue, Figure 3.36b. As it can be noticed by the first scenario, this hypothetical material is capable of reducing drastically the amplitude of the curve and to reach frequency slightly larger than 2.1 GHz, far away from the ones obtained without it. If the second situation is considered, the minimum frequency achieved is now around 2.2 GHz, but the metamaterial is still effective for limiting the drop in amplitude achieved when it is not present, allowing a minimum of approximately -12 dB (if widths smaller than 0.5 mm are discarded).

To have a general overview of the analyzed scenarios a recap graph containing all the peak locations is shown in Figure 3.37; few more scenarios not described before are present, but since they were not adding any valuable information, their description is simply reported in the legend of the figure. On the top left quadrant the peak's RL is expressed as a function of its frequency; on top right it is function of the skin thickness; on bottom left the skin width is a function of the peak's frequency. Except for the three curves related to limit cases (metamaterials and only fat layer), all the peaks have a minimum at 2.25 mm. Instead, for all the scenarios the frequency trend with respect to the skin thickness is a shift to higher frequencies when the width is decreased.



(a) Metamaterial presence: $t_m = t_skin$



(b) Metamaterial presence: $t_m = 1/2 * t_skin$

Figure 3.36: The presence of a metamaterial is tested by varying its thickness in proportion with the one of the skin, either being equal (a) or being half of it (b).

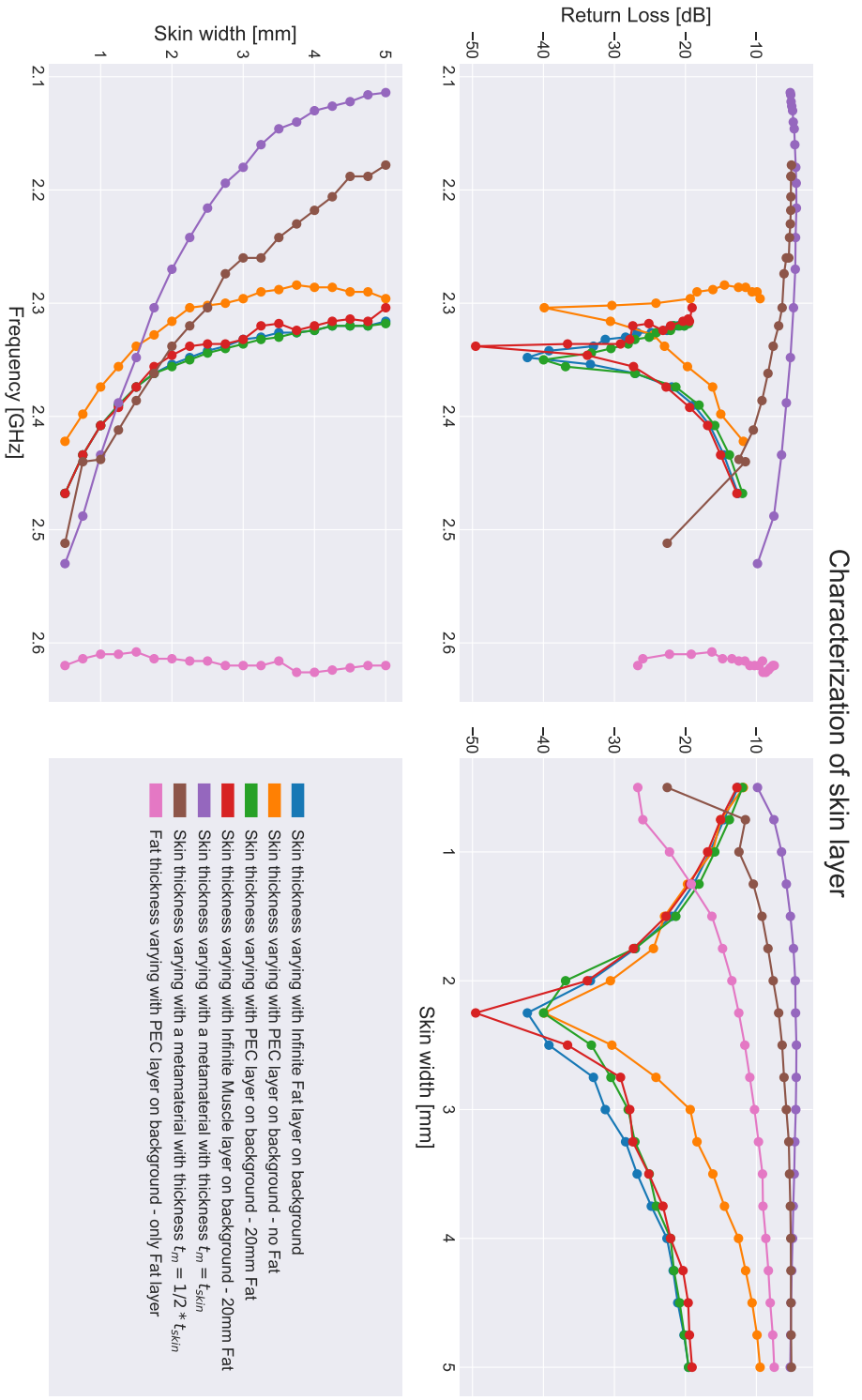


Figure 3.37: Peak location while skin thickness is varying in different scenarios.

Chapter 4

Clustering and classification

This section is dedicated to the characterization of the dataset through ML algorithms deployed. In particular, two sub-sections will be presented explaining the methodology adopted, first for the clustering part and secondly for the classification one; in the end another sub-section with the results will be included.

4.1 Clustering

Unfortunately, a clear scheme could not be found with the procedure described in Section 3.1.5, thus another solution was investigated: the main idea was to cluster different curves in order to have groups of similar data. The similarity criterion is found by the algorithm and the output will be a class to which the data point will belong to.

Since the dataset is composed with both categorical and continuous data, two different clustering algorithms were tested: **KMeans** and **KModes**. While the first one can only take into consideration continuous data, the second is capable of dealing with categorical data too, as reported in section 2.5. The dataset used for this process is the one seen in Table 3.3 and Table 3.4, discarding those columns not useful for the ML, e.g. `NAME` which is used only for plotting the curves.

Even though KMeans cannot exploit categorical data, the idea was to perform the clustering on some features in order to find if any correlation between the RL and the metadata at disposal. In fact, the inputs to this algorithm are the minimum value, its frequency and the bandwidth of the curve; however, since the algorithm requires the number of cluster as input, but this is not an information known a priori, the intuition was to run the algorithm with different number of clusters to find the most suitable number. Nonetheless, since the number of data points

is limited, it may happen that a cluster is constituted of less than ten samples and those cluster were not considered. To implement this algorithm the `KMeans` library provided by `sklearn.cluster` was used. A scaling process was tested for `POS`, `DATE` and `AFFECTED` columns, those used in the algorithm: since it showed slightly better results, it was deployed for the clustering process.

Concerning the KModes algorithm, as mentioned in Section 2.5, it allows to take into consideration categorical features and can be addressed as a more complete tool for the clustering phase. The idea is to exploit this algorithm to cluster data into different classes by looking at all the available information, including those that were not taken into account before. Since this algorithm is based on the previous one, it is necessary to specify the number of desired clusters and the procedure adopted is the same cited above. The implementation of this algorithm was possible exploiting an independently developed library called `kmodes`, licensed by MIT [1]. The aim of this phase was to move from an unsupervised problem to a supervised one: for each data point a class was associated, and this will be used to perform the classification. It is important to note that the clustering is very sensible with respect to the initial points used to perform the algorithm, chosen randomly. The randomness of these points is linked to one parameter called `random_state`: an integer number that sets the seed to the random number generator. In this way the obtained results are reproducible by using the same number and different results can be compared using different numbers but maintaining other parameters unchanged.

4.2 Classification

For the other problem associated to the classification performed using the output of the clustering, Decision Tree classifier has been chosen: this supervised ML algorithm requires a target variable to be trained with, which is obtained by the clustering process performed in advance. The interesting quality of this method is the capability of creating decision rules in the form of a tree: the deeper it is, the more complex these rules are. Each time the clustering algorithm is run, a different DT is obtained, trained on the dataset updated either with the KMeans or KModes. The output will be a tree that considers the most important features used to obtain the output class: each leaf will divide the dataset according to a simple inequality ($feature \gtreqless value$) until a dead leaf is found, where the cluster is classified. In this way it is possible to use the tree from a measurements and follow the branches to classify it.

[1] Documentation available at `https://pypi.org/project/kmodes/`

It must be noticed that beside the problem presented here, another possible classification has been performed by using the `AFFECTED` column as the class to predict. This would correspond to the possibility of detecting the disease directly from SRR sensor readings. However, given the results presented in the Section 3.2.4, the influence of the layer in direct contact with the sensor (skin) is fundamental and if a method that could cancel out or normalize this information is not used, it is likely difficult to obtain any valuable result. To perform this classification the Ensamble method was used, putting together Logistic Regression, Random Forrest and Gaussian Naive Bayes classifiers.

4.3 Results

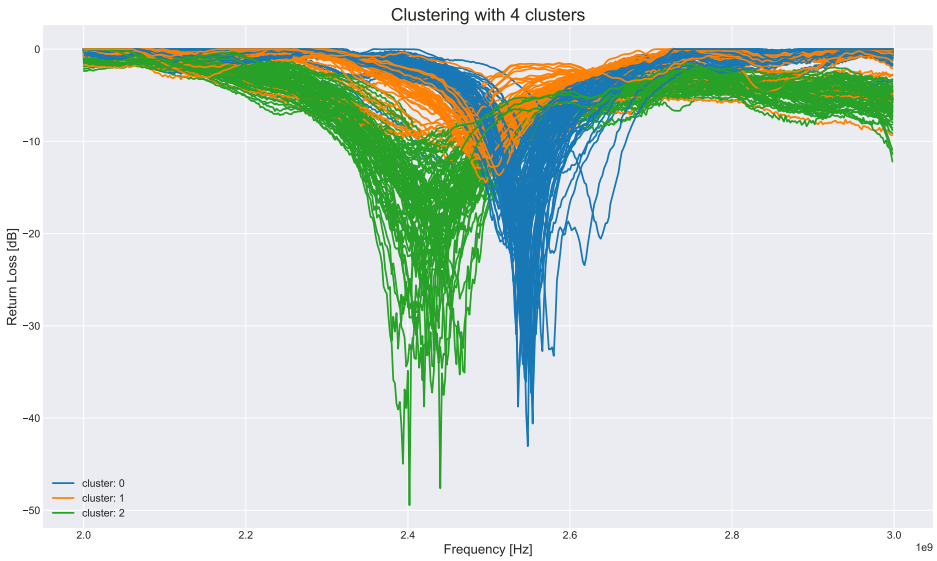


Figure 4.1: All curves clustered in different colors for KMeans algorithm performed on 3 out of 4 clusters.

The first result obtained by the clustering process is the division of the RL curves into different clusters, each one associated to a different color as shown in Figure 4.1. Because one cluster was not having enough data points (10 is the minimum threshold used to consider it valid), here only three out of four clusters are present: remarkably the three groups of curves are well separated and characterized by different trends, having peaks located in three well distanced regions. For example, cluster 0 and 2 have a similar peak amplitude but their frequency location is quite different.

To highlight what stated before, another graph characterizing the distribution of continuous parameter is presented in Figure 4.2: except for the Body Mass Index (BMI), which seems not enough informative and useful for this process, the other features are well separated. The black dot represents the mean value of the cluster for each parameter while the standard deviation is described by a red line. The fact that all the standard deviations do not overlap is significant, especially it is considered that the only features used by the algorithm are `MIN IDX` and `BANDWIDTH`, as shown by the histogram representing the feature importance used for the DT in Figure 4.3.

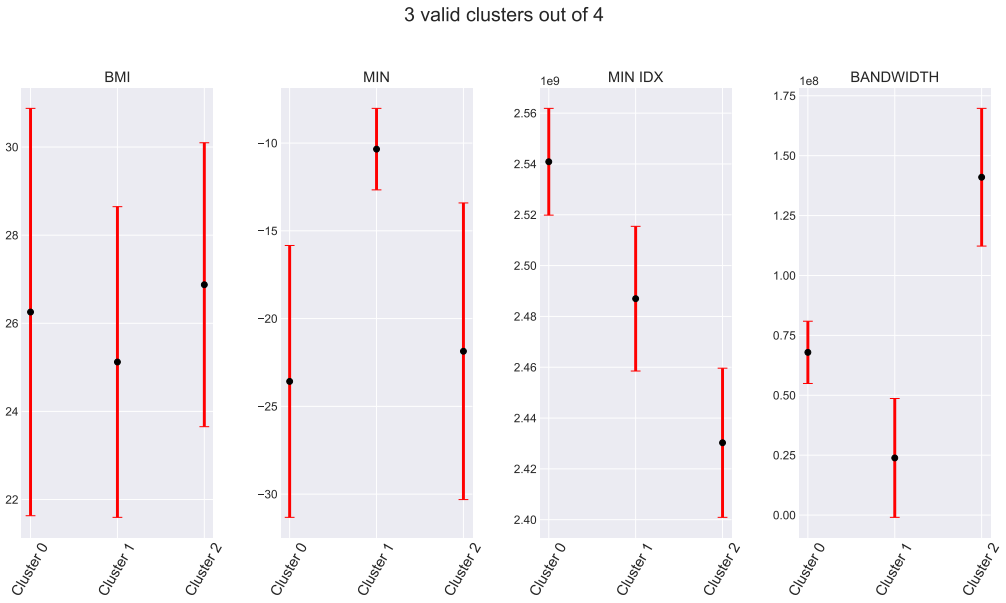


Figure 4.2: Main continuous parameter distributions for a clustering using KMeans algorithm performed on 3 out of 4 clusters.

To see how the curves in the same cluster are characterized, referring to the same situation analyzed before (3 out of 4 clusters), the mean curve of each cluster is plotted in Figure 4.4 alongside with the cloud of RL curves belonging to that cluster. Once again it is possible to notice how the shape of these three clusters is well distinguished and spatially separate one from the other.

The classification performed using the DT tool yields as output the tree represented in Figure 4.6, where most important features are used to cluster different sub-groups of data samples until divisions are not anymore possible and the proper cluster is defined.

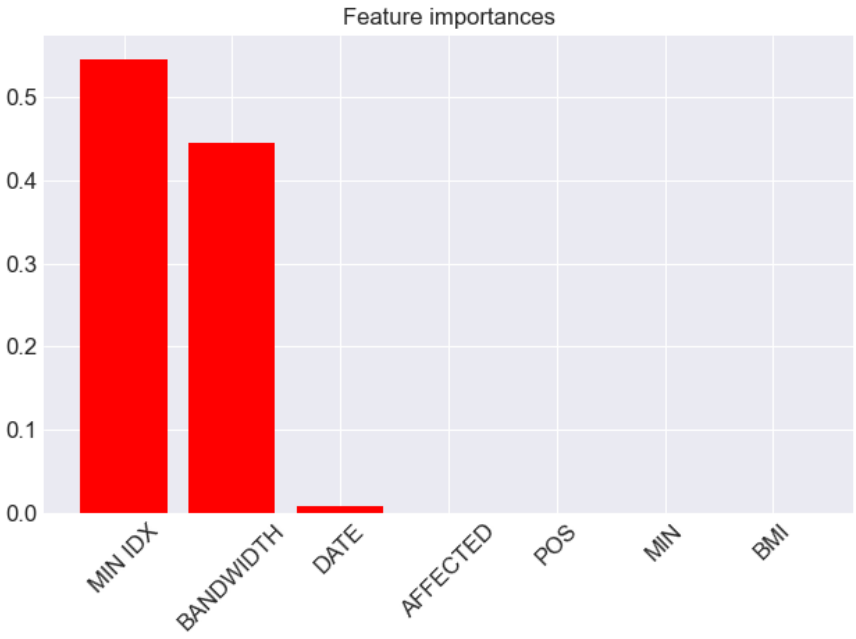


Figure 4.3: Feature importance of KMeans algorithm performed on 3 out of 4 clusters.

This procedure provides an interesting result with respect to continuous variables, but lacks the capacity of clustering categorical information; in particular, being `AFFECTED` one of the most important feature to be classified, i.e. to predict if patients have LO or not, the aim would be to find clusters with high/low amount of "affected" measurements.

To judge how good this algorithm was to perform such work, histograms present in Figure 4.5 depicts the distribution of each cluster for different categorical features. Starting from the `POS` variable, the three distributions seem quite balanced and peaks are not present, meaning that this clustering was not able to characterize the body location of the measurement; similar consideration could be done for `AFFECTED` where, even if a small gap for clusters 0 and 1 are present, they are not well separated and it is not possible to associate one cluster to measurements on affected or reference limbs; considering instead the `DATE` feature, the clustering seems to be more effective: cluster 2, the green one, is mainly located in the period near the operation (*preop*, *postop* and *1 month*) while cluster 0 and 1 are predominantly characterizing later stages from the operation. This consideration is in line with what has been discussed before, thus peaks with smaller frequencies are associated with measurements nearer to the operation.

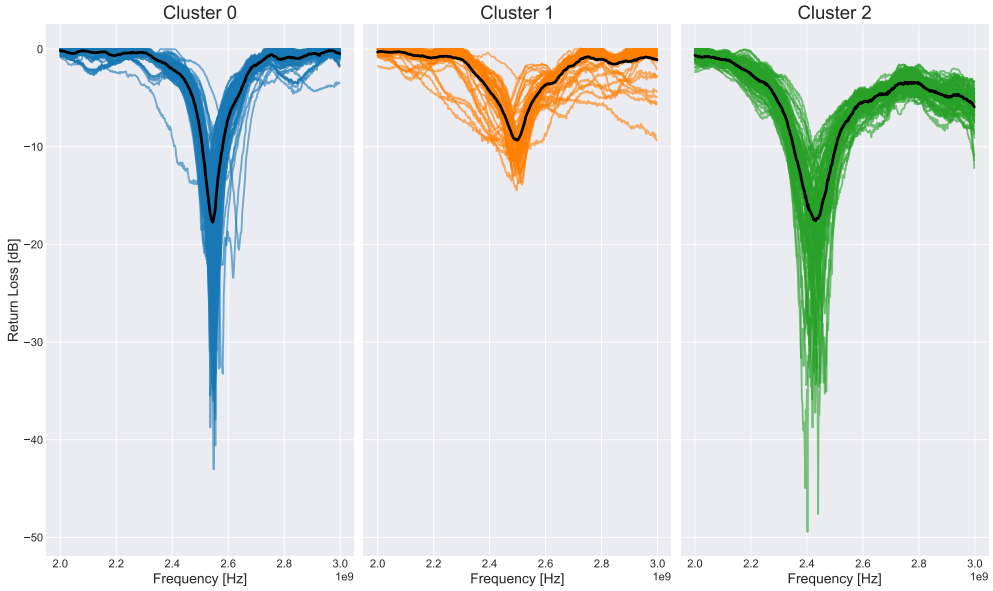


Figure 4.4: Each cluster with the corresponding mean value for KMeans algorithm performed on 3 out of 4 clusters.

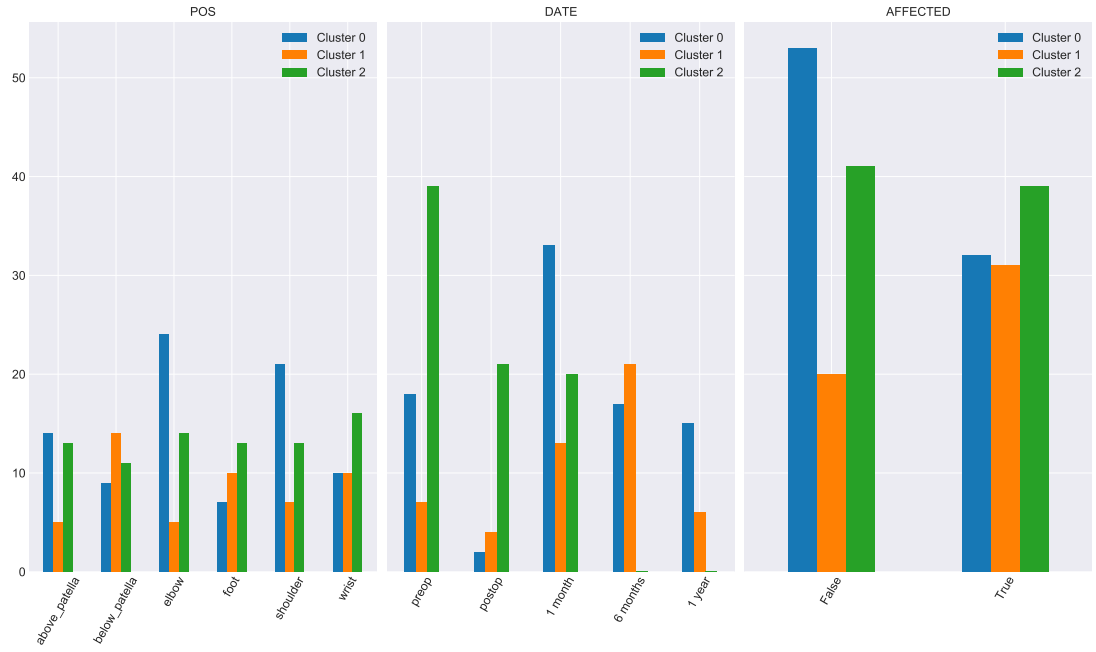


Figure 4.5: Categorical features histogram distribution for KMeans algorithm performed on 3 out of 4 clusters.

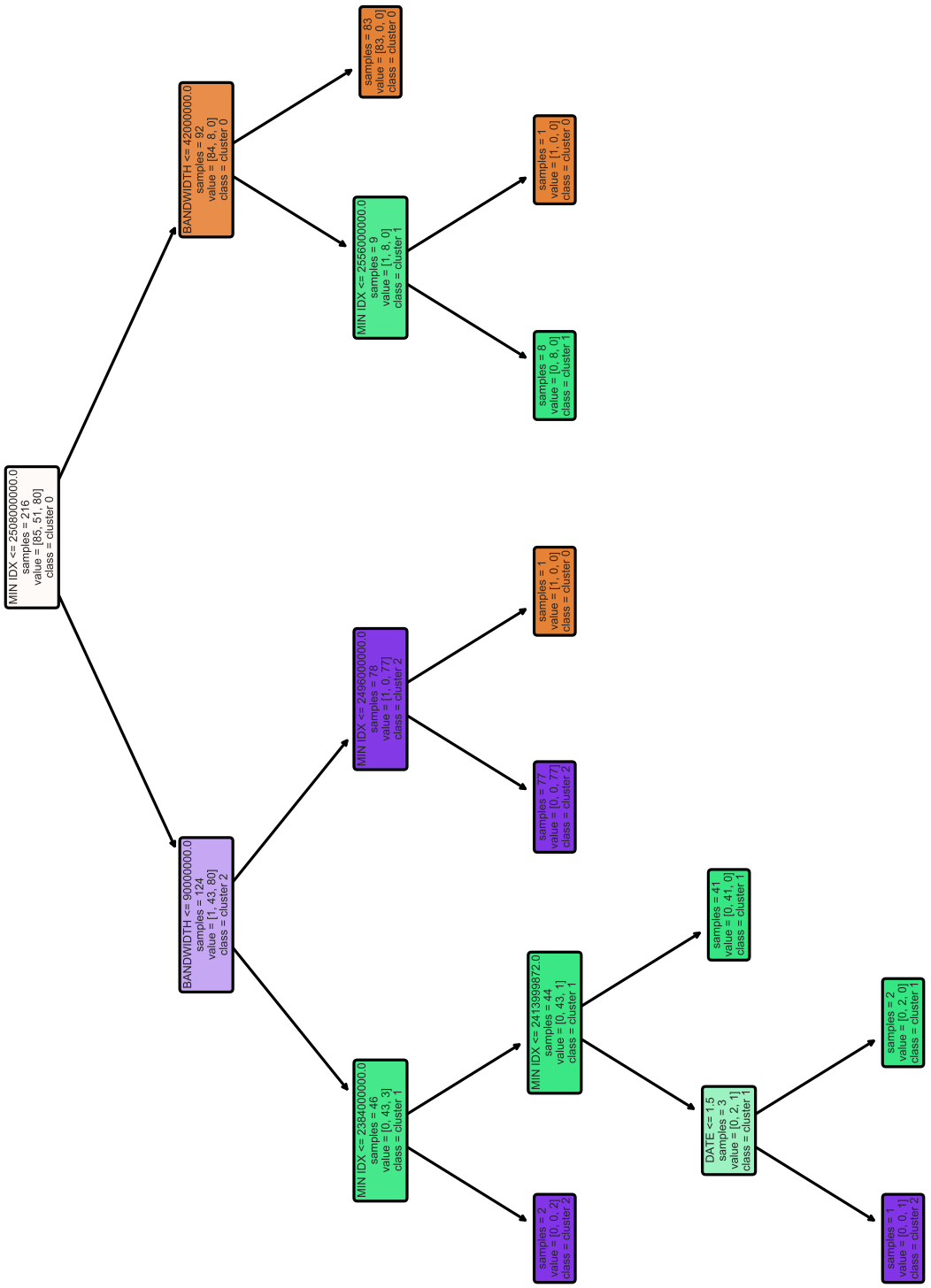


Figure 4.6: Output of the decision tree algorithm using four clusters obtained with KMeans.

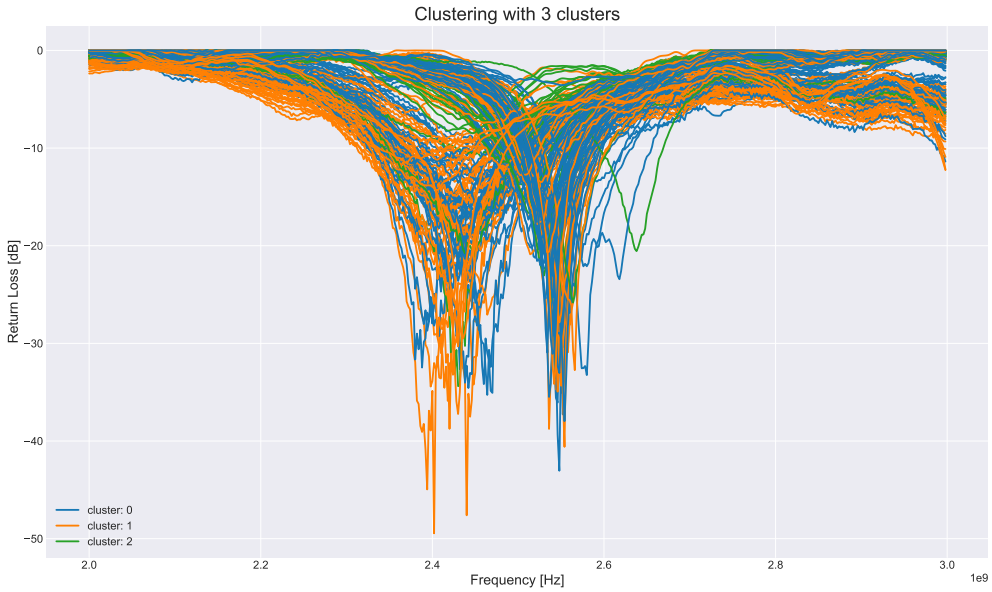


Figure 4.7: All curves clustered in different colors for KModes algorithm performed on 3 out of 3 clusters.

Now the results obtained from KModes algorithm are analyzed: as already mentioned, the supposed enhancement consists in the fact that also categorical variables are taken into account. In this case, instead of studying the case in which the input number of cluster is four as done before, the situation under investigation is with 3 out of 3 input clusters; this is done for two reasons: first, the final number of clusters is the same and a similar comparison could be made; secondly, this case is one of the best scenarios obtained by this algorithm.

From Figure 4.7 it is immediately evident that the clustering process is not anymore effective on continuous features, resulting in a more confused and overlapping cloud of curves. In order to examine better the peak distribution Figure 4.8 is reported: as expected the mean values are very close to each other and the standard deviation is practically covering the same regions.

Predictably, the feature importance extracted from the DT (Figure 4.9) classifier is much wider: not only categorical features were considered, but all variables are now having a small weight on the final output. For this situation the `DATE` feature is the most significant one, having almost 30% of importance, followed by the others, distributed around 15% and 6%.

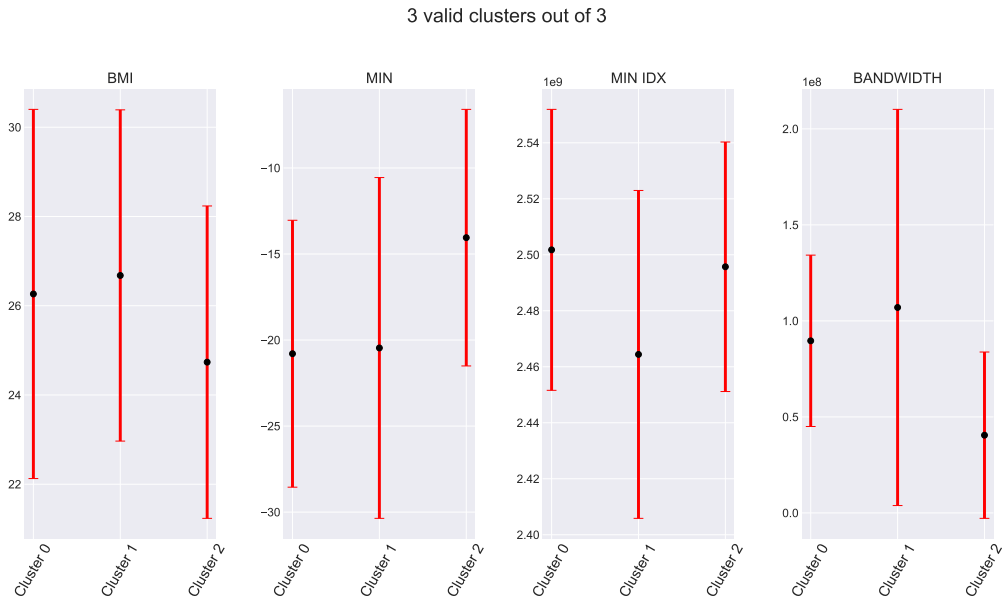


Figure 4.8: Main continuous parameter distributions for a clustering using KModes algorithm performed on 3 out of 3 clusters.

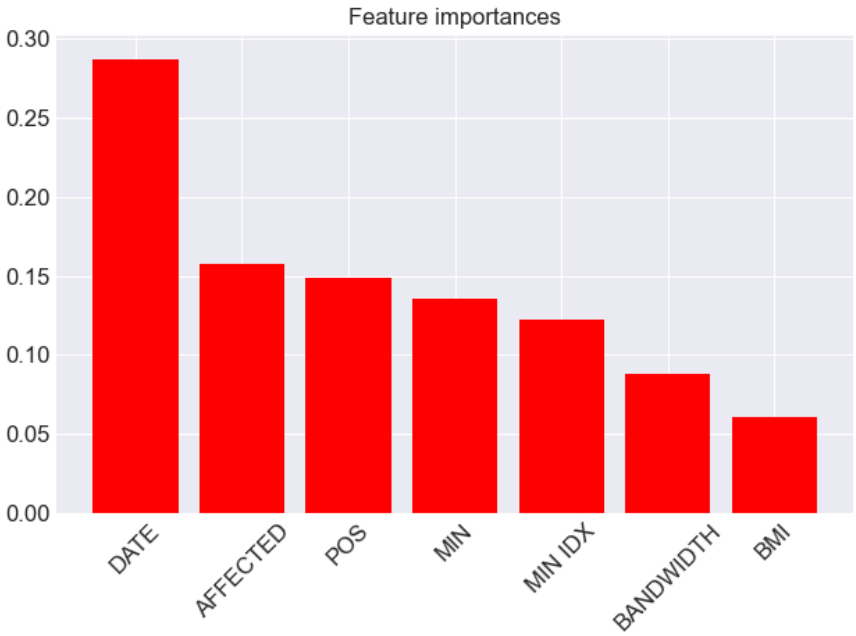


Figure 4.9: Feature importance of KModes algorithm performed on 3 out of 3 clusters.

As a consequence of a more superimposed behaviour of curves, Figure 4.10 highlights how the mean value of each cluster is not having anymore a clear peak: especially considering cluster 1, in which two groups of curves are present, each characterized by a different peak, resulting in an average curve with two local minima. Nevertheless, if the distribution of categorical features is now considered, histograms in Figure 4.11 show peaks for different clusters except for the measurement body position, which seems uninformative and, therefore, equally distributed. If the `DATE` feature is analyzed, it is possible to notice how distribution of cluster 1 is centered on *preop* and *postop* readings, while cluster 0 on *1 month*; cluster 2 is quite different because of the less amount of data belonging to it, but it is visible how most of its curves are located on *6 months* and *1 year*. Interestingly, also `AFFECTED` feature seems to be well distinguished: measurements belonging to cluster 0 and 2 are mainly done on affected limbs, those belonging to cluster 1 are done on reference limbs.

Finally, the DT is plotted in Figure 4.12: it is evident that the output in this case is more complex, with a lot more branches and dead leaves, even if the number of cluster is the same as before. This is because of the higher number of features taken in consideration for this case. Given the considerations done before, it may seem reasonable to eliminate few branches and, consequently, some readings, with the aim of removing those outlier curves. However, this should be done when enough data had been collected and there is certainty on not losing any valuable information.

In conclusion, in order to exploit gathered and processed data to predict the presence of the disease, Ensamble method was deployed. The core elements chosen for this algorithm are Logistic Regression, Random Forrest and Gaussian Naive Bayes as explained before; in particular, the second one was executed separately to compare the result with the one obtained from Ensamble. To do so a grid search was performed in parallel only to retrieve best parameters of the Random Forrest. The train and test division was done randomly taking 75% and 25% of the total 218 available data points, resulting in 163 and 55 data points respectively. Unluckily, for both cases the accuracy was around 50%, with a slightly higher percentage for the Ensamble method: the mean of accuracies is always around 50%, but in some cases it gets up to 60% depending on the grid search and the random seed used.

Another interesting behaviour should be noted regarding the output of the Ensamble algorithm classification. Since it allows to retrieve the probability associated to each prediction output, only those values with a probability higher than 70% were considered (frequently less 50% of all predicted values): almost the totality of these predictions were belonging to the class associated with reference measurements (absence of disease).

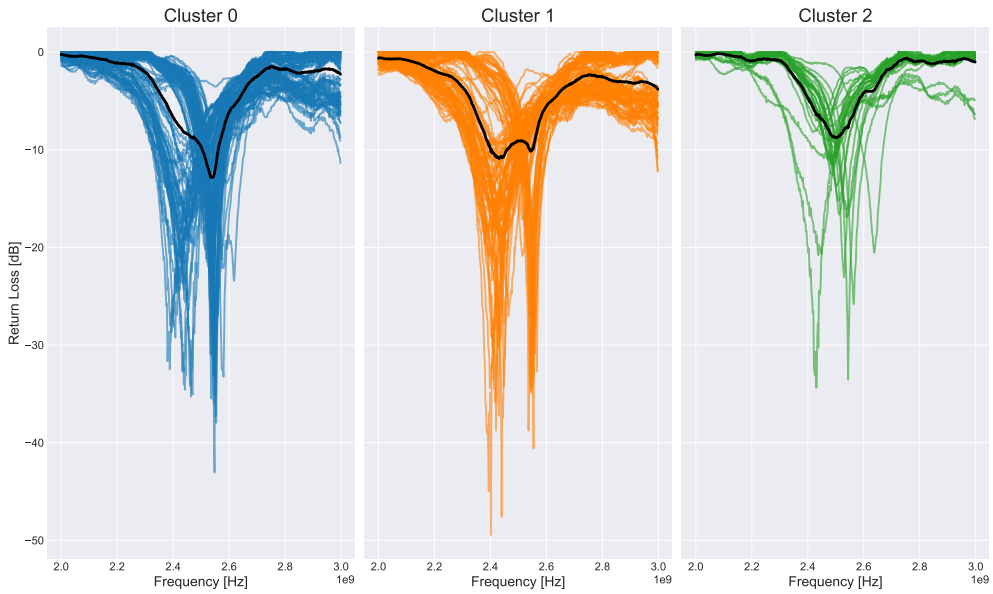


Figure 4.10: Each cluster with the corresponding mean value for KModes algorithm performed on 3 out of 3 clusters.

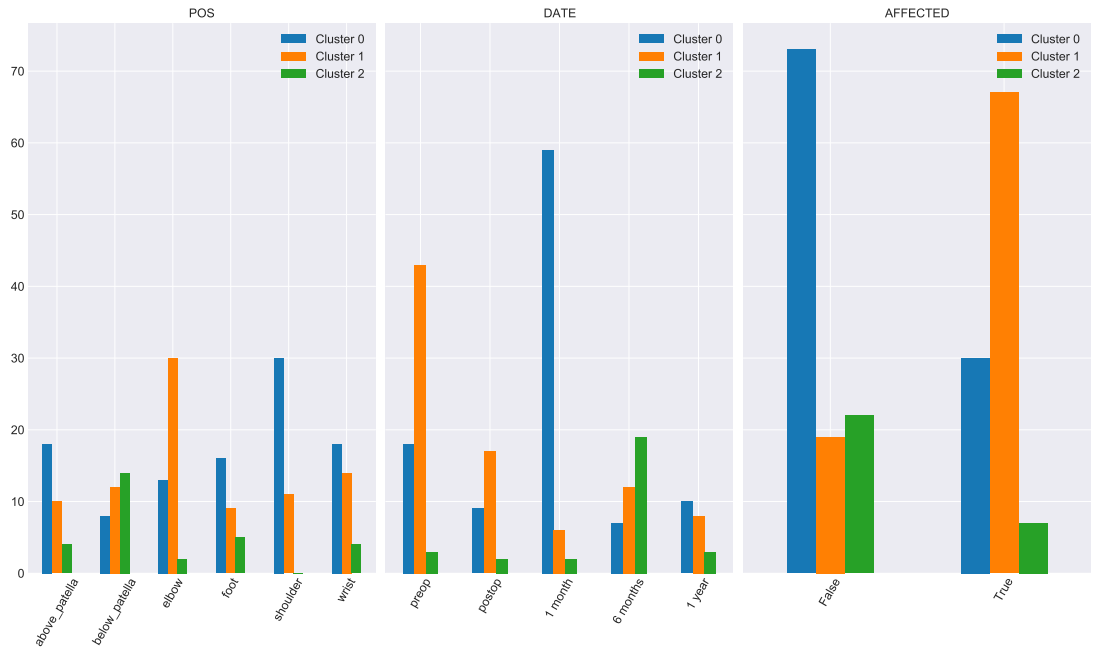


Figure 4.11: Categorical features histogram distribution for KModes algorithm performed on 3 out of 3 clusters.

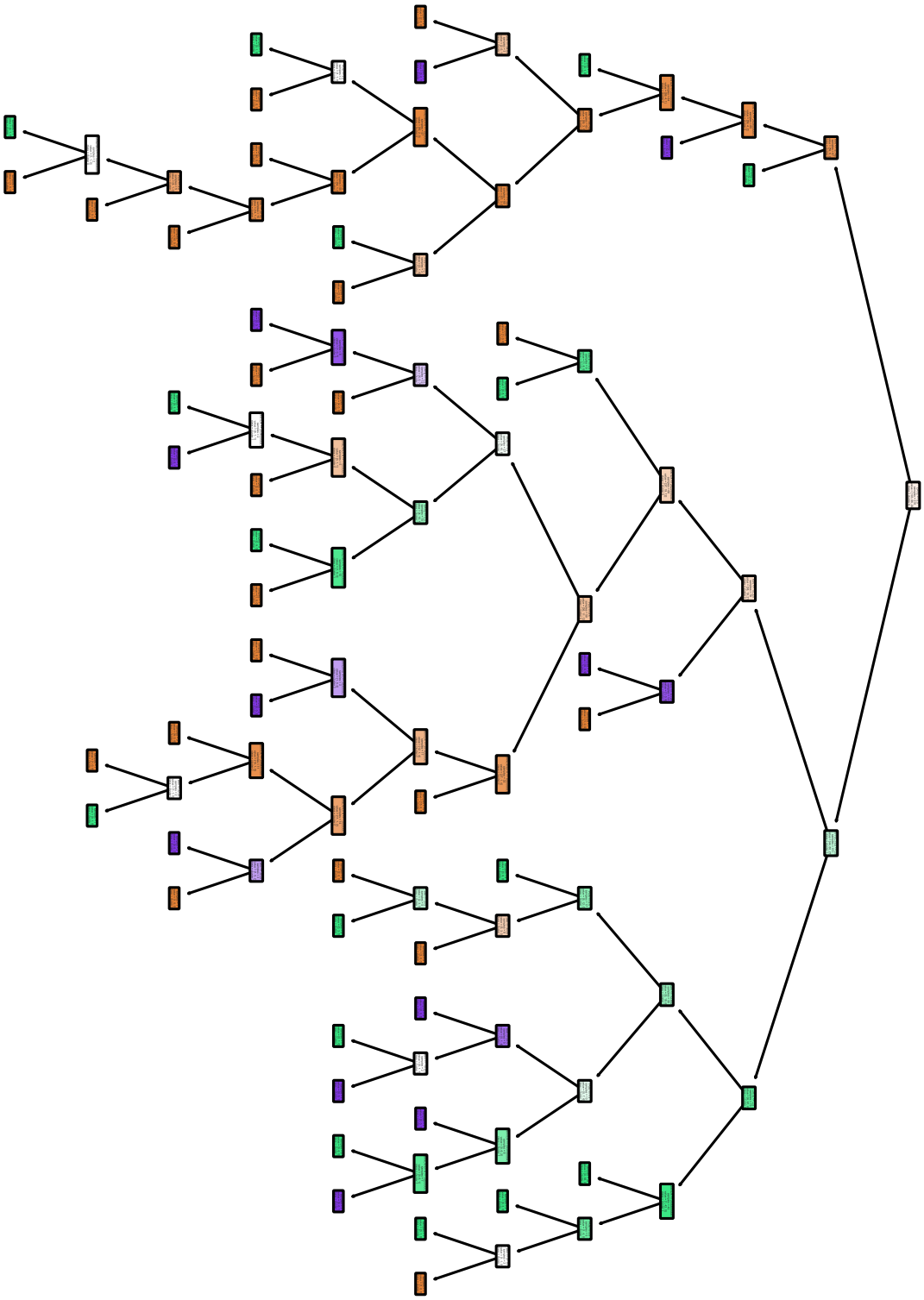


Figure 4.12: Output of the decision tree algorithm using three clusters obtained with KModes.

Chapter 5

Conclusions & Future works

The work done for this thesis is focused on the possibility of using non-invasive, cheap and simple MW equipment that would allow detection or extension of the LO disease. To do so, a dataset composed of SRR sensor measurements on LO patients has been investigated: the main target was to discover if any correlation between the disease and the measurements was present.

The first step was about the processing of data to improve the quality of dataset; as described in Section 3.1.3, these were the stages:

1. *Preparation*: where folders and file names were re-organized and uniformed.
2. *Standardization*: where RL curves having different frequency sampling were standardized, considering 500 frequencies in the 2 to 3 GHz region; Excel files were extracted merging curve data with information about patient's alias, body location and date of the measurement.
3. *Filtering*: where only data complying with certain conditions were kept for future processing; these conditions were related to (i) the number of zeros, (ii) to the prevalence of the peak and (iii) to the difference between readings building same measurement; in this section the average of three readings (or less, depending on the filtering itself) related to same measurement was evaluated.

The filtering process allowed to save 60% of the dataset. Nevertheless, as seen in detail in Section 3.1.4, few experiments were performed to assess the correctness of data: it was discovered that some readings had erroneous behaviours that could not be corrected with processing. In particular, after examining experiments' results, problems were associated to a broken cable and a wrong calibration procedure, frequently present in the dataset. To avoid these issues for future data collection, it follows a series of suggestions:

1. **Correct calibration**: the calibration process is fundamental and must be followed rigorously, i.e. putting the calibration loads attached to the copper cable and not to the miniVNA DUT port;

2. **Clean the surface**: first step needed is to clean and dry the surface over which the sensor will be laid;

3. **Manage instruments carefully**: noticing that the miniVNA and the equipment were carried in the same small plastic box, leading to an unnatural bending of the cable, the carefulness needed when handling the hardware must always be high;

4. **File naming and ordering**: the names of the files and their disposition in the folders must be chosen and followed throughout all measurements, possibly collecting patient metadata in the same (or at least similar) format;

5. **Environmental and default measurement**: it could be a very good practice to perform some reference measurements before those on the patient, e.g. measuring the air or same other material whose EM properties do not change; this could be used to assess the proper functioning of the miniVNA and sensor, and could help in the comparison of two different measurements done on different patients and in different scenarios (the environment influence in this way should be taken into account);

6. **Reference measurement**: it could be useful if more reference data were collected, even if those are not related to patient with LO; this would allow to better understand the range distribution of RL peaks and, potentially, improve the classification process;

In addition to these precautions, another crucial matter to be tackled is the effect of the skin: as throughfully explained in Section 3.2, where different scenarios were simulated thanks to CST Studio Suite software, the first material in direct contact with the SRR sensor has a huge impact on the measurement, determining where and with which amplitude will be located. Consequently, the skin thickness becomes a very important parameter, whose knowledge may be essential in order to extract certain information from real measurements. Unfortunately, although many simulations were performed, it was not possible to obtain a curve distribution similar to the one of real data; the reasons given to explain it are the following:

1. **Materials are not enough heterogeneous**: their properties are uniformly distributed over the whole model and this may not be a good representation of reality, in which, beneath the area covered by sensor, different distributions may result in an overall unpredicted effect;

2. **Model is too simple**: similarly to the previous point, the usage of a three layered structure may not take in consideration all the smaller elements present in human body, like veins, nerves and so on;

3. **SRR antenna is not identical**: the sensor was designed to be identical, but the simulation may be affected by smaller inappreciable differences, resulting in a different output;

4. **Material EM parameters are not suitable for this kind of simulation**: the IFAC database used to determine materials parameters could not be perfectly suitable for this kind of study and more variations on the parameters should be examined;

Always regarding simulations, one possible future implementation could be their integration with real measurements. The idea should be to better investigate the human tissue structure, possibly exploiting other technologies, like the BIA, that could improve the 3D modeling and could allow to associate the real data to simulated ones. In this way it would also be possible to evaluate and comprehend the differences between simulated and measured RL curves. This process would eventually allow to find a one-to-one correspondence, leading to the possibility of simulating a huge quantity of data that could be used for the ML process.

Concerning the clustering performed in the latter part of this thesis described in Section 4.1, the main focus was to find any correlation between the available metadata of patients and the measurements carried out. As already highlighted previously in Section 3.1.5, the correlation with time distance from operation was only found discriminating between those really close to the operation (at most 1 month after) and the others (6 months and 1 year), having peak location closer to 2.4 GHz and 2.5 GHz respectively.

This was confirmed with results obtained for KMeans clustering algorithm, where the attempt of clustering 4 groups of curves allowed the identification of 3 robust and well-separated clusters. Even though one cluster was composed of only early measurements (up to 1 month), the histogram distribution of categorical variables did not highlight any peculiar pattern for body position and presence of the disease. Thus, to deeply investigate those categorical features, KModes algorithm was tested: results were almost the opposite, where the RL curves were not distinctly separated, but here it was possible to distinguish one cluster mainly characterized by reference measurements performed after one month and another one with affected readings carried out before and right after the operation.

In such a way to have a tool that could be used in the hospital and could help the classification of patient status, a Decision Tree classifier was used based on the output of these two clustering algorithms. This was done also to evaluate the feature importance and their weight on determining which cluster a curve is associated to. For the KMeans algorithm, the extracted tree is readable and simple to understand, mainly because categorical features are not considered. On the opposite, the tree derived from KModes is incredibly chaotic and big; it would deserve a better investigation if the output of the clustering was sufficiently descriptive and useful, but since a clear behaviour associated to each cluster was not found, few considerations are needed.

Before presenting them it is necessary to discuss results obtained with the Ensamble method. As described in the final part of Section 4.3, these were not promising: an average of 50-55% of accuracy for a binary class is not sufficient to consider the classification process effective. Different ML techniques were tested but the prediction was not improving; despite it might be thought that this problem could not be solved with the proposed method, it should also be taken in consideration that the available data were not enough to perform a proper classification (~200 data points is a meager value for any ML technique). Besides this, another interesting element is the difficulty of predicting affected measurements: almost any of the correct predictions with a confidence higher than 70% are classified as *not affected*.

These highlighted points lead to the following considerations:

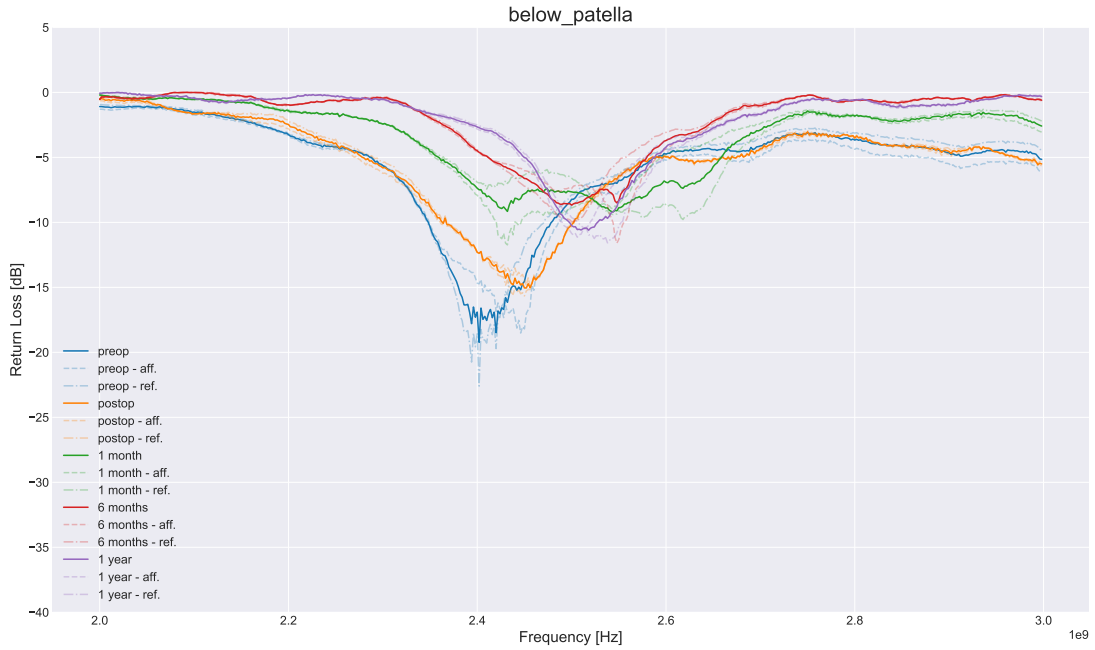
1. **Metadata are not enough**: categorical features used for this clustering may not be the ones that could be actually useful for this kind of process; BMI, for example, may not be the the best parameter to analyze or to be associated with the disease. On the contrary, evolution of the illness may be characterized by the circumference and/or its evolution in time.

2. **Presence of outliers**: despite a meticulous processing performed, the training of these ML techniques was done on a dataset containing few outliers that could not be eliminated; in particular, the fact that an erroneous calibration led to a valid curve but with a peak whose position was not exact could have prejudiced the classification.

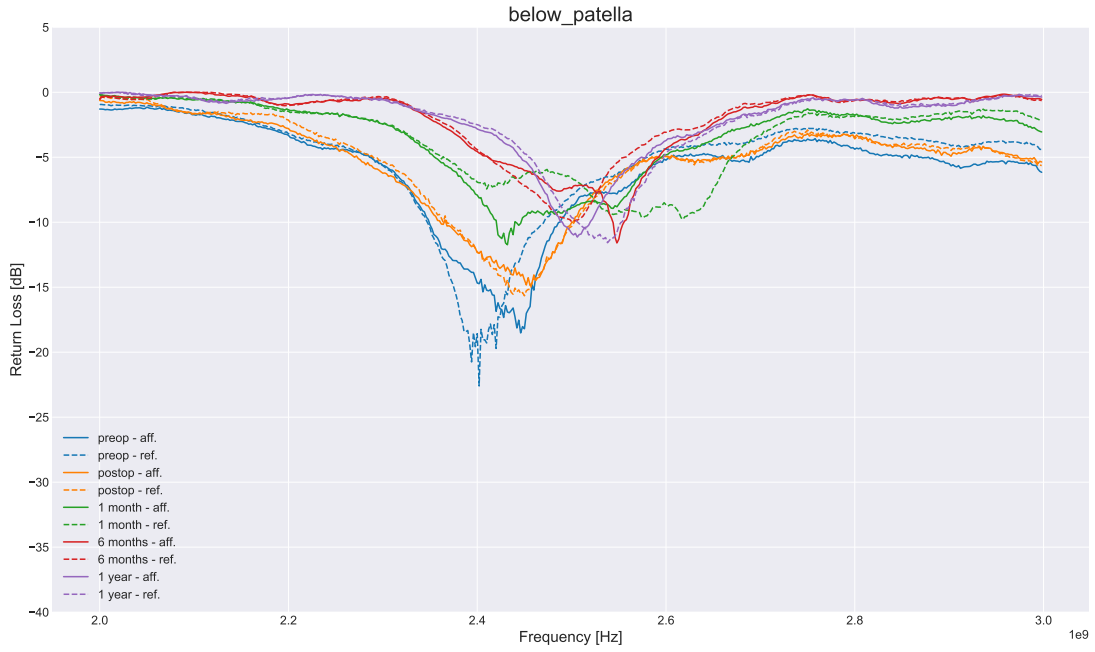
3. **Affected measurements are difficult to predict**: probably due to the more heterogeneous composition of measured area, the affected RL curves are more sparse and a clear pattern seems more difficult to be detected.

Appendix A

Extra Plots

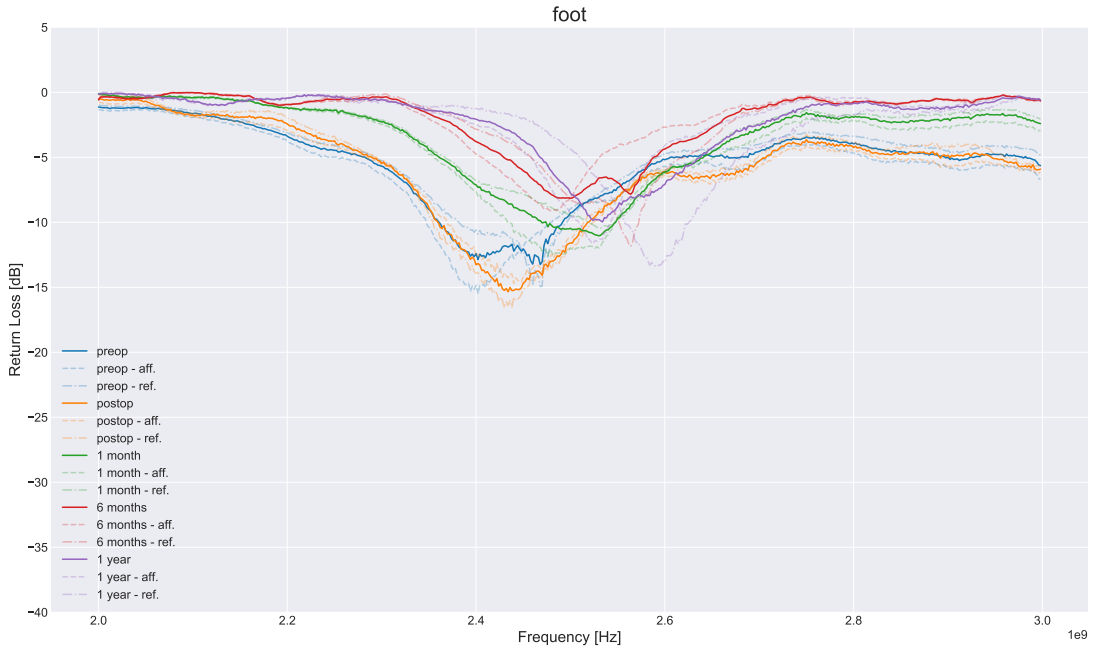


(a) Below patella measurements

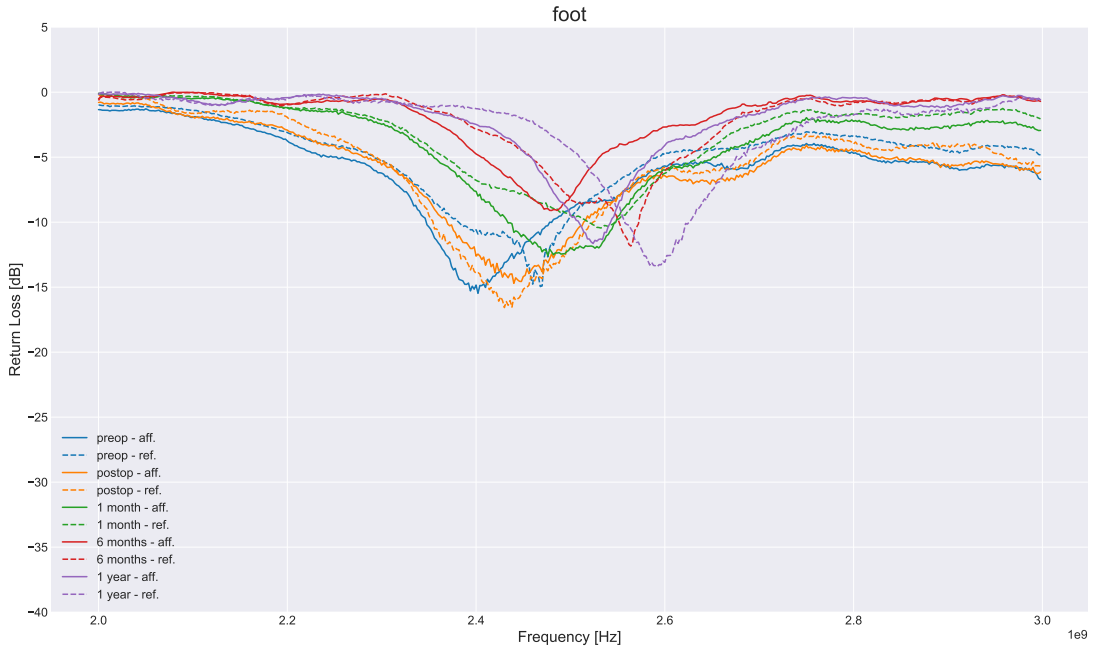


(b) Below patella measurements - reference vs affected

Figure A.1: Evolution plots (a), also considering ref. vs aff. measurements (b) for below patella measurements.

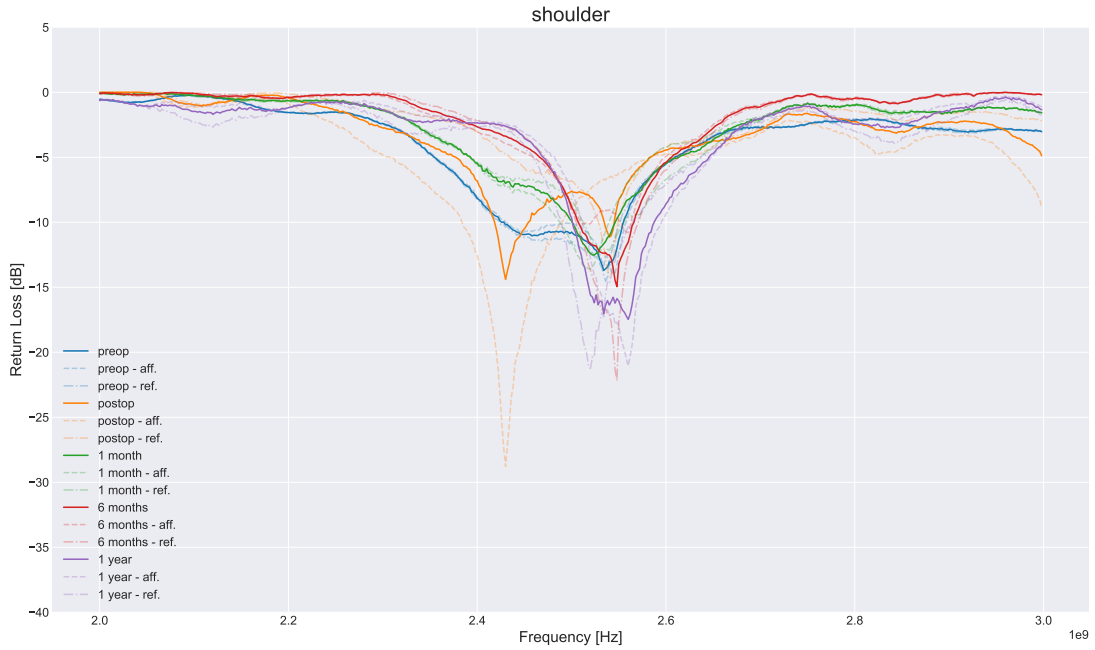


(a) Foot measurements

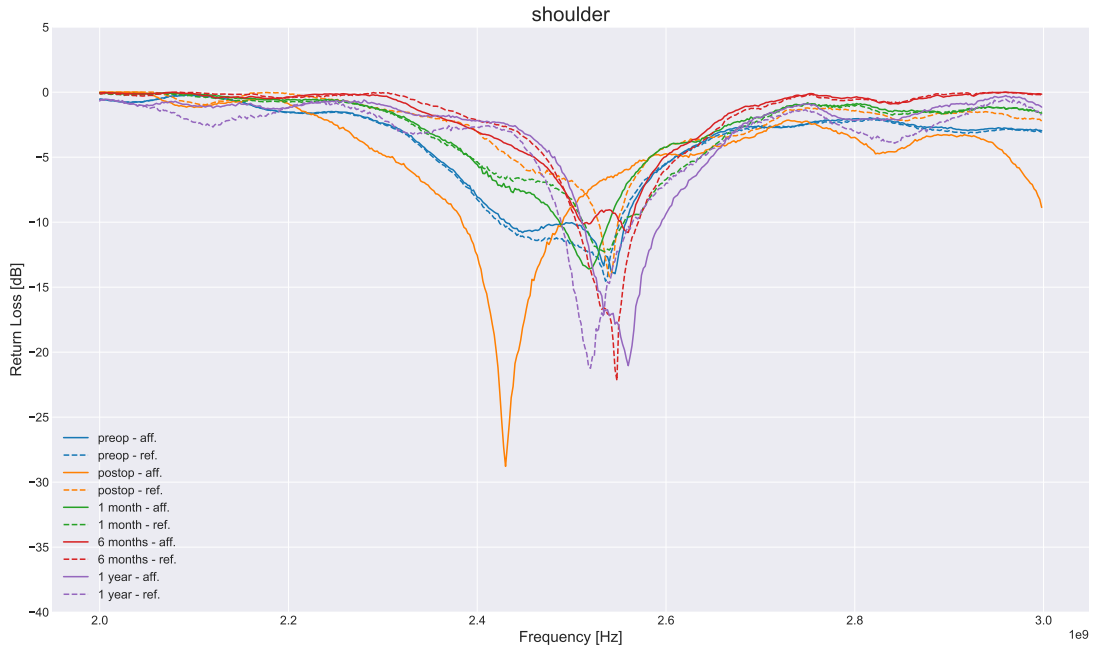


(b) Foot measurements - reference vs affected

Figure A.2: Evolution plots (a), also considering ref. vs aff. measurements (b) for foot measurements.

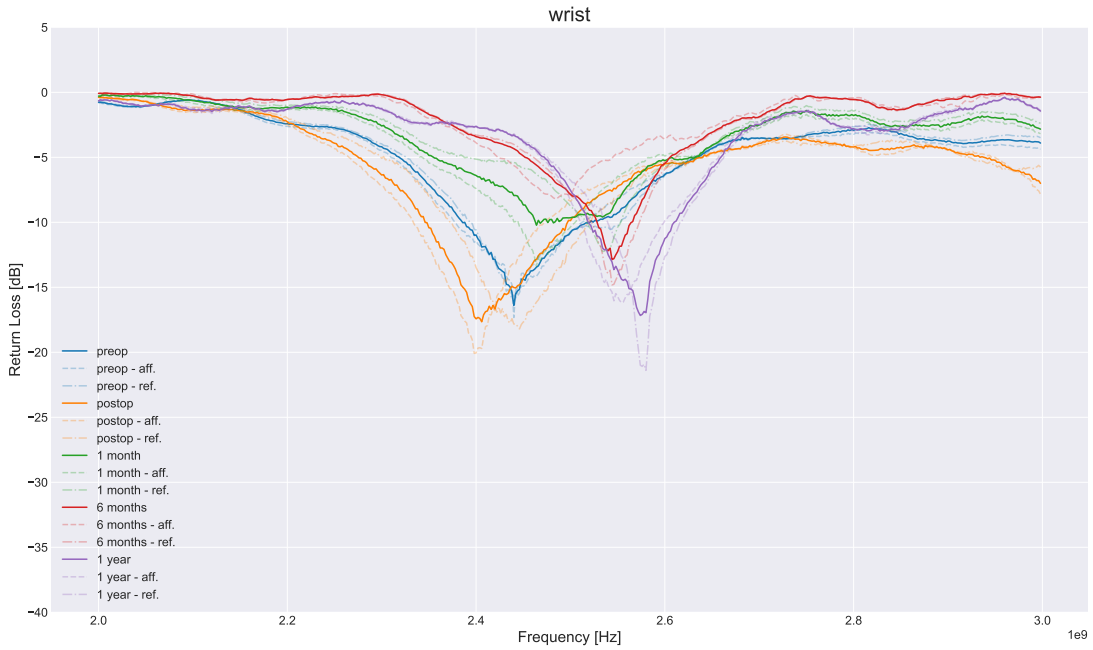


(a) Shoulder measurements

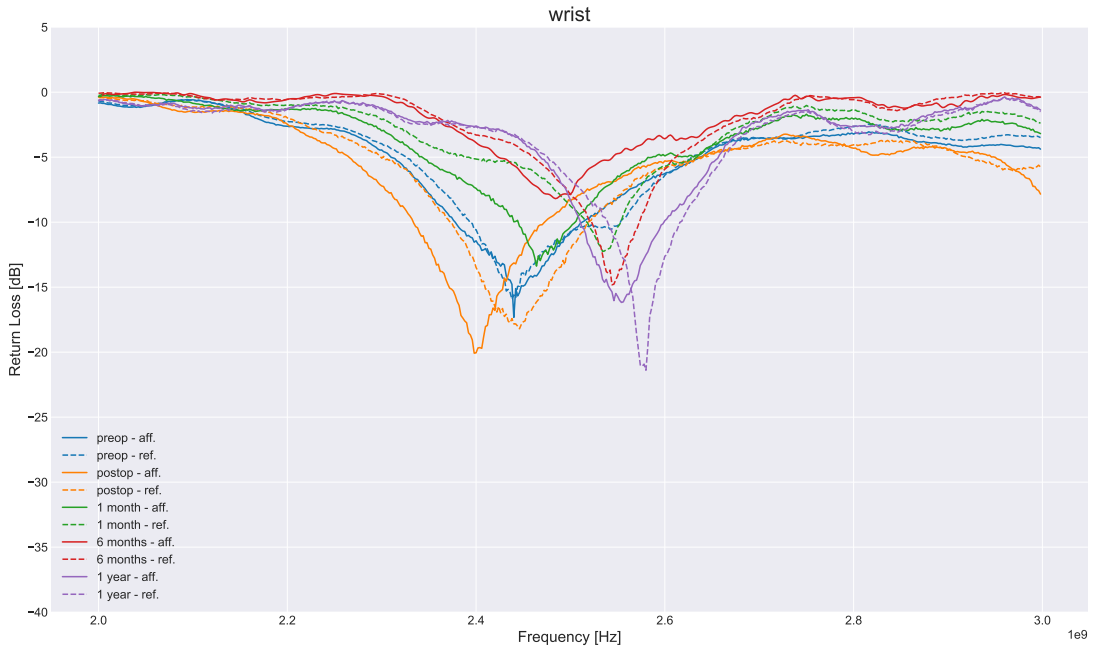


(b) Shoulder measurements - reference vs affected

Figure A.3: Evolution plots (a), also considering ref. vs aff. measurements (b) for shoulder measurements.



(a) Wrist measurements



(b) Wrist measurements - reference vs affected

Figure A.4: Evolution plots (a), also considering ref. vs aff. measurements (b) for wrist measurements.

Bibliography

[1] E. Joos, P. Bourgeois, and J.P. Famaey. «Lymphatic disorders in rheumatoid arthritis». In: *Seminars in Arthritis and Rheumatism* 22.6 (1993), pp. 392–398. ISSN: 0049-0172. DOI: https://doi.org/10.1016/S0049-0172(05)80031-9. URL: https://www.sciencedirect.com/science/article/pii/S0049017205800319 (cit. on p. 1).

[2] Stanley G Rockson and Kahealani K Rivera. «Estimating the population burden of lymphedema». In: *Annals of the New York Academy of Sciences* 1131.1 (2008), pp. 147–154 (cit. on p. 1).

[3] R. DiSipio, S. Rye, and al. «Incidence of unilateral arm lymphoedema after breast cancer: a systematic review and meta-analysis». In: *The Lancet. Oncology*. Vol. 14. 6. 2013, pp. 500–515 (cit. on pp. 1, 8).

[4] Mark V. Schaverien, Malke Asaad, Jesse C. Selber, Jun Liu, Dawn N. Chen, Melissa S. Hall, and Charles E. Butler. «Outcomes of Vascularized Lymph Node Transplantation for the Treatment of Lymphedema». In: *Journal of the American College of Surgeons* (2021). ISSN: 1072-7515. DOI: https://doi.org/10.1016/j.jamcollsurg.2021.03.002. URL: https://www.sciencedirect.com/science/article/pii/S1072751521001721 (cit. on p. 1).

[5] Mei R. Fu, Sheila H. Ridner, Sophia H. Hu, Bob R. Stewart, Janice N. Cormier, and Jane M. Armer. «Psychosocial impact of lymphedema: a systematic review of literature from 2004 to 2011». In: *Psycho-Oncology* 22.7 (2013), pp. 1466–1484. DOI: https://doi.org/10.1002/pon.3201. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pon.3201. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pon.3201 (cit. on p. 1).

[6] Marten N. Basta, Justin P. Fox, Suhail K. Kanchwala, Liza C. Wu, Joseph M. Serletti, Stephen J. Kovach, Joshua Fosnot, and John P. Fischer. «Complicated breast cancer–related lymphedema: evaluating health care resource utilization and associated costs of management». In: *The American Journal of Surgery* 211.1 (2016), pp. 133–141. ISSN: 0002-9610. DOI: https://doi.org/10.

1016/j.amjsurg.2015.06.015. URL: https://www.sciencedirect.com/science/article/pii/S0002961015004286 (cit. on p. 1).

[7] Robert J. Damstra et al. «The Dutch lymphedema guidelines based on the International Classification of Functioning, Disability, and Health and the chronic care model». In: *Journal of Vascular Surgery: Venous and Lymphatic Disorders* 5.5 (2017), pp. 756–765. ISSN: 2213-333X. DOI: https://doi.org/10.1016/j.jvsv.2017.04.012. URL: https://www.sciencedirect.com/science/article/pii/S2213333X17302299 (cit. on p. 1).

[8] Fu MR et al. «L-dex ratio in detecting breast cancer-related lymphedema: reliability, sensitivity, and specificity». In: *Lymphology* 46.2 (2013), pp. 85–96. URL: https://pubmed.ncbi.nlm.nih.gov/24354107 (cit. on p. 2).

[9] G. Eysenbach. «What is e-health?» In: *J Med Internet Res* 3.2 (July 2001), e20. ISSN: 1438-8871. DOI: 10.2196/jmir.3.2.e20. URL: http://www.ncbi.nlm.nih.gov/pubmed/11720962 (cit. on p. 5).

[10] Eunil Park. «User acceptance of smart wearable devices: An expectation-confirmation model approach». In: *Telematics and Informatics* 47 (2020), p. 101318. ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2019.101318. URL: https://www.sciencedirect.com/science/article/pii/S073658531930810X (cit. on p. 6).

[11] Huseyin Yildirim and Amr M.T. Ali-Eldin. «A model for predicting user intention to use wearable IoT devices at the workplace». In: *Journal of King Saud University - Computer and Information Sciences* 31.4 (2019), pp. 497–505. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2018.03.001. URL: https://www.sciencedirect.com/science/article/pii/S1319157817304706 (cit. on p. 6).

[12] D. Ousaka, N. Sakano, and al. «A new approach to prevent critical cardiac accidents in athletes by real-time electrocardiographic tele-monitoring system: Initial trial in full marathon». In: *Journal of Cardiology Cases* 20.1 (2019), pp. 35–38. ISSN: 1878-5409. DOI: https://doi.org/10.1016/j.jccase.2019.03.008. URL: http://www.sciencedirect.com/science/article/pii/S1878540919300386 (cit. on p. 6).

[13] Xiaoqing Li, Yu Lu, Xianghua Fu, and Yingjian Qi. «Building the Internet of Things platform for smart maternal healthcare services with wearable devices and cloud computing». In: *Future Generation Computer Systems* 118 (2021), pp. 282–296. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2021.01.016. URL: http://www.sciencedirect.com/science/article/pii/S0167739X21000261 (cit. on p. 7).

[14] Haridimos Kondylakis et al. «Patient empowerment for cancer patients through a novel ICT infrastructure». In: *Journal of Biomedical Informatics* 101 (2020), p. 103342. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2019.103342. URL: http://www.sciencedirect.com/science/article/pii/S1532046419302618 (cit. on p. 8).

[15] H. Wiig and MA. Swartz. «Interstitial fluid and lymph formation and transport: physiological regulation and roles in inflammation and cancer». In: *Physiological reviews*. Vol. 92. 3. 2012, pp. 1005–60 (cit. on p. 8).

[16] SC. Hayes, M. Janda, and al. «Lymphedema following gynecological cancer: Results from a prospective, longitudinal cohort study on prevalence, incidence and risk factors». In: *Gynecologic oncology*. Vol. 146. 3. 2017, pp. 623–629 (cit. on p. 8).

[17] C. Balakrishnan, LM. Bradt, and al. «Lymphedema of the upper extremity following circumferential burns». In: *The Canadian journal of plastic surgery*. Vol. 12. 2. 2020, pp. 79–80 (cit. on p. 8).

[18] BC. Sleigh and B. Manna. *Lymphedema*. 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK537239 (visited on 07/22/2020) (cit. on p. 9).

[19] A. Baz, T. Hassan, and al. «Role of contrast enhanced MRI lymphangiography in evaluation of lower extremity lymphatic vessels for patients with primary lymphedema». In: *The Egyptian Journal of Radiology and Nuclear Medicine* 49.3 (2018), pp. 776–781. ISSN: 0378-603X. DOI: https://doi.org/10.1016/j.ejrnm.2018.06.005. URL: http://www.sciencedirect.com/science/article/pii/S0378603X18301505 (cit. on p. 9).

[20] M. Kim, DH. Suh, and al. «Identifying risk factors for occult lower extremity lymphedema using computed tomography in patients undergoing lymphadenectomy for gynecologic cancers». In: *Gynecologic Oncology* 144.1 (2017), pp. 153–158. ISSN: 0090-8258. DOI: https://doi.org/10.1016/j.ygyno.2016.10.037. URL: http://www.sciencedirect.com/science/article/pii/S0090825816315141 (cit. on p. 9).

[21] Ebbe Nyfors. «Industrial Microwave Sensors—A Review». In: *Subsurface Sensing Technologies and Applications* 1 (2000), p. 103342. ISSN: 1573-9317. DOI: https://doi.org/10.1023/A:1010118609079. URL: https://link.springer.com/article/10.1023/A:1010118609079 (cit. on p. 10).

[22] M. Persson, A. Fhager, and al. «Microwave-Based Stroke Diagnosis Making Global Prehospital Thrombolytic Treatment Possible». In: *IEEE Transactions on Biomedical Engineering* 61.11 (2014), pp. 2806–2817. DOI: 10.1109/TBME.2014.2330554 (cit. on p. 11).

[23] S. Redzwan. «Prospective Applications of Microwaves in Medicine : Microwave Sensors for Orthopedic Monitoring and Burn Depth Assessment». PhD thesis. Uppsala University, Solid State Electronics, 2019, p. 96. ISBN: 978-91-513-0753-4 (cit. on p. 11).

[24] Kenneth Jian Wei Tang, Candice Ke En Ang, Theodoros Constantinides, V. Rajinikanth, U. Rajendra Acharya, and Kang Hao Cheong. «Artificial Intelligence and Machine Learning in Emergency Medicine». In: *Biocybernetics and Biomedical Engineering* 41.1 (2021), pp. 156–172. ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2020.12.002. URL: http://www.sciencedirect.com/science/article/pii/S0208521620301431 (cit. on p. 11).

[25] Cameron R. Olsen, Robert J. Mentz, Kevin J. Anstrom, David Page, and Priyesh A. Patel. «Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure». In: *American Heart Journal* 229 (2020), pp. 1–17. ISSN: 0002-8703. DOI: https://doi.org/10.1016/j.ahj.2020.07.009. URL: http://www.sciencedirect.com/science/article/pii/S0002870320302155 (cit. on p. 11).

[26] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. «Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis». In: *Procedia Computer Science* 83 (2016). The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops, pp. 1064–1069. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.04.224. URL: http://www.sciencedirect.com/science/article/pii/S1877050916302575 (cit. on p. 11).

[27] K. P. Sinaga and M. Yang. «Unsupervised K-Means Clustering Algorithm». In: *IEEE Access* 8 (2020), pp. 80716–80727. DOI: 10.1109/ACCESS.2020.2988796 (cit. on p. 13).

[28] Zhexue Huang. «Clustering large data sets with mixed numeric and categorical values». In: *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 1997, pp. 21–34 (cit. on p. 13).

[29] A. N. Reddy and S. Raghavan. «Split ring resonator and its evolved structures over the past decade: This paper discusses the nuances of the most celebrated composite particle (split-ring resonator) with which novel artificial structured materials (called metamaterials) are built». In: *IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*. 2013, pp. 625–629 (cit. on p. 16).