University of Coimbra - Faculty of Sciences and Technology

Master's degree in Informatics Engineering

Dissertation: Final Report

Prediction Model for Women Breast Cancer Recurrence

Bruno Filipe Aveleira Andrade

(bandrade@student.dei.uc.pt)

Advisor:

Professor Doctor Pedro Henriques Abreu (pha@dei.uc.pt)

Co-Advisor:

Professor Doctor Daniel Castro Silva (dcs@dei.uc.pt)

2014 / 2015

"Science is much more
than a body of knowledge,
it's a way of thinking"

- Carl Sagan, 1996

Abstract

Breast Cancer (BC) is the second most frequently diagnosed cancer and the fifth cause of cancer mortality worldwide. Among women, it is the leading cause of cancer deaths, with more than 500 000 registered deaths in 2012, and Portugal also reflects that reality. Survival prediction plays a crucial role in diseases with associated high mortality rates, since it has the power to help clinicians to define each patient's prognosis, thus allowing to personalize the corresponding treatments. Particularly for BC, prognosis is related to the patterns of recurrence (cancer that reappears after treatment), and it even differs depending on the local involved.

This work analyses the data of a cohort of 97 patients, with a total of 27 characteristics, more than 50% of them incomplete. Therefore, the first step is to handle Missing Data (Imputation or Deletion), to perform Classification afterwards. The purpose is to study the prognostic factors that define recurrence of female BC, to try to build a model that accurately predicts recurrence patterns, which would create the possibility of more targeted treatments.

The application of machine learning algorithms to the prediction of recurrence in different sites seems to be a novel application of these methodologies, and the results can lead the way to a better understanding of the pathways of BC recurrence.

Resumo

O Cancro de Mama (CM) é o segundo cancro mais diagnosticado no mundo, e o quinto com maior mortalidade. Nas mulheres é a maior causa de mortes relacionadas com cancro, com mais de 500 000 mortes registadas em 2012, e Portugal também reflete essa realidade. A análise de predição de sobrevivência tem um papel crucial em doenças com taxas de mortalidade elevadas, uma vez que tem o poder de ajudar médicos a definir melhor , permitindo a personalização dos tratamentos correspondentes. Particularmente no caso do CM, o prognóstico está relacionado com os padrões de recorrência (cancro que reaparece depois do tratamento), e até difere consoante o local afetado.

Este estudo analisa dados de um conjunto de 97 pacientes, com um total de 27 características, mais de 50% incompletas. O primeiro passo é portanto resolver o problema dos dados em falta (Imputar ou Apagar), para poder Classificar mais tarde. O objetivo é estudar os fatores prognósticos que definem a recorrência no CM feminino, para tentar construir um modelo que consiga prever corretamente os padrões de recorrência, o que traria a possibilidade de tratamentos mais direcionados.

A aplicação de algoritmos para a predição de recorrência em diferentes locais parece ser uma nova aplicação destas metodologias, e os resultados podem liderar o caminho para uma melhor compreensão dos mecanismos de recorrência do CM.

Contents

Acknowledgments

I relish this opportunity to thank all the people that helped me through this year of work. However, if only a few names will be mentioned, many other people also made this possible, not only during this thesis, but also for some years now.

Firstly, I want to thank my advisors, especially Professor Pedro Abreu and his continuous encouragement. His guidance helped me to overcome the challenges that my research kept throwing at me, and his enthusiasm in this work was invaluable.

The partnership with IPO-Porto was also extremely important, and I want to thank them for the opportunity to work with a real dataset from such a distinguished center. The visit to the facilities helped to give me the strength to proceed in an important part of the process. The direct contact with Dr. Miguel Abreu was also a source of privileged information about the work developed in IPO, and details about Breast Cancer research.

I am really grateful that I was not alone in this journey, and I thank my relatives for supporting me in my endeavors, especially my sister, Mariana Andrade, for staying home taking care of what was necessary while I was gone working. My "other" family was also very helpful in working through this year, in moments of relaxation that always seemed to short and to apart in time. Particularly, to Diogo Andrade (who even took me to the hospital one night) goes my apologies for a year of a little more distance, but also my sincere gratitude for all this years as a "brother from another mother", that I know I can count on and vice-versa. Finally, not so long ago, I have been lucky to meet one of the most kind, intelligent, and hard-working human beings I have ever encountered. So, I thank Miriam Santos for giving me the honor of sharing with her so many moments.

List of Abbreviations and Acronyms

AUC Area Under the receiver operating characteristic Curve

ANN Artificial Neural Network (ML algorithm)

AvAp Average Accuracy in percentage

AvAU Average AUC

BC Breast Cancer

CA15-3 Cancer Antigen 15-3

DM Data Mining

DT Decision Tree (ML algorithm)

ER Estrogen Receptor (protein)

HER2 Human Epidermal growth factor Receptor 2 (protein)

IHC ImmunoHistoChemistry (process for protein detection)

IPO[-Porto] Portuguese Institute of Oncology [regional center in Porto]

kNN k-Nearest Neighbors (ML algorithm)

L1QP L1 soft-margin minimization by quadratic programming (SVM routine)

MATLAB from "MATrix LABoratory" (programming language and tool)

MD Missing Data

ML Machine Learning

NB Naive Bayes (ML algorithm)

PR Progesterone Receptor (protein)

RBF Radial Basis Function

ROC Receiver Operating Characteristic

SEER Surveillance, Epidemiology, and End Results program

SMO Sequential Minimal Optimization (SVM routine)

SVM Support Vector Machine (ML algorithm)

WEKA Waikato Environment for Knowledge Analysis (programming tool)

WHO World Health Organization

List of Figures

List of Tables

Chapter 1

Introduction

This starting chapter is organized as follows. The first section pertains to the global theme, Breast Cancer, presenting an overall view of this disease as well as some statistics, and also mentioning a partner of this project. Section 1.2 shows the primary goals of this work, while Section 1.3 presents the time plan to accomplish them, and the mitigation strategies for possible risks are enunciated in Section 1.4 The last section contains the structure of this document.

1.1 Context

Breast Cancer (BC) is a major cause of concern worldwide. According to the latest statistics by GLOBOCAN [1], it was the second most frequently diagnosed cancer and the fifth cause of cancer mortality worldwide, responsible for 6.4% of all deaths. Among women, it is associated to the highest number of deaths due to cancer, with 521 907 registered deaths in 2012 [1]. Though predominantly in women, BC can also occur in men. However, male BC is rare: it represents less than 1% of all cases [2]. Further references to BC will pertain to female BC except where noted, since it is what this work will focus in.

Portugal follows these global trends, with BC being among the top three most frequently diagnosed cancers. Particularly for women, it was the cancer with highest rates of incidence and mortality. Solely in 2012, 6088 women were diagnosed with this disease, and 1570 died, which confirms the alarming scenario in Portugal [1]. According to WHO (World Health Organization) projections, these number are expected to rise, with 1620 deaths by BC predicted for 2015 [3].

In diseases with high mortality rates, such as this one, survival prediction assumes an important role, since it aids clinicians to better define each patient's prognosis and the corresponding treatments to be attempted. In particular for BC, prognosis is related to the patterns of recurrence [4]. Cancer Recurrence (or Relapse) describes cancer that reappears after treatment, and in the specific case of BC, recurrence is very common, being experienced by about one third of patients after initial diagnosis [4]. Therefore, establishing the patterns of recurrence is a crucial task to accurately predict the clinical behavior of this pathology. This enables a more personalized treatment for the patients, avoiding undesired overtreatment and adverse complications.

Despite the considerable advances in the study of BC in the last couple of decades, the underlying processes of recurrence have not yet been completely understood [5]. Encompassed in this reality, this work conducts a data-driven research, attempting to construct a model of recurrence for patients with this condition. As

detailed in the following section, our goal is to study the prognostic factors that define female BC recurrence, clarify the correlation between such factors and relapse patterns, and lastly, to provide a model to predict recurrence for a particular patient, based on her personal characteristics as well as her tumor expression.

Portuguese Institute of Oncology

The *Instituto Português de Oncologia* (Portuguese Institute of Oncology), known as IPO, is a corporate public entity with the mission of providing timely patient-centered health care services, focusing on treatment, prevention and research in Oncology [6]. IPO is a tertiary cancer center in this area, meaning that its patients have already been diagnosed with this pathology when they are admitted.

The data for this work were obtained from IPO-Porto, one of three Regional Centers. This institution has been distinguished for its dynamism, top position in quality standards, development of highly credible scientific activity and quality of education in Oncology. IPO-Porto treats about 10 000 new oncology patients each year, from which around 1000 are BC patients [6]. It is the largest unit in Portugal regarding this latter number, and one of the top units in Europe. IPO-Porto is also a reference center for the highest level of Clinical Trials performed in Portugal [6].

1.2 Objectives

Our project aims to construct a model of metastatic BC. This is achieved by examining the behavior of BC relapses, in terms of the localization of the tumor and its other features. The primary goals of this work are the following:

1. **Evaluate the pattern of metastatic dissemination in patients with BC**
   The first objective is to understand how the relapses are physically distributed, and their respective characteristics. BC prognosis is related to recurrence, but it even differs according to the site affected, namely bone only, visceral non hepatic, and visceral hepatic. Therefore, it is important to assess the behavior of BC recurrence metastases.

2. **Establish the relation between the patterns of metastatic proliferation, patients characteristics and BC subtypes**
   After analyzing the metastatic spread of BC, it will be measured the correlation between these data and the characteristics of the patients and their tumors. BC subtypes are defined via Immunohistochemistry (IHC) studies, used to determine the tumor features. The purpose of this goal is to determine how these characteristics affect the patterns of BC recurrence.

3. **Build a model of BC recurrence**
   This work intends to define a recurrence pattern, based on the characteristics of both patients and tumors. To achieve this, we must construct a structure that generalizes the relations found in the previous goal. The fact that it is based on a real-world dataset means that this model may be able to support the decision-making process of clinicians, establishing more accurate predictions, following the paradigm of Personalized Medicine.

1.3 Planning

This section refers to the presentation of a time-planning diagram, prepared to guide our work during the year. In Figure 1.1, the time expected to complete each task can be compared to the real period spent to complete it.

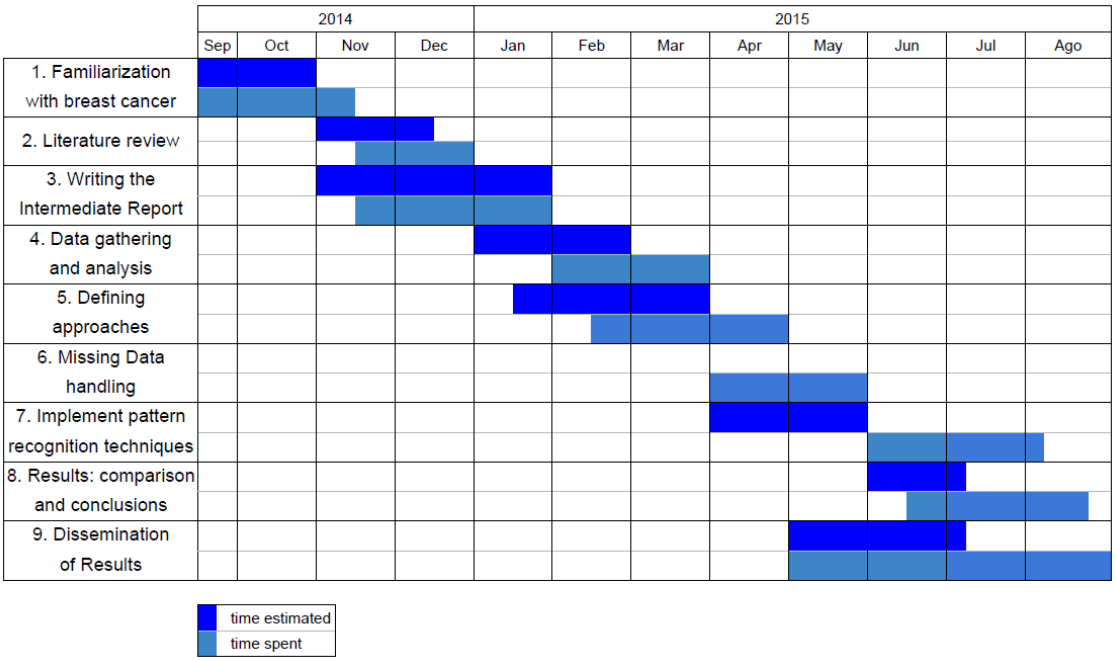

Figure 1.1: Work Scheduling

The diagram shows the work plan, divided into 9 tasks, namely:

1. **Familiarization with BC**
   Firstly, before the work was completely defined, some reading had to be done, to better understand the subject. This task consisted of searching and reading topic-related papers, to familiarize with cancer aspects, both medical and technical. Being a medical subject, it has to be considered a long period to assimilate ideas associated with this topic. This involved reading about BC, and the related terminology, and its recurrence, also having specific concepts associated with it.

   Since this work is being developed by an Informatics Engineering student, whose background is not in the medical field, this was a very time-consuming task, requiring much effort, and taking even more time than initially planned.

2. **Literature review**
   There has been an increase in awareness regarding BC. However, though there is much information about scattered in the Internet, it was necessary to study several scientific papers to understand the work developed in this area with detail. It is always required to study the state of the art, and in this case, the analysis of papers regarding BC relapses is an towards the correct implementation of pattern recognition techniques, as proposed in this work. This task focused on the important step of reading articles dealing with BC recurrence, including several studies analyzing the metastatic behavior of BC.

3. **Writing the Intermediate Report**
   The purpose of this task is to summarize all the work developed during the first semester, writing the present report. It also presents the work planned for the second semester. This report is submitted at the end of January, followed by a public presentation, which will take place in the beginning of February.

4. **Data gathering and analysis**
   The data used in this work were received from IPO-Porto, containing patients' information. These data not only characterize the patients but also their malignancies. Patients' information included variables such as age, gender and ethnicity, while the tumor is characterized in terms of subtype or site of metastases, among others.

5. **Defining approaches**
   Due to the complex nature of BC, and cancer tumors in general, we need to determine the aspects in which we will focus our analyses. As such, BC recurrence site was chosen as the primary splitting point, which guided the division of sections in this report.

6. **Missing Data handling**
   This task was not initially planned. However, it is accounted for in Section 1.4 (Task 4), since we know this could be a potential problem. The task involves reading articles about BC where Missing Data (MD) imputation methods were used. Afterwards, the necessary code is developed to impute MD in our dataset, including a simulation with the originally complete variables.

7. **Implementing pattern recognition techniques**
   This step involves the use of several approaches to extract correlation information from data. As explained in chapter 4, the goal is to find links between the different variables, establish relations between characteristics of the patients and tumors features, and among each of this groups.

8. **Results: comparison and conclusions**
   After the approaches are implemented and tested, it is necessary to evaluate and analyze the results. The purpose of this task is to draw conclusions from the work developed in the previous ones.

9. **Dissemination of Results**
   The thesis' Final Report is written in this task, including all of the work developed during the year, in both semesters. Moreover, a scientific article is also be produced.

In the first semester, the primary differences between the planned and real times concern the time given to understand BC recurrence. Since the original disease is already a complex pathology, understanding its process of recurrence is even more time-consuming. Therefore, task 1 took a longer period than expected.

1.4 Risk Analysis and Mitigation

As with any project planning, there are associated **risks**. This section presents the process of developing options and actions to reduce the potential impact of those

threats to the goals of this work. In case this events occur, they may jeopardize the entire project, or at least delay its execution and reduce its quality. This shows the importance of trying to prevent these incidents, or at least prepare a backup-strategy.

To achieve the proposed, the planning phases will be analyzed, assessing the possible risks for the individual tasks of the project (in less formal terms, "what could go wrong" with each one). For each risk, the **Mitigation Strategy** (to circumvent the problem at hand) proposed will be presented, and in some cases, the **preventive steps** (to avoid these risks) will also be indicated. This way, all the stages of the project are covered, and the risks are organized in an structured manner.

The first two tasks are "Familiarization with BC" and "Literature Review". Both of them consisted of reading and compiling information about the previous work developed, respecting the subject of this thesis. These were the risks found in the analysis:

1. [**Task 1**] **BC recurrence papers are too specific**
   The construction of a solid medical state of art depends on the existence of papers in this area. Although there are many proven developments in BC, the same is not verified when dealing with its recurrence phenomena. Even when such information is found, it is often too specific, and its understanding becomes severely difficult without a medical background.

   Impact: Medium

   Mitigation Strategy: In addition to the available papers about BC recurrence, it is important to read about the primary disease itself.

2. [**Task 2**] **Techniques not applied in the medical context**
   Several techniques that are intended to be implemented in this work have not been yet applied to the subject in study. Some of them may have not been used in the medical context at all.

   Impact: Medium

   Mitigation Strategy: To ensure the completeness of the analysis of existing methodologies, it might be necessary to study some papers of other areas.

The tasks for the second semester are more risk-prone. Since the work depends on external data and technologies, there are more possible sources of threats.

3. [**Task 4**] **Dataset not available**
   The data for this work is received from IPO-Porto. As previously explained, it is one of the most influential organizations of its kind, not only in the country, but also internationally. Besides the study of BC, it has a huge reputation in clinical trials too. Moreover, there is a team of several doctors dedicated to this task. The prevention consists in using a dataset from such a reliable source, compiled by a team of multiple doctors.

   Impact: High

   Mitigation Strategy: If the dataset from IPO-Porto is not provided for our study, there are others available on the internet. One example is the SEER Research website (Surveillance, Epidemiology, and End Results program), in which a dataset from the United States can be requested [7].

4. **[Task 4] Dataset requiring preprocessing**
   When the dataset arrives from IPO-Porto, it may need previous preparation. While the multiple doctors involved add a layer of trust to the data gathering process, there can always be problems. Problems in data values can consist of noise, contradictions and missing values, among others. Furthermore, the attributes can be irrelevant, or its values can be imbalanced,for example.

   Impact: Medium

   Mitigation Strategy: These problems are addressed in chapter 4. Should they be noted when the dataset is received, the preprocessing tasks are already prepared. Some examples are the elimination of attributes/patients, the normalization, the imputation (estimation) of missing values.

5. **[Tasks 4 and 5] Dataset delivery is delayed**
   If the dataset doesn't arrive on time, it is not needed to apply the mitigation strategy immediately. While the project goals can be achieved, we may tolerate some level of delay. However, this change has the potential to delay the whole project.

   Impact: Medium

   Mitigation Strategy: Before we receive the data, there is some preparation work that may be done regarding the methodology. Although we don't know the exact problems we will have, just like this risk analysis, it is possible to anticipate the foreseeable situations, providing alternatives to prevent them. The Data Mining (DM) approaches can be previously enumerated, as well as the preprocessing methods and validation tasks. When the data arrives, is it only needed to choose the approach according to its characteristics, but the possibilities are already defined. If the delay is too long, we consider the dataset as "not available" (risk 3).

The work planned for the second semester also includes the implementation phase. Regarding this step, the following risks were found:

6. **[Task 7] Algorithms are too complex**
   Some of the computational techniques used in this work may have the potential to lose computability. For example, in Neural Networks (one of the possible techniques), there are many possibilities of variation: learning rate, learning function, activation function, number of hidden (virtual) neurons, among others.

   As a preventive step, the search for good results is not a brute-force application of all the configurations of the methods proposed. Instead, some possibilities can be tested beforehand, and subsequent trials will be focused on variations of specific configurations (based on the analysis of the previous).

   Impact: High

   Mitigation Strategy: Using the example of Neural Networks, and more concretely, the number of virtual neurons used in its configuration, we may choose to use numbers with a certain interval x (in our implementation). If the time doesn't allow the use of many values, it is possible to increase this interval, and test less possibilities, while still covering the range intended. It is also possible to focus the attention in the best-performing techniques, increasing our efforts to optimize these algorithms.

7. **[Task 7] Algorithms' code is not available**
   Several techniques are studied during this work. The existence of theoretical explanations, or even previous work, doesn't guarantee that their implementations are available. To prevent this situation, *MATLAB* will be the programming language used. The tool with the same name is proprietary software from *Mathworks*, for which we possess a license. This commercial tool provides a large number of toolboxes for different applications, with Neural Networks and Evolutionary Algorithms among the possibilities, for example.

   Impact: Medium

   Mitigation Strategy: The code for a certain implementation can be created for our work, if the approach is believed to be very important. In addition, if that is not even possible, there are other options, for example:

   - *WEKA:* Machine Learning (ML) tool (in Java);
   - *R:* statistical programming language;

1.5 Document Structure

The following chapters show the remainder of our work: Chapter 2 contains more detailed information regarding the main topics of this thesis: BC as a pathology, DM techniques, and evaluation metrics of classification systems. Chapter 3 reveals an analysis of recent related literature regarding BC Recurrence. This State of Art is divided into two sections, focusing on the recurrence sites on one hand, and DM techniques on the other. Chapter 4 presents the proposed approaches to be analyzed and compared, serving as a basis for further work to be developed. The results of such experiments are then exposed in Chapter 5. Finally, Chapter 6 summarizes the thesis, also providing possible directions for the continuation of this work.

Chapter 2

Background Knowledge

The information contained in this chapter represents the basis of all the work developed throughout this project, in two distinct areas: a clinical overview of BC as a disease (Section 2.1) and a technical explanation of DM methods (Section 2.2).

2.1 Breast Cancer

Cancer is the name given to the phenomenon of uncontrolled growth of abnormal cells. BC is the name given to malignant tumors that originate in the breast, hence the name. The most important statistics have already been mentioned in Section 1.1.

However, many patients that have BC do not have serious symptoms, or may associate fatigue and weight loss (possible cancer symptoms [8]) to a number of other causes (stress, different diet, less sleep). The mammogram, an X-ray image of the patient's breast, plays an important role in the early detection of BC, detecting cancer much before any symptoms show up. External signs of BC may include a lumps in the breast, or general changes. When a patient discovers an anomaly in the breast (via self-examination or in a doctor's appointment) or a mammogram reveals it, the suspicion of cancer appears. A biopsy is then performed, and a pathologist examines it to confirm the diagnosis, while radiology can be used to detect distant involvement in other organs by cancerous cells (metastases).

Invasive BC can be divided according to the starting local of the tumor inside the breast, and the two most frequent are ductal and lobular. These names originate in the names of the ducts, channels that carry the milk from the producing glands to the nipple, and the lobules, the glands themselves. Invasive ductal carcinomas begin in a duct of the breast and grows into the surrounding tissue. It is the most common form of BC, accounting for approximately 80% of invasive BC. Invasive lobular carcinomas start in the lobules, representing about 10% of invasive BC.

BC subtypes are a way of categorizing patients based on some important features of the tumors. The variables used to distinguish these subgroups are assessed in a chemical precess called immunohistochemistry (IHC), and represent the presence or absence of different protein in the tumor (respectively positive, $^+$, and negative, $^-$). Estrogen Receptors (ER) are receptors of the hormone Estrogen, while Progesterone Receptors (PR) are receptors of the hormone Progesterone. HER2 (human epidermal growth factor receptor 2) is another important protein, linked with the progression of BC tumors.

The most common distinction is shown in the following list, identified with the terminology used:

- Luminal: ER $^+$ or PR $^+$ (at least one of them) and HER $^-$

- HER2-enriched: HER^+
- Triple-negative: ER^-, PR^- and HER^-

Occasionally, a new subtype is considered for patients with ER^+ or PR^+ (at least one of them) and HER^+, called Luminal HER2. There can also be a distinction of Luminal patients based on a proliferation index, Ki-67, into Luminal A (Ki-67^-) and Luminal B (Ki-67^+) patients. This categorization of patients is regarded as the most probable explanation for why patients have different outcomes [9].

BC is commonly treated by one or several combinations of what has been mentioned before: surgery, radiation therapy, chemotherapy, and hormone therapy. The selection of therapy may be influenced by the the characteristics of the patient and those specific of the tumor, e.g.:

- Menopausal status of the patient
- Stage of the disease
- Grade of the primary tumor
- ER and PR status of the tumor
- HER2 overexpression
- Histological type

Adjuvant therapy for BC is any treatment given after the primary therapy: Chemotherapy is the use of drugs to try to kill malignant cells. Often, more than one drug is given during adjuvant chemotherapy; Hormonal therapy tries to block BC cells from receiving the hormone estrogen; Tamoxifen, for example, blocks estrogen's activity in the body. Trastuzumab is a targeted drug, focusing on cells that everexpress HER2; Radiation therapy is usually given after breast-conserving surgery and may be given after a mastectomy (it is a local therapy, while the others are systemic therapies, because they travel to the whole body through the bloodstream). Neoadjuvant therapy, on the other side, is given before the primary therapy, for example, to try to diminish the size of an inoperable tumor. With the advancements in the area of medical sciences, new medicines and therapies have been developed, bringing renovated hope to BC patients, including those with recurrence.

When relapse is diagnosed in a patient, the median survival time is expected to be between 1,5 and 2,5 years. It is extremely difficult to pinpoint the exact causes for the variation, and this range can even be a result of different characteristics of the patients included in each study. In spite of all this, some patients can survive several decades even after a relapse episode [10], which means that it is not the end of the road for these people. There are features associated with BC relapse, some of these variables are lymph node involvement, large tumors, low levels of ER and PR, and higher histological grade.

2.2 Data Mining

Data are everywhere, and the volume never stops increasing. As new repositories are created, new gathering methodologies are also developed, which keeps feeding this cycle. The hobby of photography is something that changed over the years. Instead

of having to own a dedicated camera, one can simply grab the smartphone (if not holding it already!) and take a picture. Nowadays, many synchronization services automatically save this into online storage space. And as long as new repository services and new technologies become more available, the amount of data keeps growing.

In the medical field, there is also a constant search for new techniques to capture data about the patients. Whether it is a wearable accessory that monitors your heartbeat 24/7, or a new diagnostic method with High Definition three-dimensional resolution, all this adds to the toll. To extract information of this incommensurable world of "zeros and ones", it is necessary to develop intelligent computational ways of transforming these data into real human-understandable information. Without this process, all we get are values, while the (possibly useful) information remains hidden.

Data Mining (DM) is the answer for this problem, as it involves methodologies of Machine Learning (ML). This means that a computer will receive examples of data and try to understand the underlying patterns, thus getting the knowledge to predict future examples. Pattern recognition is natural to the human being, and even the "machine" part is not that new, but that are more opportunities to use them than ever. The goal of ML, more than simply compile the information about the examples seen, is to generalize for future data. The algorithms used in this work are Supervised, meaning that the system takes a known set of responses to the known input data (although some of them may also have Unsupervised versions). In Unsupervised algorithms, the predictor wouldn't know the response, and would try to "draw inferences" from the inputs [11]. Figure 2.1 shows the two steps of the process of Supervised Learning.

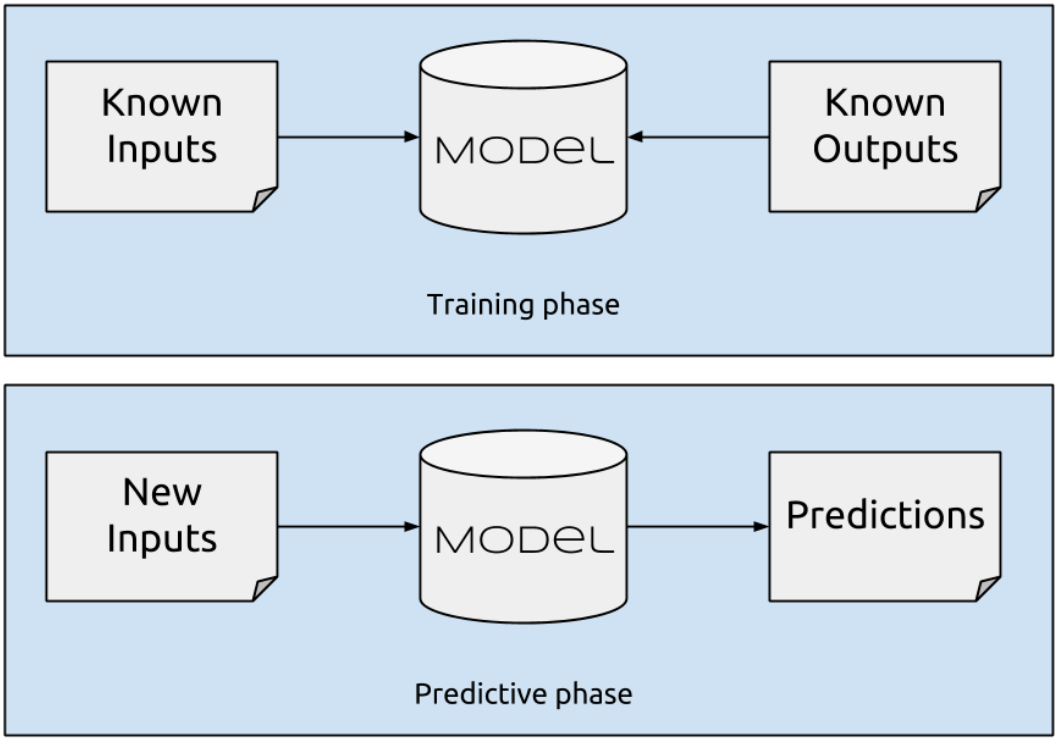

Figure 2.1: Depiction of the Supervised Learning process

There are many ML algorithms, and the following are the ones used in this work. The next subsections only present the computational techniques used, while

the implementation details of the both the imputation and the classification are explained in the next chapter.

2.2.1 k-Nearest Neighbors

This algorithm, also known as kNN, is based on the concept that similar examples should be associated with similar outputs. There is an unsupervised version [11], but the one used in this work is supervised. In *MATLAB*, there is even a direct function to impute, called knnimpute(), which replaces MD in a dataset using this algorithm, allowing to vary parameters such as the value of k, for instance. In theory, kNN starts by choosing the closest k examples in the training set to the new data, retrieving also their response values. The classification label for the new example is based on the labels of the different values. A different value of k will make the decision be based on more or less neighbors. Moreover, there are alternative ways of finding the closest points, instead of the basic euclidean distance. The ten available distance metrics are [12]:

- euclidean: Euclidean distance
- seuclidean: Standardized Euclidean distance; each distance is scaled
- cityblock: City block (or Manhattan) distance
- minkowski: Minkowski distance
- chebychev: Chebychev distance (maximum coordinate difference)
- cosine: One minus the cosine of the included angle
- correlation: One minus the sample linear correlation
- spearman: One minus the sample Spearman's rank correlation
- hamming: Hamming distance, percentage of coordinates that differ
- jaccard: One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ

Instead of focusing only on the most used distances, this study aims to compare all of them, to try to get the best possible result, both in imputation and classification.

2.2.2 Artificial Neural Networks

ANNs were created with the purpose of resembling how the brain works internally. In our case, the architecture we are going to use is the Multi-Layer Perceptron, in which the network transforms inputs in outputs by means of layers of neurons with weighted interconnections [13].

There is always an input layer (where data entries), and an output layer; additionally, there are intermediate layers with a variable number of nodes, called hidden layers. It has been stated [13, 14] that a neural network can approximate any function with only one layer of hidden nodes, as accurately as desired, as long as there are enough neurons there (it is a universal approximator). This way, we will vary the number of hidden nodes, but with only a single hidden layer. Another

parameter to vary in this algorithm is the training function, and there are thirteen choices:

- trainlm: Levenberg-Marquardt
- trainbr: Bayesian Regulation
- trainscg: Scaled Conjugate Gradient
- trainrp: Resilient BackPropagation
- trainbfg: Broyden–Fletcher–Goldfarb–Shanno quasi-Newton
- traincgb: Conjugate Gradient with Powell-Beale restarts
- traincgf: Conjugate Gradient with Fletcher-Reeves updates
- traincgp: Conjugate Gradient with Polak-Ribiere updates
- traingd: Gradient Descent
- traingda: Gradient Descent with adaptive (variable) learning rate
- traingdm: Gradient Descent with momentum
- traingdx: Gradient Descent with momentum and adaptive learning rate
- trainoss: One step secant

A comparison between them, spanning different problems, has been performed and is available online [15].

2.2.3 Decision Trees

Theoretically, DTs are a representation of classes through a series of yes/no questions. These correspond to the binary splits in the branches of the tree that represents such a model [16].

There were two parameters in this algorithm that we changed: the minimum leaf size (minLeaf), the minimum number of instances in a leaf (end of a branch, with the class of the instances that follow the path to it); and the criterion used to create the splits, namely:

- Gini's diversity index
- twoing rule
- deviance reduction

2.2.4 Support Vector Machines

Data are represented by a set of feature vectors. When the data have two classes, SVM can try to divide them with an hyperplane. To do this, the algorithm projects the examples in a higher-dimensional space, simplifying the problem of creating a division. SVM also tries to maximize the distance between this boundary and the training examples of either side (class). If they are not separable, it will try to separate most of them (called a soft-margin). Among the parameters that we changed is the Kernel Function:

- Linear
- Radial Basis Function (RBF)
- Polynomial (order 1)
- Polynomial (order 2)

Another parameter is the Optimization Routine (parameter 'Solver'):

- L1 soft-margin minimization by quadratic programming (L1QP)
- Sequential Minimal Optimization (SMO)

Some parameters of SVM may use subsampling (picking a subgroup of patients from the original group). This involves random processes, which lead us to restart the random generator to the same number before each imputation. This way we ensure that all imputations start from the same point of randomness, allowing the results to be replicable.

2.3 Conclusion

This chapter provides the fundamental knowledge required to understand the different steps of this thesis. It covers both the clinical and technological points of view. The terminology and core concepts of BC are explained in the first section, and clearly show the complexity of this pathology. The intricate details of relapse are also shown, particularly for the case of BC. It is more difficult for someone outside the medical community to understand, but the collaboration has the potential to be very rewarding. The DM section explains the theory of the computational aspects of this thesis.

Chapter 3

Literature Review

This chapter will present a selection of research papers about BC Recurrence, divided into two categories. While Section 3.1 contains articles with a focus on the different recurrence sites, Section 3.2 shows the use of predictive machine learning in BC recurrence. Our work fits in the two categories, but to the best of our knowledge, this was never attempted. The first section features clinical statistical studies in the first section, and none of the authors try to use machine learning approaches. On the other hand, the articles in the second section deal with recurrence as an atomic event. Nevertheless, these show some of the applications of Data Mining techniques to the study of recurrence of malignant breast tumors. Finally, a brief conclusion of this review work will be provided.

3.1 Recurrence Sites

The goal of this section is to present a review of state-of-the-art articles in the field of BC recurrence. For terminology and background knowledge about BC, see Section 2.1.

There has been significant progress in the characterization of BC. However, it is often still difficult to accurately predict its behavior [17]. In luminal like disease (hormonal receptor positive, HR^+), this is especially noted, since the mechanisms of treatment resistance, late relapse and dormancy are not well-understood [18, 19]. Moreover, BC metastatic behavior not as well studied as breast cancer itself. Searching in a platform from *Thomson Reuters* [20] for documents with the expression "breast cancer" in the title yields more than 330 000 results. In comparison, the value associated with recurrence is of around 20 000, obtained when including the results for any combination of the terms "recurrence(s)", "relapse(s)" and "metastasis(es)". This starts to show the novelty of this work, and the development of new therapies and approaches might decrease the incidence of relapse, as described by Hurk et al. [21] in a Dutch population-based analysis. After a direct contact with IPO-Porto, it was decided that this work will focus on the prediction of recurrence in the different metastatic locations, trying to assess the relation of different characteristics of patients and tumors with different prognoses.

Even though cancer tumors are not completely homogeneous masses, it was found that the characteristics of the primary tumor are usually preserved in metastases [22]. Several studies analyze the impact of BC subtypes [4, 23–25] and the tumor's hormone receptor status [26–29] in its relapse patterns. However, the effect of HER2 status (Human Epidermal growth factor Receptor 2) on distant recurrence in early stage breast cancer differs according to the metastatic site [23], for example.

Throughout this review, Overall Survival (OS, the time interval between diagnosis and death or last contact), and Progression-Free Survival (PFS, the period from the diagnosis of metastases to progression of disease, death, or last follow-up) are common metrics used. The results of that type of analysis may have important implications in our understanding of the disease process, allowing more aggressive treatment in a subgroup of patients [30], while avoiding overtreatment in others.

In a study with 3726 female patients, Kennecke et al. [31] showed that the luminal/HER2$^+$ and HER2 enriched tumors are associated with higher rates of brain, liver and lung recurrence. On the other hand, triple negative tumors are associated with a significantly higher rate of brain and lung metastases but a significantly lower rate of liver and bone relapses. In this particular study, the goal was to compare characteristics from patients and their tumors, throughout the several BC subtypes. Patients' data were retrieved from the British Columbia Cancer Agency.

The tests used were χ^2 (chi-squared, for categorical variables) and Wilcoxon rank sum (for continuous). The method used for the estimation of cumulative incidence curves was the "competing risks methodology" (to estimate a single event when several competing ones exist: the patients who died before developing recurrence, and those who hadn't died at cutoff date). Having established the cumulative curves across BC subtypes, Gray's test was the choice to compare them, testing them for statistically significant differences. Survival (from initial and recurrence diagnoses) was estimated with Kepler-Meier method, and was later compared with the log-rank test. The site of relapse was tested for association with the BC subtype with chi-squared, and also with multivariate models using logistic regression (dependent variable: presence or not of relapse in a determined site; covariates: characteristics of patients/tumors). It was used the software SAS (Statistical Analysis System) and also the R Statistical Language (Programming) Language.

The primary distinction of locations is between bone-only metastases and visceral sites. Visceral metastases were classically related with worse prognosis. Some patients and tumor characteristics could be linked with this type of recurrence namely age, menopausal status, tumor size, lymph node involvement, stage, Estrogen and Progesterone Receptors (ER and PR), and HER2 pattern [30]. Among visceral sites, three sites will be considered in the following sections, according to the most observed categories.

The relapse sites were considered bone, liver, brain, and lung, each one in the next subsections.

3.1.1 Bone

In 2010, Kennecke et al. [31] reports bone as the most commonly diagnosed metastatic site. This study included more than 3000 female patients with recurrent BC. Mendonza et al. [30] showed that the development of bone metastasis was dependent on primary tumor characteristics like size, lymph node involvement, lymphovascular invasion, stage, estrogen and progesterone receptors and HER2 pattern.

In a cohort of 351 breast cancer patients with bone-only metastasis, Niikura et al. [32] intended to determine what factors influence the outcome of patients with bone-only metastases, while comparing PFS and OS of these patients, according to their treatments. The study defines as favorable prognostic factors: performance status 0-1, asymptomatic bone disease and single metastasis. Patients with metastatic disease at diagnosis (vs recurrence), a single metastasis and asymptomatic bone dis-

ease had also a prolonged PFS. In luminal patients (HER2 negative) combination therapy (chemotherapy and endocrine therapy) was associated with better outcomes (OS and PFS). In multivariate analysis, combination therapy (chemotherapy with endocrine therapy) was not superior in terms of PFS or OS to endocrine therapy alone. In patients with HER2+ disease, trastuzumab was associated with longer PFS with no impact in OS [32].

To achieve this, several methods were used: to sum up the age at diagnosis, μ and σ (mean and standard deviation) were used; mean age was compared between treatment groups using variance analysis; Pearson's χ^2 and Fischer's exact tests were performed to test the association of these treatments with categorical clinical features; these features were presented using frequencies and proportions; to estimate PFS and OS, Kaplan-Meier product-limit was the method chosen, while Kaplan-Meier curves were used for presentation; Cox regression models (uni- and multivariate) had the purpose of testing the effect of predictive factors. The software used was SAS (analysis) and S-PLUS (plots).

3.1.2 Liver

In 2003, Wyld et al. [33] analyzed 145 patients liver metastases at the time, from a total of 506 that presented to the Nottingham Breast Unit over a 5-year period. Patients' age ranged from 24 to 92 years, with a median of 61 years. None of them was able to survive at least 5 years, and only one patient managed to survive more than 3 years. Kruskal–Wallis ANOVA and Mann–Whitney U-test were used to compare survival times between patients with different metastatic distribution patterns, while Kaplan–Meier cumulative survival plots for survival with liver metastases according to a serum concentration level.

Local therapies for hepatic metastasis might provide a survival benefit in some patients. Surgical resection is a safe procedure in specialized centers, with a 5-year overall survival after that of 21-61% [34, 35]. Local ablation, has also a survival benefit (5-year overall survival of 27-41%). Macroscopically radical resection (R0) is the only favorable prognostic factor described in the studies and a major surgical effort to obtain this with a low mortality is essential. Factors like: age $\leq$50 years [36], time to hepatic recurrence $\leq$1 year after diagnosis [35–37], metastases size (overall metastases diameter higher than 3.5cm) [36] and number [37], absence of expression of estrogen and progesterone receptors (primary tumor) [36], and the presence of extrahepatic disease [34, 37] were described as worst prognostic factors, but data were not consistent.

Weinrich et al. [38] added in 2014 that high T and N stage and high grade (of the primary tumor) could also be one of the worst prognostic factors. Their study focused on the University Hospital of the Saarland (Germany), analyzing 29 operations performed on 24 patients suffering from isolated liver metastases of breast cancer (3 patients required two surgical interventions and 1 patient required three). SAS software was used to perform the statistical calculations. Survival rates were compared with the logrank test. Multiple regression analysis was performed using a Cox regression. Mortality rates of two groups at fixed time points were compared with Fisher's exact test. Test results with p values of less than 0.05 were considered statistically significant and results with p values between 0.05 and 0.10 were statistically only slightly significant.

3.1.3 Brain

Breast cancer is the second most common cause of metastasis to the Central Nervous System (CNS) [39]. In a review of 420 patients, Altundag et al. [39] described that these patients had more frequently tumors T2 (40.1%), N1 (59.7%) and G3 (81.4%), at diagnosis and CNS was the first site of recurrence in 12%.

Comparing to other tumors that also affects CNS (lung, GI, renal and melanoma), breast cancer patients seemed to be younger, had more frequently the primary tumor controlled and had better survival [40]. Better prognosis was associated with: younger age, no skull base involvement, radiation therapy, luminal A tumors, good basal performance status (ECOG 1-2 vs 3-4) and bone metastasis only vs hepatic or skin metastases [40]. The increase on the risk of dying was independently related to lack of systemic therapy and liver involvement [40].

Survival differs also according to BC subtypes, with a worse survival related to triple negative tumors, according to Anders et al. [25]. In their study, comprising 119 patients who had BC-derived metastases in the CNS, the goals were: to assess the clinicopathologic traits of the tumors, evaluating primary BC and metastases in the CNS; report local and systemic treatments in patients with brain metastases; and evaluate the correlation between patients' outcomes (after CNS recurrence) and their characteristics, both intrinsic traits and the BC subtype. To compare survival curves, the Kaplan-Meier method and log-rank test were used, while Cox regression analysis was the choice for the evaluation of possible predictors. The authors built 4 different multivariate models, since it was not possible to put together all the information into a single one, due to the presence of missing data. The software used was SAS.

3.1.4 Lung

Overall, lung is the second more frequent site of breast cancer recurrence [41]. Considering only basal-like tumors, lung was the most common relapse site [31, 41], and this subtype represents around 40% of metastatic BC in the lungs [41].

Similarly to other visceral locations, metastatic BC to the lung is not commonly associated with luminal A tumors. However, metastases in the lung derive from luminal B cancer in a significant percentage (around one third) [41].

3.2 Data Mining approaches

The classification of cancer patients into groups with different prognoses is essential for providing customized treatment, and automated systems can aid clinicians in the decision-making process [42]. Tumor characteristics are not enough to assess the patient, as it was regarded in the past, since the classification through tumor morphology is only representative in less than 25% of patients with invasive breast carcinomas [43]. But allowing clinicians to predict the outcome of this disease helps them to make more informed decisions to improve the efficiency of the treatments. Due to the high dimensionality of databases, it is necessary to develop intelligent strategies to find meaning in such data [44, 45].

Statistical techniques were the traditional approach to discover hidden relations among data variables, but Data Mining techniques have been gradually adopted, and have been applied in several fields including medical research [46], obtaining good results. Paliwal and Kumar reported in 2009 that Artificial Neural Networks

(ANN), probably the most commonly applied data mining modeling example, had been used for prediction and classification tasks, for which statistical methods used to be the typical choice [47]. Most authors apply only one of the methodologies, although some comparisons also exist in the literature [48, 49].

The patterns of relapse of BC are yet to be fully studied with application of machine learning methodologies, but some research has been published, especially using private databases. This section provides a review of some of these articles, developed for the prediction of the prognosis of breast tumors, regarding recurrence.

In 1997, Subramani Mani et al. [50] used a database from a Breast Care Center, with 887 patients, to find tumor features associated with recurrence of BC. About 10% (85) of these patients experienced this event during follow-up, while the remaining 90% (802) had no evidence of it (10% rate of recurrence). Since the two classes were imbalanced, 6 different subdatasets were created, each with 148 relapse-free patients and all of the 85 with recurrence (64%/36%). From many initial features, six were hand-picked by a surgeon. The algorithms used included DT (C4.5 and CART) and Association Rules (C4.5rules and First Order Combined Learner [FOCL]). According to the authors, the extracted trees and rules (respectively) provide crucial information, especially in this medical context [50–52]. In this paper, a comparison is made with the Naive Bayes algorithm, but all of the other algorithms failed to surpass its accuracy results (average of ≈68.3%). To properly evaluate the techniques, 50 runs were conducted for each subdataset, splitting into different partitions of training set (n=155) and test set (n=78). Averaging the accuracy of the 300 runs (50×6), the second best value was achieved by FOCL (≈66.4%). However, this was the only metric used, which does not allow a full comparison of performance.

José Jerez-Aragonés et al. [53] employed a hybrid model, combining ANN and DT, to a database from a hospital in Málaga, Spain. There were 14 variables chosen by doctors beforehand (from 85 fields), and information of this data is presented in the article (range, mean, standard deviation, median). The authors apply a neural network to predict recurrence in BC patients at 7 given intervals (10-month periods: 0-10, 10-20,...,50-60 and more than 60 months), using a subset of the 14 variables as input. To choose which variables to use, they employ a new decision tree algorithm (CIDIM: Control of Induction by Sample Division Method). This dataset was constituted by 1035 patients, but records with Missing Data were discarded, resulting in 845, 741, 681, 600, 520, 466 and 466 patients, respectively for each interval. To evaluate the performance of the algorithm, the authors partitioned the data (Holdout method) in 80% for training and 20% for testing purposes. Using an holdout method (partition train/test) with 20% of the data for testing purposes, the accuracy values found range from 93.4% to 96%, while sensitivity varied between 78.7% and 88.7%, and specificity between 94.5% and 97.2%.

Amir Razavi et al. produced two papers in 2005 [54, 55] concerning the application of Canonical Correlation Analysis (CCA) to the study of BC recurrence. In the first study [54], associated with a Swedish Breast Cancer Study Group, the purpose was to try to find risk factors for both local and distant recurrence. The database used was local, with 637 patients and 18 variables (17 binary and 1 with three values). The idea stated is that CCA could be applied as a feature selection method, without decreasing the predictive performance. The advantage indicated for CCA is the possibility of analyzing the correlation of sets of multiple variables, which allows the evaluation of several outcomes simultaneously. To the best of our knowledge, no other authors applied this technique to the subject in question. They didn't validate their results analytically, but the system seemed to detect known

risk factors, according to the authors, specifically for the time intervals of 0-2 and 2-4 years. The impact of CCA on an actual classification task and the associated performance metrics is the focus of the next paper.

In the other article [55], Razavi et al. applied CCA as a preprocessing method, to predict BC relapse using DT. The dataset used included 3949 patients with BC, obtained from a Swedish regional center. Unlike in many other articles, handling of Missing Data was performed in this study, instead of removing these patients. For this purpose, Expectation Maximization (EM) method [55] was used, to estimate the missing values of incomplete data. From more than 150 variables, the first step was to select 13 predictors with the help of medical experts, which resulted in 17 inputs. The outcomes were local and distant recurrence, both before and after a five-year threshold (from time of diagnosis). CCA application resulted in a reduced system, with 8 inputs and 1 output (distant metastases in the first five years). The best accuracy results are obtained using the proposed preprocessing (67%), higher than without (54%) or just Missing Data imputation (57%). There is however a slight decrease in sensitivity (83% to 80%), meaning a lower capability of detecting positive recurrence. Nevertheless, it is still an important result, considering this solution yields trees with only 10% of the size of those without preprocessing (27 nodes with 14 leaves, instead of 273 nodes with 137 leaves). This results in a simpler system, improving interpretability.

In 2007, the same authors applied once again DT to predict recurrence in BC [56]. This time, the main goal was to compare its performance with two medical experts' diagnoses. From the dataset with 3949 patients, repeated entries were removed, resulting in 3699 registries. The authors left 100 cases aside for comparison (selected randomly, with the same class proportion as the original dataset), and the DT performance was based in 10-fold cross-validation of the remaining 3599. CCA was used again to select the variables, the outcome chosen was "distant metastasis or death because of breast cancer within 4 years", and the Missing Data imputation method used was Multiple Imputation (MI). This is a combination of EM with "a data augmentation [...] procedure" [56]. Despite a better accuracy, DT had lower AUC values, but the differences were not statistically significant. In terms of predictive power of Recurrence, DT was better than one of the doctors, but worse than the other. The results obtained for DT were 82% for accuracy and 0.755 for AUC. A good point of this article is the presence of the confusion matrices of both oncologists and DT, and the ROC curve and AUC values.

In the same year, Yijun Sun et al. [57] combined clinical information with genetic features to try to obtain better predictive results. The dataset is publicly available in Nature website [58] (and it was used by Laura van't Veer et al. [59] to create a 70-gene signature of BC, to predict patients' outcome and treatment responses). In this study, the authors use 97 registries in their analysis, in which they try to predict distant recurrence of BC in the first five years. As preprocessing, the data is normalized to the range of 0 to 1, and feature selection (I-RELIEF method) is applied. To evaluate the methodology and compare it against the previous approaches, the authors set a 90% sensitivity threshold, and analyzed the specificity values. In fact, the proposed algorithm achieves the best performance, with 67%, better than the genetic-only study (47%) and clinical-only (48%) studies. The AUC value was also better (visible from the ROC curve), although no concrete values were not provided. However, it would be useful to carry this analysis using a larger dataset, which is more difficult to compile (hence the small size of the one used), given the hybrid nature of the system (both clinical and genetic data). It is still a good result, given

the difficulty shown in previous studies in combining these data [60, 61].

In terms of number of different algorithms tested, the most comprehensive study was found to be published by Thora Jonsdottir et al. [62], in 2008. With 17 different algorithms, they tested a wide range of techniques: Naive Bayes classifier, different DT and several Rule Inducers, among others, although these algorithms were only used with one configuration. Using a dataset named *Rose*, from Iceland, with a total of 400 variables reported, they chose the top 98 features using knowledge from literature, creating a base dataset (*Base-DS*). *Med-DS*, which contained the best 22 features chosen by a doctor, and *Small-DS*, including only the top 5, were also used in this study, along with many others (over 100) generated with feature selection methods. The number of patients included is 257. One of the goals was to predict whether a BC patient would develop recurrence during a 5-year period after diagnosis. Then, the authors tried to predict the same, but with an added subjective variable, a Risk group (low, intermediate, high; attributed by a doctor). Finally, a secondary goal was to predict the Risk variable from the remaining variables. The datasets yielded similar results, with better performance for Naive Bayes and J48 DT, without significant improvement from the addition the Risk variable. Despite the accuracy reported being around 75% to 80%, the value of sensitivity was only around 40%, which is especially bad in the medical context (indicates a large number of wrong predictions of "Recurrent" class). A better way to assess the performance is with the AUC, for which Naive Bayes had the best value (0.77), for *Small-DS*. All of the values were validated using 10-fold cross-validation, a strength of this study.

The study developed by Qi Fan et al. [63] in 2010 targeted the internationally available SEER dataset [7]. Records with Missing Data were ignored, but the authors didn't mention the number of patients in the final dataset. As feature selection, medical consulting resulted in 13 attributes being selected as inputs. The algorithms used included ANN and DT, with four variants of the latter. Dividing the dataset into training and test partitions (80%/20%) showed that C5 DT had the best accuracy (71%), but ANN provided better predictive power for the Recurrence class (78%, higher than 72% by C5 DT), although with lower accuracy (66%). However, the data was only partitioned once, which may mean that results are not representative of the real performance. Moreover, a single configuration of the algorithms was used, and the authors did not provide details about the architecture used nor about the reasons to use it.

Smaranda Belciug et al. (2010) [64] used a clustering approach to predict recurrence in a public BC database, WPBC (Wisconsin Prognostic Breast Cancer dataset) [65], with 198 patients. It is known that there are From a total of 34 features included in the original dataset (numerical variables, continuous), the authors chosen 12 to be considered inputs, though no methods for this selection were made explicit in the text. The output class was the presence of relapse. The three algorithms used were k-means, self-organizing map, and cluster network. The latter obtained the best results, by comparing the test performance. The system had 78% accuracy, obtained through 10-fold cross-validation.

Leila Ahmad et al. [66] compared in 2013 three different methods to predict recurrence of malignant breast tumors. The data used were retrieved from a national center in Iran. From a total 1189 records, 642 were removed because important data was missing, resulting in a cohort of 547 patients. Then, an imputation method was applied to estimate the values of other continuous variables, namely Expectation Maximization. Using ANN (MLP), DT (C4.5) and SVM, the final result was obtained through 10-fold cross-validation. To evaluate the performance the authors

presented the accuracy, sensitivity and specificity values. In all metrics, the SVM method had the best values (95.7%, 97.1% and 94.5, respectively), and was thus considered the best performing algorithm in this study.

Alberto Pawlovsky and Mai Nagahashi (2014) [67] applied the well-known kNN clustering algorithm to BC Prognosis. This method creates groups (or clusters) of labeled data, so that new data can use the nearest neighbors (data with more similarities) to induce its own class. Moreover, the article explains how to select a good setting for the algorithm. The dataset used was WPBC [65] (198 patients, 4 with Missing Data removed), and the data was used in three ways: raw data (without preprocessing), standardized data (mean = 0, standard deviation = 1), normalized data (range from 0 to 1). The authors then tested the classification with kNN with several configurations, by using a different number of neighbors and different percentage of training data, and running it several times. It is explained that a "good" setting would retrieve a good average prediction value, but also not very low minimum. The first conclusion was that the preprocessing does not seem to influence the results obtained significantly. Then, the average results increase as the k value increases (with drops at even k values), but it stabilizes around 70%, and the minimum is still 20%. Bearing all this in mind, the authors build a score mechanism, in which they assign weight values to several parameters (three) and resulting metrics (five). Considering the ideal result, scores were given for the different possibilities: a higher score is given to a higher mean, but also to a lower range of accuracy values, for example. Finally, from 5130 different settings, the conclusion is that the best setting includes 80% of data (raw/standardized better than normalized) for classification, and a k value of 19. Also, 100 simulation trials seem to be enough to assess the settings' performance. The accuracy achieved with this methodology was 76%, ranging from 62% to 90%.

In the same year, Zahra Behesti et al. [68] used a more modern approach in nine different medical databases. Among them is the WPBC [65] (198 patients), for the prognostic of patients with malignant breast tumor. To handle Missing Data (4 records), the authors used the Mean method (statistical) [68]. The methods used are based in Particle Swarm Optimization (PSO), in which a population of candidate solutions moves gradually towards a global solution, by following the best positions of the "swarm" (the group). Besides more common approaches (in this field), a novel one is shown, namely a Centripetal Accelerated PSO (CAPSO), which takes advantage of Newton's motion laws. Moreover, the authors implement a fusion of CAPSO (and other three methods) with ANN (MLP), resulting in a hybrid learning strategy. The settings used to configure the parameters of the architecture used were said to be based in the literature. To evaluate these algorithms, several metrics are presented: Mean Square Error (MSE), Accuracy, Sensitivity, Specificity and AUC. In addition, statistical tests between the accuracy values of the approaches considered are also performed (Wilcoxon's signed ranks and t-test). Particularly for Breast Cancer, CAPSO-MLP had significantly better results than two of the others (mean 80.25%, ranging from 77.5% to 82.5%). The only close result was obtained by Gravitational Search Algorithm (GSA-MLP), but its sensitivity values only averaged less than 8%, compared to 52.33% of CAPSO-MLP, which also obtains the best specificity (83.38%) and AUC (0.63). Each algorithm was run 10 times, and the best, worst and mean results were provided by the authors. The values presented were based in the application of the Holdout method. For training purposes, 80% of the data was used, while the remaining 20% constituted the test partition. The latter originated the resulted observed in this review.

Table 3.1 presents an overview of the datasets and methodologies used in the studies analyzed, and results achieved for several metrics. To identify each article, it is presented the first author and publication year. Concerning the dataset, its availability and number of records is shown. The main metrics are also in the table, as well as the algorithm that achieved them. Moreover, validation methods used are displayed in this table. The last column shows if Missing Data was observed in the dataset, and if so, how the authors handled this problem.

3.3 Conclusion

As can be seen in the two previous sections, there have been many studies regarding the pathways of BC recurrence. The main purpose of the first section was to systematize the previous work in this area. The authors of these research studies use statistical algorithms to find the characteristics of the patients in each study population. After an overview of the recent evidence about this topic, a more specific analysis to the work developed for each recurrence site is presented.

Machine Learning algorithms have also been applied to the study of BC recurrence, with the capacity of unveiling information hidden in the data, generalizing from its underlying patterns. The authors either use binary response variables in the classification task, or try to predict periods of time, which means that these studies do not exactly match the goal of the present thesis. However, the referred articles can help understand what kind of algorithms may be used in this area, and there have been interesting developments in this field.

Dataset description and Evaluation Metrics results

Author	**Year**	**Dataset**	**Patients**	**Variables**	**Best Algorithm**	**Acc**	**Sen**	**Spe**	AUC	**Validation**	**Missing Data**
Mani	1997	Private	887	D	NB	68.3%	-	-	-	Holdout	No
Jerez-Arag.	2003	Private	845-466	C	ANN	95.6%	88.7%	96.5%	-	Holdout	Yes: removed
Razavi	2005	Private	3949	D	DT (C4.5)	67%	80%	63%	-	10-fold CV	Yes: EM
Razavi	2007	Private	3699	D	DT (C4.5)	82%	21.1%	96.3%	0.76	10-fold CV; Holdout	Yes: MI
Sun	2007	Public (Nature) [58]	97	G+D	LDA	-	90%	67%	-	Leave-one-out CV	No
Jonsdottir	2008	Private	257	D	NB	79%	36%	96%	0.77	Stratified 10-fold CV	No
Fan	2010	Public (SEER) [7]	-	D	DT (C5.0)	71.2%	71.7%	70.7%	-	Holdout	Yes: removed
					ANN	65.8%	77.8%	53%	-		
Belciug	2010	Public (WPBC) [65]	198	C	Cluster network	78%	-	-	-	10-fold CV	Yes: unknown
Ahmad	2013	Private	547	D	SVM	95.7%	97.1%	94.5%	-	10-fold CV	Yes: removed
Pawlovsky	2014	Public (WPBC) [65]	198	C	Cluster(k-means)	76%	-	-	-	100 repetitions	Yes: removed
Behesti	2014	Public (WPBC) [65]	198	C	CAPSO-MLP	80.3%	52.3%	83.4%	0.63	Holdout	Yes: Mean

Table 3.1: Description of methodology and results.

Acc = Accuracy, ANN = Artificial Neural Network, AUC = Area Under the (ROC) Curve, C = Continuous, CV = Cross-Validation, D = Discrete, DT = Decision Tree, EM = Expectation Maximization, G = Genetic, LDA = Linear Discriminant Analysis, MI = Multiple Imputation, NB = Naive Bayes, Sen = Sensibility, Spe = Specificity, SVM = Support Vector Machine

Chapter 4

Experimental Setup

This chapter presents the methodology used in this thesis. The first section describes the dataset used in this work, while the other two present implementation details, namely the plan for the handling of incomplete records (Section 4.2) and the construction of a classification model (Section 4.3), respectively.

4.1 Dataset characterization

The dataset used in this work was retrieved from IPO-Porto. The study population is composed of female patients, older than 18 years of age, with breast carcinoma histologically confirmed in all of this patients. To protect the confidentiality of the patients, we never had access to their names, using an ID (IPO number) as distinguishable identifier. From a database with 99 patients, two of them did not contain the necessary information about recurrence. Those were removed immediately, leaving a final cohort with a total of 97 patients.

The next step was to analyze the distribution of MD among variables. It was found that 12 features were complete for all patients, while the remaining had MD rates in the range of 1%-91%. After removing some variables with MD rates above 70%, the final number of variables was 27, of which 12 are complete and 15 are not. This left the database with only 28 complete patient records (28.85%), while the remaining 69 (71.13%) had at least one missing value.

4.1.1 Inputs

Table 4.1 shows the distribution of the missing values. About the table:

- Ca 15-3 was transformed into a binary feature, with a cut-off value of 30 U/ml, based on the literature [69–74];
- The variable Age_Dx_years contains the age of the patient, in years, at the time of diagnosis of BC (range = 27-84 years, median = 48 years);
- When the variables concerning the histology of the tumor (whether it is Ductal and Lobular, respectively) are both *true*, the tumor is considered "Mixed", while the combination of both features as *false* means "Other" type, as defined by the doctors from IPO-Porto;
- Patients in this study have disease of either stage *I*, *II* or *III*;
- ER, PR and HER2 expression were determined via IHC;

Table 4.1: MD rates in the input features used (in percentage).

Variable name	Variable type	% MD	(N)
Menopausal	binary	8.25	(8)
CA15-3_initial	binary	40.21	(39)
Age_Dx_years	discrete	0	(0)
Type_Sx_1	binary	1.03	(1)
Type_Sx_2	binary	1.03	(1)
Hx_ductal	binary	0	(0)
Hx_lobular	binary	0	(0)
T	ordinal	5.15	(5)
N	ordinal	3.09	(3)
Grade	ordinal	11.34	(11)
DCIS	binary	40.21	(39)
Stage	ordinal	2.06	(2)
ER	binary	0	(0)
PR	binary	11.34	(11)
HER2	binary	19.59	(19)
Tx_Neoadj	binary	0	(0)
CT_neoadj	binary	0	(0)
HT_neoadj	binary	8.25	(8)
Trast_neoadj	binary	5.15	(5)
N_cycles_neoadj	discrete	0	(0)
Tx_adj	binary	0	(0)
CT_adj	binary	0	(0)
RT_adj	binary	0	(0)
HT_adj	binary	0	(0)
Trast_adj	binary	4.12	(4)
agLHRH	binary	0	(0)
N_cycles	discrete	3.09	(3)

Adj: Adjuvant, agLHRH: Luteinizing Hormone-Receptor Hormone agonist (therapy combined with tamoxifen), CA15-3: Cancer Antigen 15-3, CT: Chemotherapy, DCIS: Ductal Carcinoma In Situ, Dx: Diagnostic, ER: Estrogen Receptor, HER2: Human Epidermal growth factor Receptor 2, HT: Hormone Therapy, Hx: Histology of tumor, N: part of the TNM staging, N_cycles: Number of (chemotherapy) cycles, Neoadj: neoadjuvant, PR: Progesterone Receptor, RT: Radiotherapy, Sx: Surgery type, T: part of the TNM staging, Trast: Trastuzumab, Tx: Therapy (in general)

- In the TNM staging, the T status represent the diameter of the tumor (T1, T2, T3 or T4), while the N status indicates the invasion of cancer cells to the lymph nodes involvement (N0, N1, N2 or N3). M would be the Metastatic characteristics;

- The Grade feature refers to how differentiated the tumor cells are (1 to 3);

- Considering the tumor size, adjacent organs invaded, regional lymph nodes it has spread to, and whether it metastasized, a value 1 to 3 (*I*, *II* or *III*) is atributed to the feature Stage.

4.1.2 Outputs

Table 4.2 shows the output variables. Each of the variables refer to a single location, with exception of "Rec_other", which represents all the other relapse sites. In the same table, the number of patients in the positive class (with metastasis in that site) is indicated.

Variable	Variable type	Positive
Rec_bone	binary	63
Rec_ganglia_local	binary	11
Rec_ganglia_distant	binary	15
Rec_local	binary	12
Rec_pleural	binary	13
Rec_pulmonary	binary	22
Rec_hepatic	binary	19
Rec_cerebral	binary	8
Rec_other	binary	5

Table 4.2: Outputs respective to recurrence sites.

4.2 Missing Data handling

Missing values may have different origins, but for the purposes of this work, it will be assumed that all MD is missing completely at random (meaning that its real value is uncorrelated to being absent). The methods used to handle MD included Deletion and Imputation methods: with the first, patients or variables with MD are deleted, to generate a smaller complete dataset; the second attempt to estimate those missing values using statistical and ML techniques.

4.2.1 Missing Data simulation

To assess which imputation methods performed better, a simulation of the several available algorithms was prepared. This consisted in using only the complete variables of the original dataset, removing some values at random After making a selection of the best imputation methods, the classification step can be done in much less time.

The MD percentages to test were decided to be 5%, 10%, 15%, 20%, 25%, 30%, 50% and 70%, to cover a spectrum of percentages without overcharging the simulation, until an acceptable maximum. However, performing a brute-force analysis would generate

$$(m+1)^v - 1 = 9^{12} - 1 = 282\,429\,536\,480$$

datasets for each imputation configuration, where m is the number of MD rates ($m+1$ includes the 0%), v is the number of variables, and the "-1" at the end of the formula removes the combination where none of the variables has missing values. Therefore, it was decided to perform feature selection, to determine the most important features, in which we would introduce missing data.

Feature Selection

The purpose of using feature selection at this stage is to diminish the number of combinations of MD rates to analyze. To do so, four feature selection methods were used (code was available), and a rank system was built based on them. The four methods were based in AUC (Area Under the receiver operating characteristic Curve), F1-score (harmonic mean of two other evaluation metrics), information gain, and the point-biserial correlation coefficient, respectively.

Firstly, each method was applied to each complete variable, in relation to each binary output at a time. Then, we averaged the results of each feature selection for

each variable through all the outputs. Then, we ranked them from higher scores to lower, awarding more points to higher positions. Finally, we added the points from each feature selection algorithm.

After choosing the most important variables, the simulation of MD is ready to start.

Imputation

When the new datasets are created, the system can start imputing them with the desired approaches, whether they are statistical or apply ML techniques. Two statistical methods were used, Mean Imputation and Median Imputation, which are exactly what they seem: replacing each missing value for the mean or median, respectively, of the non-missing elements of the same feature. These results cannot be improved, because there are no parameters to change.

On the other side, we have ML algorithms, namely kNN, ANN, DT and SVM (defined in Section 2.2). For each one, the methodology of the implementation was the same, both in the inner working of these methods and the search for the best architecture.

Considering the inner working, all algorithms use the complete patients for training, while the testing occur with the incomplete patients. For both phases, the complete variables are the input, while the incomplete ones are the target/output. In terms of our search for the best settings, they were also used in the same way: starting with a combination of some values for the parameters, a group of the best is chosen according to the evaluation metric desired, to then explore more around the same search space.

In the case of kNN algorithm, the first iteration used only five values for k (number of neighbors), but all possible values for the distance. The values chosen for k were sufficiently apart to cover the interval from one to the number of patients: 1, 10, 20, 50 and 90. By the end, we could estimate what were the best distances, and then proceed for a more thorough search in other k values. Afterwards, the same was done for ANN, DT and SVM, each one with their own parameters, but the search method was the same.

4.2.2 Validation

When creating each dataset with random missing data, the same dataset is used for all imputations. To ensure that random processes did not play a role in the different performances, the set of all datasets created is the same for every imputation architecture. This single iteration may lead us to think that randomness could still play a role, since we do not repeat the process: however, it does not lose the robustness since the final value for the evaluation metric is the average of thousands of values from thousands of imputations.

The choice for metric, due to its simplicity, is Accuracy, given by the following formula:

$$Accuracy = (TotalCorrect)/(TotalValuestoImpute)$$

This gives us a general idea of how much the system is learning, as a proportion of the total missing values to impute in each dataset. The final metric to use is an average Accuracy over the total datasets, and the results are shown in Section 5.2.

4.3 Classification

"Prediction is an attempt to accurately forecast the outcome of a specific situation, using as input information obtained from a concrete set of variables that potentially describe the situation" [53]. Our task is to make a model learn the underlying patterns in the data. To that end, we applied several ML algorithms to create models that tried to accurately predict the output variables for new, unseen data. Averaging the metric of choice over the several outputs after cross-validation was the method used to evaluate and validate the classifiers.

4.3.1 Classification Algorithms

The methodology applied in this step is the same as the Imputation task: we start each algorithm with a set of algorithms, evaluate them, and proceed to another round with a different set of parameters. For more information, see Section 4.2. Besides the classifiers used in imputation, NB was also used for classification, searching the best solution in the same way. In this case, the Kernel Smoother type was the parameter changed.

4.3.2 Validation

To validate the models created in the previous step, there are several possible validation processes and evaluation metrics. As for the process, 10-fold Cross-Validation was chosen, for its acceptance as a standard [62, 75].

Regarding evaluation metrics, Accuracy is in practice the most used metric [76]. In fact, that is used for the imputation phase, as described in Section 4.2, because it did not matter what the model predicted correctly, as long as it did. With thousands of datasets to impute, the training and test cases have much variation. However, in the case of our outputs, in some cases very imbalanced, the accuracy can be misleading: if three quarters of the outputs belong to the negative class, the model can have 75% accuracy just by assigning every patient to that class.

Sensitivity and Specificity give us the definition we would like for Negative and Positive classes. Their formulas are:

$$Sensitivity = TP/(TP + FN)$$

and

$$Specificity = TN/(TN + FP)$$

where:

- TP = True-Positives (elements of the positive class correctly classified)
- TN = True-Negatives (elements of the negative class correctly classified)
- FP = False-Positives (elements of the negative class incorrectly classified)
- FN = False-Negatives (elements of the positive class incorrectly classified)

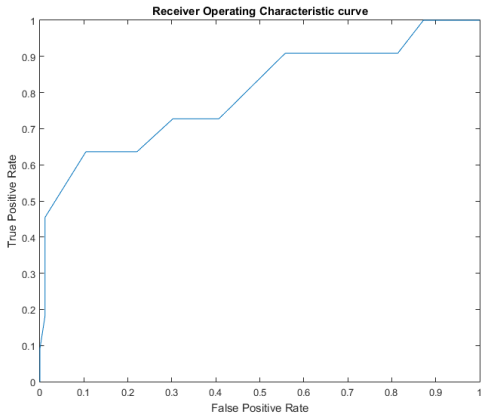

Figure 4.1: Example of a ROC curve

At one extreme, all outputs can be considered part of the positive class, originating a sensitivity of one and a specificity of zero; at another, all patients can be considered in th negative class, originating the opposite. This leaves us with the problem of having two metrics instead of one: if two models have only one of the measures higher than each other, how can we decide that one is better than the other? What is needed is "an unbiased measure of the accuracy of the model", that can also account for both classes and how much we lose or gain by changing the thresholds of decision. The ROC (Receiver Operating Characteristic) curve is plotted by associating each value of sensitivity to the correspondent of specificity. The Area Under this Curve is called AUC, and weighs both sensitivity and specificity. The final value for each architecture of classification was the averaged AUC over the nine outputs in use.

Moreover, we made sure that randomness was "controlled", by using the same partitions (folds of cross-validation) for every creation of a classification model, besides the restart of the random number generator.

4.4 Conclusion

The methodology of this thesis shows the steps taken during the implementation phase. Starting with a raw dataset, it was preprocessed manually, and then MD was computationally handled. After this, the dataset was ready to start building the classification model. The results of the MD simulation and the Classification are shown in Chapter 5.

Chapter 5

Results

There were several implementation steps in the course of this thesis. This chapter covers all of them, presenting the actual results of the experiments already described. In the next sections, the results of Feature Selection (for MD simulation), Imputation of missing values, and Classification will be in Sections 5.1, 5.2 and 5.3, respectively.

5.1 Feature Selection

The process of Feature Selection, as a preparation for MD Imputation, is explained in Section 4.2.

Variable	**Variable Points**
Age_Dx_years	10
Hx_ductal	12
Hx_lobular	12
ER	**22**
Tx_Neoadj	**30**
QT_neoadj	**36**
N_cycles_neoadj	**36**
Tx_adj	11
QT_adj	14
RT_adj	16
HT_adj	10
agLHRH	11

Table 5.1: Feature Selection ranking points for each variable: the higher the score, the more important the variable is

Table 5.1 It can be seen that four of the variables had considerably better results. Therefore, these were the variables chosen for the next step of the work, the imputation of MD. It is good that not all of the chosen variables are binary, since the initial dataset also contained non-binary features which will have to be imputed afterwards.

This four features represent a total number of $9^4 - 1 = 6\,560$ datasets. The next section addresses the imputation of all of them with different algorithms.

5.2 Imputation

In this section, the results of the imputations are presented, culminating in the group of datasets to use in the classification phase.

5.2.1 Imputations by algorithm

The metric considered was accuracy, and the final value was calculated as the average of the 6560 datasets. The next subsections will display the results of the imputations performed during the imputation phase. For background information about the ML algorithms, see Section 2.2.

Statistical methods

The first imputations were Mean and Median Imputations, and the results are registered in Table 5.2.

Method	**Average accuracy in percentage (AvAp)**
Mean	57.85
Median	**75.97**

Table 5.2: Imputation results for Mean and Median

The better result of Median is probably explained by the presence of a non-binary feature, with values up to 8, that increase the mean, while 0 is the most frequent value. The median, on the other side, accurately predicts the 0's.

kNN

The first iteration was performed with the following parameters:

- Mean vs Median
- $k = \{1, 10, 20, 50, 90\}$
- distance = {euclidean (1), seuclidean (2), cityblock (3), minkowski (4), chebychev (5), cosine (6), correlation (7), spearman (8), hamming (9), jaccard (10)}

The average accuracies, by parameter, are shown in Tables 5.3, 5.4 and 5.5.

Mean/Median	Mean	Median
Average Accuracy	68.68	**76.14**

Table 5.3: First iteration of kNN Imputation: Mean and Median

k	1	10	20	50	90
AvAp	**74.19**	73.26	73.12	71.42	70.03

Table 5.4: First iteration of kNN Imputation: k

Having in mind the previous results, the parameters selected for the second iteration were these:

Distance	1	2	3	4	5
AvAp	70.86	**73.23**	71.22	70.86	69.65
Distance	6	7	8	9	10
AvAp	**74.97**	**75.03**	**75.03**	71.45	71.18

Table 5.5: First iteration of kNN Imputation: Distance

- Median
- k = {1,2,...,30}
- Distance [metric] = {seuclidean,cosine,correlation,spearman}

The best 5 settings in this iteration are the ones in Table 5.6, ordered by average accuracy (also indicated).

k	**Distance**	**Mean/Median**	**AvAp**
5	correlation	median	79.92
9	cosine	median	79.73
11	cosine	median	79.61
12	cosine	median	79.56
20	spearman	median	79.56

Table 5.6: Final results for kNN Imputation

The best setting for kNN is the first one: using a k value of 5, and correlation as the distance metric.

ANN

To start the ANN imputation, the parameters were:

- Hidden [nodes] = {1,2,3}
- Train [function] = all

The three best results all contained 'trainbr' as a training function, as it is visible in Table 5.7.

Hidden	**Train**	**AvAp**
1	trainbr	77.09
2	trainbr	72.56
3	trainbr	69.94
1	trainlm	69.43
2	trainlm	68.40

Table 5.7: First iteration for ANN Imputation

Using 'trainbr' as the train function, the attempt to use different numbers of hidden nodes to improve the accuracy was unsuccessful, as can be seen in Table 5.8.

hidden	1	2	3	4	5
AvAp	**77.09**	72.56	69.94	69.35	68.77
hidden	6	7	8	9	10
AvAp	68.42	67.89	67.90	67.80	67.78

Table 5.8: Second iteration of ANN Imputation

DT

In the case of DT as an imputation algorithm, the first try was made with these parameters:

- MinLeaf = {1,2,3}
- Split [criterion] = all

The results for the AvAp of the several different split criteria are displayed in Table 5.9.

Split	Gini's diversity	twoing	deviance
AvAp	79.59	79.57	**79.82**

Table 5.9: First iteration of DT Imputation

As we can see, deviance reduction seems to be the best criterion, and we lock it for the next cycle. Then, we try to discover the best value for minLeaf by running many numbers, and Table 5.10 presents the result.

MinLeaf	1	2	3	4	5
AvAp	79.65	**80.17**	80.02	80.05	68.77
minLeaf	6	7	8	9	10
AvAp	68.42	67.89	67.90	67.80	67.78

Table 5.10: Second iteration of DT Imputation

We can see that minLeaf is worse with value 1 than with 2 or 3. This is probably due to overfitting, because the tree is allowed to have leafs for just one patient. Next, we remembered that the 6560 dataset are not all equal, and how much the size of the training partition can change between different datasets. Our idea was to use a relative minLeaf value: instead of setting an integer directly, we could set as a proportion of the training input. The results are displayed in Table 5.11

MinLeaf	1	1/2	1/3	1/4	1/5
AvAp	75.97	75.97	75.90	76.05	77.18
MinLeaf	1/6	1/7	1/8	1/9	1/10
AvAp	79.04	80.08	**80.28**	80.19	80.09

Table 5.11: Third iteration of DT Imputation

In fact, there is a slight improvement, and the new best is now minLeaf = 1/8, with deviance reduction as split criterion.

SVM

The last imputation algorithm is SVM. The starting configuration had several values for three parameter at the same time:

- Kernel [Function] = {linear, Radial Basis Function (RBF), polynomial (orders 1 and 2)}
- Optimization Routine = {L1QP, SMO}
- Standardize = {false, true}

Kernel	linear	RBF	polynomial (order 1)	polynomial (order 2)
AvAp	78.16	**81.61**	**81.61**	79.10

Table 5.12: First iteration of SVM Imputation: Kernel Function

Optimization Routine	L1QP	SMO
AvAp	79.85	**80.43**

Table 5.13: First iteration of SVM Imputation: Optimization Routine

Standardize	false	true
AvAp	79.85	**80.43**

Table 5.14: First iteration of SVM Imputation: Standardize

The average accuracies, by parameter, are in Tables 5.12, 5.13 and 5.14.

The best configuration would be to Standardize, use SMO as Optimization Routine, while the Kernel Function has a tie. The best and the function to classify the variable with more than two classes (see Section 5.1). For this latter, we can use one of the previous, already tuned, algorithms. Comparing the association of SVM with

- Kernel [Function] = {radial basis function, polynomial (order 1)}
- Multiclass [Function] = {L1QP, SMO, }

The best 5 results of this simulation are listed in Table 5.15

Multiclass	**Optimization**	**Kernel**	**AvAp**
DT	SMO	polynomial (order 1)	82.2822
DT	L1QP	RBF	82.2809
DT	SMO	RBF	82.2808
DT	L1QP	polynomial (order 1)	82.2782
kNN	SMO	polynomial (order 1)	81.7157

Table 5.15: Final results for SVM Imputation

The best setting for SVM is then: Standardize, SMO as Optimization Routine, Polynomial Kernel Function of order 1, and DT to classify the multiclass variables.

5.2.2 Final datasets for classification

There were two types of datasets created, depending on whether patients or variables are deleted, or missing values are imputed.

Two complete datasets were created using Deletion methods. By eliminating patients with MD, the dataset generated was *completePatients*, while deleting the variables with MD yielded the dataset *completeVariables*. The information of each dataset can be understood from the statistics of Section 4.1, but is presented in Table 5.16 for a more direct visualization:

Dataset	**Number of Variables**	**Number of Patients**
completePatients	27	28
completeVariables	12	97

Table 5.16: Complete datasets resulting from the use of Deletion

Instead of using the overall best, we would like to see how different imputation algorithms behave in the classification phase. The imputed datasets to be used are then the best setting for each of the best three algorithms:

- completeSVM - Kernel Function: polynomial (order 1); Kernel Scale: 1; Solver: Sequential Minimal Optimization; Standardized; DT to classify multiclass
- completeKNN - k = 5; Distance: correlation; Median
- completeDT - minLeaf = 12, criterion = deviance reduction

5.3 Classification

This section has the purpose of showing the results of the classification phase of this thesis.

5.3.1 Classifications by algorithm

The values shown were obtained from an average of nine values, each one corresponding to one output. In turn, each of those individual values were obtained from cross-validated models, to ensure the validity of results. The metric used was AUC, shown here as Average AUC (AvAU).

kNN

The first test to a classification algorithm was done with kNN, with the following parameters:

- k = {1, 10, 20, 30, 40, 50, 60, 70, 80, 90}
- Distance [metric] = {euclidean (1), seuclidean (2), cityblock (3), minkowski (4), chebychev (5), cosine (6), correlation (7), spearman (8), hamming (9), jaccard (10)}

The reasons for this choice were the same as in the previous section, when also using kNN. The 5 best settings in this cycle are present in Table 5.17, ordered by AvAU (also indicated),

k	**Distance**	**AvAU**	**Dataset**
20	7	0.6104	completeKNN
60	9	0.6081	completeKNN
20	8	0.6050	completeKNN
20	6	0.6011	completeKNN
10	6	0.6004	completeKNN

Table 5.17: First iteration of kNN Classification

Considering this results, a new iteration was performed, to cover all possible values for k, with the following parameters:

- k = {1, 2, ..., 97}
- Distance [metric] = {cosine (6), correlation (7), spearman (8), hamming (9)}

The results of this second attempt are the ones in Table 5.18

The optimal configuration for this algorithm is k=15 (or near) and the use of 'correlation' as the distance metric. It is interesting that the best MD handling mechanism to combine with this classification algorithm also uses kNN, and the similarities may not be a coincidence, since even the distance metric is the same.

To compare the performance of the different imputation methods, Table 5.19 contains the best setting for each complete dataset.

k	**Distance**	**AvAU**	**Dataset**
15	7	0.6446	completeKNN
16	7	0.6388	completeKNN
17	7	0.6378	completeKNN
18	8	0.6285	completeKNN
16	8	0.6279	completeKNN

Table 5.18: Second iteration of kNN Classification

Dataset	k	**Distance**	**AvAU**
completePatients	18	9	0.5891
completeVariables	16	8	0.6125
completeKNN	15	8	0.6446
completeDT	18	8	0.6244
completeSVM	17	8	0.6232

Table 5.19: Best results of kNN Classification for each Complete Dataset

ANN

The first iteration of ANN involved the following setting:

- Hidden [nodes] = {1,2,3,4,5,10,15,20,25,30}
- Train [function] = all

The reasons to choose this configuration were the same as in the imputation phase. Ordered by AvAU, the 5 best settings are registered in Table 5.20.

Train	**Hidden**	**AvAU**	**Dataset**
traincgf	5	0.6770	CompletePatients
trainscg	5	0.6742	CompletePatients
traingd	5	0.6698	CompletePatients
traingdm	5	0.6674	CompletePatients
traingda	5	0.6623	CompletePatients

Table 5.20: First iteration of ANN Classification

Since all these have 5 Hidden Units, the next step was to search around this number. Numbers under 5 were already included, so the parameters were:

- Hidden [nodes] = {6,7,8,9}
- Train [function] = all

There was only one setting that achieved a better result than the previous, which is in Table 5.18

In fact, this is the optimal configuration for the algorithm. To compare the performance of the different imputation methods, the best setup for each dataset is registered Table 5.22.

Train	Hidden	AvAU	Dataset
traincgf	7	0.6855	completePatients

Table 5.21: Second iteration of ANN Classification

Dataset	Train	Hidden	AvAU
completePatients	traincgf	7	0.6855
completeVariables	traincgf	25	0.6115
completeKNN	traingd	20	0.5872
completeDT	traingda	2	0.5888
completeSVM	trainoss	20	0.5832

Table 5.22: Best results of ANN Classification for each Complete Dataset

To classify with ANN, imputing MD does not seem to be the best approach. The best results were obtained when using the dataset with only originally complete patients, connected to a network with Conjugate Gradient with Fletcher-Reeves updates as the training function, as well as a hidden layer of 7 neurons.

DT

The parameters for the first run with DT in the creation of a classification model used these values for the parameters:

- MinLeaf = {1, 5, 10, 15, 20}
- Split [criterion] = all

Once again, the reasons to choose this are the same as the imputation phase The 5 best settings of this first iteration are listed in Table 5.23.

Split	MinLeaf	AvAU	Dataset
deviance	5	0.6001	CompletePatients
Gini's diversity	5	0.5951	CompletePatients
twoing	5	0.5951	CompletePatients
deviance	10	0.5707	CompletePatients
Gini's diversity	10	0.5642	CompletePatients

Table 5.23: First iteration of DT Classification

The three best have 5 as MinLeaf. Moreover, the next best results have 10 as MinLeaf, which lead us to think that maybe the best values are between 5 and 10. It appears that there is an order in the split criterion too, but without so much impact.

- MinLeaf = {6,7,8,9}
- Split [criterion] = all

Split	MinLeaf	AvAU	Dataset
deviance	7	0.6079	completePatients
deviance	5	0.6001	completePatients
deviance	8	0.5988	completePatients
Gini's diversity	5	0.5951	completePatients
twoing	5	0.5951	completePatients

Table 5.24: Second iteration of DT Classification

Dataset	Split	MinLeaf	AvAU
completePatients	deviance	7	0.6079
completeVariables	deviance	25	0.5183
completeKNN	deviance	20	0.5595
completeDT	deviance	2	0.5262
completeSVM	deviance	20	0.5277

Table 5.25: Best results of ANN Classification for each Complete Dataset

Table 5.24 shows the respective results.

To compare the performance of the different imputation methods, the best setup for each dataset is registered Table 5.25.

Once again, the Deletion dataset yielded better results than Imputation. The best results were obtained when using the dataset with only originally complete patients, and the parameters of the algorithm were 7 for minimum size of leaves and deviance for the split criterion.

SVM

The parameters to run SVM were the combination of all those that have been changed imputation:

- Kernel [Function] = all
- Optimization [Routine] = all
- Split [criterion] = all

The 5 best settings of this first iteration are listed in Table 5.26.

Kernel	Optimization	Standardize	AvAU	Dataset
polynomial(order 1)	L1QP	true 0.6277	CompletePatients	
RBF	L1QP	true 0.6215	CompletePatients	
RBF	SMO	true 0.6215	CompletePatients	
polynomial(order 1)	SMO	true 0.6215	CompletePatients	
polynomial(order 1)	L1QP	true 0.6096	CompletePatients	

Table 5.26: First iteration of DT Classification

To compare the performance of the different imputation methods, the best setup for each dataset is registered Table 5.27.

Dataset	Split	MinLeaf	MinLeaf	AvAU
completePatients	deviance	7	7	0.6079
completeVariables	deviance	25	25	0.5183
completeKNN	deviance	20	20	0.5595
completeDT	deviance	2	2	0.5262
completeSVM	deviance	20	20	0.5277

Table 5.27: Best results of SVM Classification for each Complete Dataset

One more time, the Deletion dataset had better results than Imputation. this algorithm did not have the best results, but still had a respectable contribution.

5.3.2 Final results

This section evaluated the results of this work, concerning the classification step. Since we had to work to develop the MD simulation, the first idea that we may have is that it does not help. Had the imputation been performed only using kNN, the kNN is a good option to have a high AvAU value.

The best setting, however, is

Chapter 6

Conclusion

This final chapter summarizes the work developed during this thesis, showing a glimpse of the paths this study may lead to.

6.1 Discussion

The main purpose of this academic study was to attempt to build a model for Recurrence in BC. To the best of our knowledge, there was never an attempt to predict BC relapse sites as multiple targets, as can be read in the Literature Review. The studies in the area of BC recurrence tend to analyze whether metastases appear or not, or predicting survival.

To handle a problem like this, in real world, one must take into account the problems that may emerge from raw data. In our case, the biggest problem we have come across was MD. To address it, we first simulated missing values to choose the best imputation method. Besides the best algorithm, SVM, we also used the best settings for the second and third best algorithms, respectively, as well as two datasets created by deletion of patients or variables (one each). However, we later found out that deleting records may sometimes a better option.

After selecting the best imputation methods, the next phase involved the cross-validation of several classification models, with different combinations of parameters for each algorithm used. Then, each configuration was trained for the several output, one at a time, being evaluated for each one. To handle the multi-target situation, each output was treated like an individual binary problem, but the goal was to unify this aspect, hoping to find a good model for the recurrence as a whole.

It is also an important point that this study was created with the concern to be replicable: the repetition of the same experience would yield the same results, since the random number generator is reset to the same point when necessary, the datasets with simulated MD were all saved, and the partitions of the classification phase were also saved.

In this type of private studies, with databases not available to the public, it is too difficult to establish comparisons. Nevertheless, 0.6855 is a very respectable value for an average AUC over nine values, all of them from cross-validation.

6.2 Future Work

To continue this study, we could hope to have access to an even bigger database. This would help us validate our results, while also providing the opportunity to

build even better models. However, that is not entirely up to us, and only time can bring such an opportunity.

Meanwhile, that are some points that could be further analyzed. One of them is the issue of imbalance, particularly in the output classes. Subsampling would be one way of dealing with this problem, but it would reduce even more the database; oversampling methods that copy data can be better, but they are not generating any new information; however, some oversampling methods like Synthetic Minority Oversampling TEchnique generate synthetic data, some synthetic and are proving to be efficient in providing balanced datasets to work upon.

Many other ML algorithms could also be tested in both imputation and classification phase. Moreover, the ones at study could be further improved, for example, with a more exhaustive search, although it requires much time. Feature Selection was also something implemented in this project, even if in a small scale. It could be used again before the classification phase, or even inside the imputation (choosing a subset of the complete variables to predict the incomplete ones from).

This thesis was a longer process than initially expected, but it will hopefully help others to explore this topic even further. There is still a long way to go, but thinking about all those that this work can help, we can always find more motivation.

Bibliography

[1] International Agency for Research on Cancer. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. *http://globocan.iarc.fr*. [Accessed: 2014-12-28].

[2] L. A. Korde, J. A. Zujewski, L. Kamin, S. Giordano, S. Domchek, W. F. Anderson, J. M. Bartlett, K. Gelmon, Z. Nahleh, J. Bergh, B. Cutuli, G. Pruneri, W. McCaskill-Stevens, J. Gralow, G. Hortobagyi, and F. Cardoso. Multidisciplinary meeting on male breast cancer: summary and research recommendations. *J. Clin. Oncol.*, 28(12):2114–2122, 2010.

[3] World Health Organization. Cancer fact sheet. *http://www.who.int/mediacentre/ factsheets/fs297*. [Accessed: 2014-12-21].

[4] O. Metzger-Filho, Z. Sun, G. Viale, K. N. Price, D. Crivellari, R. D. Snyder, R. D. Gelber, M. Castiglione-Gertsch, A. S. Coates, A. Goldhirsch, and F. Cardoso. Patterns of Recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from international breast cancer study group trials VIII and IX. *J. Clin. Oncol.*, 31(25):3083–3090, 2013.

[5] J. S. Reis-Filho, B. Weigelt, D. Fumagalli, and C. Sotiriou. Molecular profiling: moving away from tumor philately. *Sci Transl Med*, 2(47):47ps43, 2010.

[6] Instituto Português de Oncologia. About IPO-Porto. *http://ipoporto.pt*. [Accessed: 2015-01-22].

[7] Surveillance, Epidemiology, and End Results (SEER) Program. SEER Research - Accessing the 1973-2011 SEER data. *http://seer.cancer.gov/data/access.html*. [Accessed: 2015-02-16].

[8] Neil K Aaronson, Sam Ahmedzai, Bengt Bergman, Monika Bullinger, Ann Cull, Nicole J Duez, Antonio Filiberti, Henning Flechtner, Stewart B Fleishman, Johanna CJM de Haes, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in Oncology. *Journal of the national cancer institute*, 85(5):365–376, 1993.

[9] N. Ribelles, L. Perez-Villa, J. M. Jerez, B. Pajares, L. Vicioso, B. Jimenez, V. de Luque, L. Franco, E. Gallego, A. Marquez, M. Alvarez, A. Sanchez-Munoz, L. Perez-Rivas, and E. Alba. Pattern of recurrence of early breast cancer is different according to intrinsic subtype and proliferation index. *Breast Cancer Res.*, 15(5):R98, 2013.

[10] G. Bonadonna, G. N. Hortobagyi, and P. Valagussa. *Textbook of breast cancer: a clinical guide to therapy*. CRC Press, 2006.

[11] Mathworks. Unsupervised Learning. http://www.mathworks.com/discovery/unsupervised-learning.html. [Accessed: 2015-06-11].

[12] Mathworks. Classification using Nearest Neighbors. www.mathworks.com/help/stats/classification-using-nearest-neighbors.html. [Accessed: 2015-06-11].

[13] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[14] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[15] Mathworks. Choose a multilayer Neural Network training function. http://www.mathworks.com/help/nnet/ug/choose-a-multilayer-neural-network-training-function.html. [Accessed: 2015-06-11].

[16] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[17] F. Beca, R. Santos, D. Vieira, L. Zeferino, R. Dufloth, and F. Schmitt. Primary relapse site pattern in women with triple-negative breast cancer. *Pathol. Res. Pract.*, 210(9):571–575, 2014.

[18] L. J. Esserman, D. H. Moore, P. J. Tsing, P. W. Chu, C. Yau, E. Ozanne, R. E. Chung, V. J. Tandon, J. W. Park, F. L. Baehner, S. Kreps, A. N. Tutt, C. E. Gillett, and C. C. Benz. Biologic markers determine both the risk and the timing of recurrence in breast cancer. *Breast Cancer Res. Treat.*, 129(2):607–616, 2011.

[19] E. Lim, O. Metzger-Filho, and E. P. Winer. The natural history of hormone receptor-positive breast cancer. *Oncology (Williston Park, N.Y.)*, 26(8):688–694, 2012.

[20] Thomson Reuters. Web of Science. http://thomsonreuters.com/thomson-reuters-web-of-science/. [Accessed: 2015-01-20].

[21] C. J. van den Hurk, R. Eckel, L. V. van de Poll-Franse, J. W. Coebergh, J. W. Nortier, D. Holzel, W. P. Breed, and J. Engel. Unfavourable pattern of metastases in M0 breast cancer patients during 1978-2008: a population-based analysis of the Munich Cancer Registry. *Breast Cancer Res. Treat.*, 128(3):795–805, 2011.

[22] A. C. Chiang and J. Massague. Molecular basis of metastasis. *N. Engl. J. Med.*, 359(26):2814–2823, 2008.

[23] K. R. Hess and F. J. Esteva. Effect of HER2 status on distant recurrence in early stage breast cancer. *Breast Cancer Res. Treat.*, 137(2):449–455, 2013.

[24] K. D. Voduc, M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke. Breast cancer subtypes and the risk of local and regional relapse. *J. Clin. Oncol.*, 28(10):1684–1691, 2010.

[25] C. K. Anders, A. M. Deal, C. R. Miller, C. Khorram, H. Meng, E. Burrows, C. Livasy, K. Fritchie, M. G. Ewend, C. M. Perou, and L. A. Carey. The prognostic contribution of clinical breast cancer subtype, age, and race among patients with breast cancer brain metastases. *Cancer*, 117(8):1602–1611, 2011.

[26] Y. H. Park, S. Lee, E. Y. Cho, Y. L. Choi, J. E. Lee, S. J. Nam, J. H. Yang, J. S. Ahn, and Y. H. Im. Patterns of relapse and metastatic spread in HER2-overexpressing breast cancer according to estrogen receptor status. *Cancer Chemother. Pharmacol.*, 66(3):507–516, 2010.

[27] I. Vaz-Luis, R. A. Ottesen, M. E. Hughes, P. K. Marcom, B. Moy, H. S. Rugo, R. L. Theriault, J. Wilson, J. C. Niland, J. C. Weeks, and N. U. Lin. Impact of hormone receptor status on patterns of recurrence and clinical outcomes among patients with human epidermal growth factor-2-positive breast cancer in the National Comprehensive Cancer Network: a prospective cohort study. *Breast Cancer Res.*, 14(5):R129, 2012.

[28] R. Nishimura, T. Osako, Y. Okumura, R. Tashima, Y. Toyozumi, and N. Arima. Changes in the ER, PgR, HER2, p53 and Ki-67 biological markers between primary and recurrent breast cancer: discordance rates and prognosis. *World J Surg Oncol*, 9:131, 2011.

[29] P. K. Idirisinghe, A. A. Thike, P. Y. Cheok, G. M. Tse, P. C. Lui, S. Fook-Chong, N. S. Wong, and P. H. Tan. Hormone receptor and c-ERBB2 status in distant metastatic and locally recurrent breast cancer. Pathologic correlations and clinical significance. *Am. J. Clin. Pathol.*, 133(3):416–429, 2010.

[30] E. S. R. Mendoza, E. Moreno, and P. B. Caguioa. Predictors of early distant metastasis in women with breast cancer. *Journal of cancer research and clinical oncology*, 139(4):645–652, 2013.

[31] H. Kennecke, R. Yerushalmi, R. Woods, M. C. Cheang, D. Voduc, C. H. Speers, T. O. Nielsen, and K. Gelmon. Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol.*, 28(20):3271–3277, 2010.

[32] N. Niikura, J. Liu, N. Hayashi, S. L. Palla, Y. Tokuda, G. N. Hortobagyi, N. T. Ueno, and R. L. Theriault. Treatment outcome and prognostic factors for patients with bone-only metastases of breast cancer: a single-institution retrospective analysis. *Oncologist*, 16(2):155–164, 2011.

[33] L Wyld, E Gutteridge, S E Pinder, J J James, S Y Chan, K L Cheung, J F R Robertson, and A J Evans. Prognostic factors for patients with hepatic metastases from breast cancer. *Br J Cancer*, 89:284–290, 2003.

[34] M. Bergenfeldt, B. V. Jensen, B. Skjoldbye, and D. Nielsen. Liver resection and local ablation of breast cancer liver metastases–a systematic review. *Eur J Surg Oncol*, 37(7):549–557, 2011.

[35] K. Hoffmann, C. Franz, U. Hinz, P. Schirmacher, C. Herfarth, M. Eichbaum, M. W. Buchler, and P. Schemmer. Liver resection for multimodal treatment of breast cancer metastases: identification of prognostic factors. *Ann. Surg. Oncol.*, 17(6):1546–1554, 2010.

[36] V. Treska, M. Cerna, V. Liska, I. Treskova, A. Narsanska, and J. Bruha. Surgery for breast cancer liver metastases - factors determining results. *Anticancer Res.*, 34(3):1281–1286, 2014.

[37] B. Elsberger, C. S. Roxburgh, and P. G. Horgan. Is there a role for surgical resections of hepatic breast cancer metastases? *Hepatogastroenterology*, 61(129):181–186, 2014.

[38] M. Weinrich, C. Weiss, J. Schuld, and B. M. Rau. Liver resections of isolated liver metastasis in breast cancer: results and possible prognostic factors. *HPB Surg*, 2014:893829, 2014.

[39] K. Altundag, M. L. Bondy, N. Q. Mirza, S. W. Kau, K. Broglio, G. N. Hortobagyi, and E. Rivera. Clinicopathologic characteristics and prognostic factors in 420 metastatic breast cancer patients with central nervous system metastasis. *Cancer*, 110(12):2640–2647, 2007.

[40] K. L. Chaichana, S. Gadkaree, K. Rao, T. Link, D. Rigamonti, M. Purtell, I. Browner, J. Weingart, A. Olivi, G. Gallia, C. Bettegowda, H. Brem, M. Lim, and A. Quinones-Hinojosa. Patients undergoing surgery of intracranial metastases have different outcomes based on their primary pathology. *Neurol. Res.*, 35(10):1059–1069, 2013.

[41] M. Smid, Y. Wang, Y. Zhang, A. M. Sieuwerts, J. Yu, J. G. Klijn, J. A. Foekens, and J. W. Martens. Subtypes of breast cancer show preferential site of relapse. *Cancer Res.*, 68(9):3108–3114, 2008.

[42] M. Karabatak and M. C. Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2, Part 2):3465–3469, 2009.

[43] X. Yang, X. Ai, and J. M. Cunningham. Computational prognostic indicators for breast cancer. *Cancer Manag Res*, 6:301–312, 2014.

[44] M.-S. Chen, J. Han, and P. S. Yu. Data mining: an overview from a database perspective. *Knowledge and data Engineering, IEEE Transactions on*, 8(6):866–883, 1996.

[45] A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications*, 36(4):8204–8211, 2009.

[46] S.-M. Chou, T.-S. Lee, Y. E. Shao, and I-F. Chen. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 27(1):133–142, 2004.

[47] M. Paliwal and U. A. Kumar. Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1):2–17, 2009.

[48] P.C. Pendharkar, J.A. Rodger, G.J. Yaverbaum, N. Herman, and M. Benner. Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications*, 17(3):223 – 232, 1999.

[49] X. Xiong, Y. Kim, Y. Baek, D. W. Rhee, and S.-H. Kim. Analysis of breast cancer using data mining and statistical techniques. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on*, pages 82–87. IEEE, 2005.

[50] S. Mani, M. J. Pazzani, and J. West. Knowledge discovery from a breast cancer database. In *Artificial Intelligence in Medicine*, pages 130–133. Springer, 1997.

[51] O. Intrator and N. Intrator. Interpreting neural-network results: a simulation study. *Computational statistics & data analysis*, 37(3):373–393, 2001.

[52] Z. H. Zhou and Y. Jiang. Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. *IEEE Trans Inf Technol Biomed*, 7(1):37–42, Mar 2003.

[53] J. M. Jerez-Aragones, J. A. Gomez-Ruiz, G. Ramos-Jimenez, J. Munoz-Perez, and E. Alba-Conejo. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artif Intell Med*, 27(1):45–63, Jan 2003.

[54] A. R. Razavi, H. Gill, O. Stal, M. Sundquist, S. Thorstenson, H. Ahlfeldt, and N. Shahsavar. Exploring cancer register data to find risk factors for recurrence of breast cancer–application of Canonical Correlation Analysis. *BMC Med Inform Decis Mak*, 5:29, 2005.

[55] A. R. Razavi, H. Gill, H. Åhlfeldt, and N. Shahsavar. A data pre-processing method to increase efficiency and accuracy in data mining. In *Artificial Intelligence in Medicine*, volume 3581 of *Lecture Notes in Computer Science*, pages 434–443. Springer Berlin Heidelberg, 2005.

[56] A. R. Razavi, H. Gill, H. Åhlfeldt, and N. Shahsavar. Predicting metastasis in breast cancer: Comparing a decision tree with domain experts. *Journal of Medical Systems*, 31(4):263–273, 2007.

[57] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23(1):30–37, 2007.

[58] Nature. Nature Journal. *http://www.nature.com/nature*. [Accessed: 2015-03-24].

[59] L. J. van't Veer, H. Dai, M. J. van De Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[60] M. Dettling and P. Bühlmann. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106 – 131, 2004. Special Issue on Multivariate Methods in Genomic Data Analysis.

[61] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[62] T. Jonsdottir, E. T. Hvannberg, H. Sigurdsson, and S. Sigurdsson. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Systems with Applications*, 34(1):108–118, 2008.

[63] Q. Fan, C.-J. Zhu, and L. Yin. Predicting breast cancer recurrence using data mining techniques. In *Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on*, pages 310–311, April 2010.

[64] S. Belciug, F. Gorunescu, A.-B. Salem, and M. Gorunescu. Clustering-based approach for detecting breast cancer recurrence. In *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*, pages 533–538, Nov 2010.

[65] M. Lichman. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2013. [Accessed: 2015-03-30].

[66] L. G. Ahmad, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi. Using three machine learning techniques for predicting breast cancer recurrence. *Journal of Health and Medical Informatics*, 4(2):1–3, 2013.

[67] A. P. Pawlovsky and M. Nagahashi. A method to select a good setting for the knn algorithm when using it for breast cancer prognosis. In *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*, pages 189–192. IEEE, 2014.

[68] Z. Beheshti, S. M. Hj. Shamsuddin, E. Beheshti, and S. S. Yuhaniz. Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis. *Soft Computing*, 18(11):2253–2270, 2014.

[69] A Berruti, M Tampellini, M Torta, T Buniva, G Gorzegno, and L Dogliotti. Prognostic value in predicting overall survival of two mucinous markers: Ca 15-3 and ca 125 in breast cancer patients at first relapse of disease. *European Journal of Cancer*, 30(14):2082–2084, 1994.

[70] Michael J Duffy, Catherine Duggan, Rachel Keane, Arnold DK Hill, Enda McDermott, John Crown, and Niall O'Higgins. High preoperative ca 15-3 concentrations predict adverse outcome in node-negative and node-positive breast cancer: study of 600 patients with histologically confirmed breast cancer. *Clinical Chemistry*, 50(3):559–563, 2004.

[71] Stephen G Shering, Frances Sherry, Enda W McDermott, Niall J O'Higgins, and Michael J Duffy. Preoperative ca 15-3 concentrations predict outcome of patients with breast carcinoma. *Cancer*, 83(12):2521–2527, 1998.

[72] Eero Juha Kumpulainen, Riitta Johanna Keskikuru, and Risto Tapio Johansson. Serum tumor marker ca 15.3 and stage are the two most powerful predictors of survival in primary breast cancer. *Breast cancer research and treatment*, 76(2):95–102, 2002.

[73] Rafael Molina, Jose M Auge, Blanca Farrus, Gabriel Zanón, Jaume Pahisa, Montserrat Muñoz, Aureli Torne, Xavier Filella, Jose M Escudero, Pedro Fernandez, et al. Prospective evaluation of carcinoembryonic antigen (cea) and carbohydrate antigen 15.3 (ca 15.3) in patients with primary locoregional breast cancer. *Clinical chemistry*, 56(7):1148–1157, 2010.

[74] G. Grassetto, A. Fornasiero, D. Otello, G. Bonciarelli, E. Rossi, O. Nashimben, M. Anna Minicozzi, G. Crepaldi, F. Pasini, E. Facci, G. Mandoliti, M. C. Marzola, A. Al-Nahhas, and D. Rubello. 18f-fdg-pet/ct in patients with breast cancer and rising ca 15-3 with negative conventional imaging: A multicentre study. *European Journal of Radiology*, 80(3):828 – 833, 2011.

[75] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[76] Qiang Wang. A hybrid sampling svm approach to imbalanced data classification. In *Abstract and Applied Analysis*, volume 2014. Hindawi Publishing Corporation, 2014.